回归分析

主讲人：刘宏志

liuhz@ss.pku.edu.cn

北京大学

回归分析的动机

- 相关分析的不足：
  - 只能分析现象之间相关关系的方向和相关的密切程度
  - 不能判断现象之间具体的数量变动依存关系
  - 不能根据相关系数估计或预测应变量y可能发生的数值
- 回归分析：
  - 对具有相关关系的两个或两个以上变量之间数量变化的一般关系进行测定
  - 确定应变量和自变量之间数量变动关系的数学表达式
  - 以便对应变量进行估计或预测的统计分析方法

- 二者的关系：
  - 相关分析的主要任务是研究变量间相关关系的表现形式和密切程度
  - 回归分析是在相关分析的基础上，进一步研究现象之间的数量变化规律

- 变量x_i与变量y的回归模型一般表示为

$$y = \varphi(x_1, x_2, x_3, \cdots, x_p) + \varepsilon$$

“应变量”或“被解释变量”(dependent variable)

“自变量”或“解释变量”(independent variables)

随机变量

为什么称为“回归”分析？

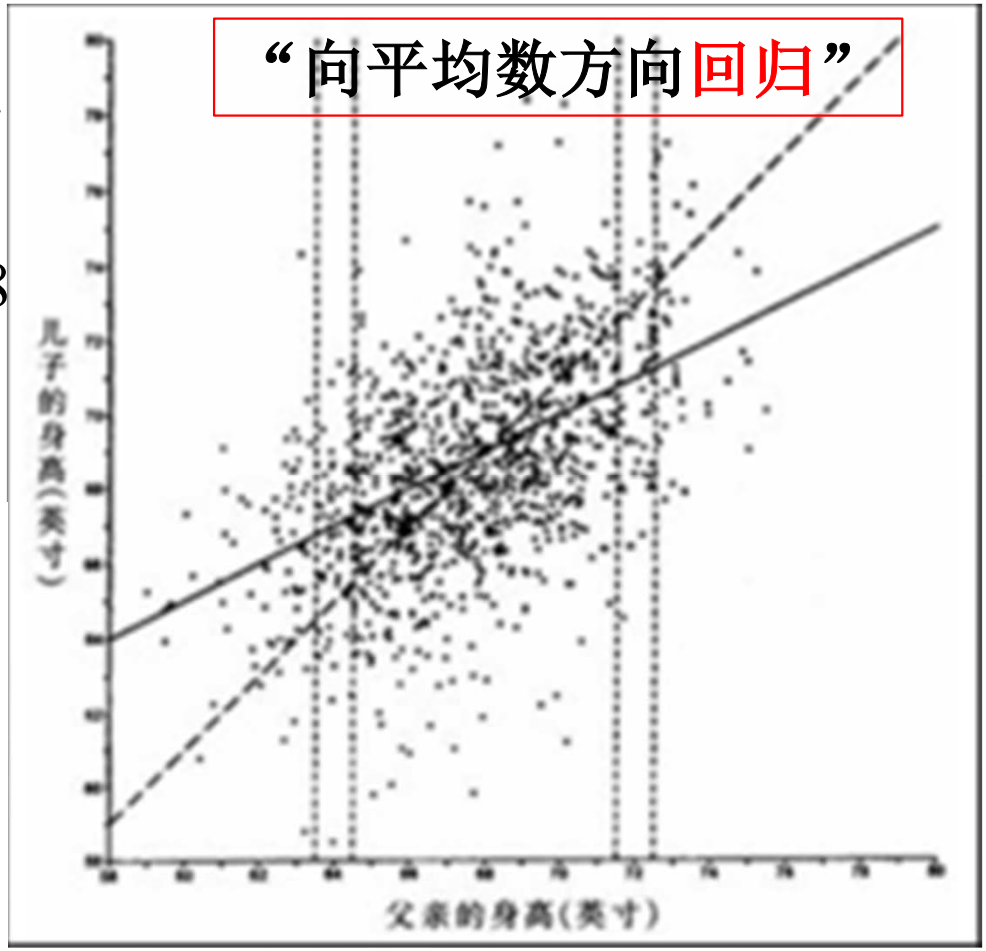
回归分析的起源

- F.Gallton和K.Pearson收集了1078个家庭的身高记录
- 寻找儿子们身高与父亲们身高之间关系的具体形式
- 下图是根据1078个家庭的调查所作的散点图

父亲们身高：
平均值：$\bar{X}\approx68$；标准差：$S_x\approx2.7$
儿子们身高：
平均值：$\bar{Y}\approx69$；标准差：$S_Y\approx2.8$

斜虚线：依平均身高推测的关系线

斜实线（回归线）：线上的点是当给定某一X_i值（父亲身高值）时，对应的若干Y_i值（儿子的身高值）与之（直线上点Y值）离差平方和最小的直线



回归分析的主要内容

- 根据研究目的和现象之间的内在联系，确定自变量和应变量
- 确定回归模型的类型及数学表达式 (曲线拟合)
- 对回归分析模型进行评价与诊断 (统计检验)
- 根据给定的自变量值确定应变量的值 (预测)

回归分析的特点

- 回归分析必须根据研究目的确定其中一个为应变量，其余为自变量
  - 而相关分析可以不用区分自变量和应变量
- 回归分析中，要求应变量是随机的，而自变量的值则是给定的
  - 而相关分析中，两个变量要求都是随机的
- 若变量之间互为因果，则可求出两个回归方程
  - 而两个变量之间只能求出一个相关系数

常见的曲线拟合方法

（1）使偏差绝对值之和最小，即

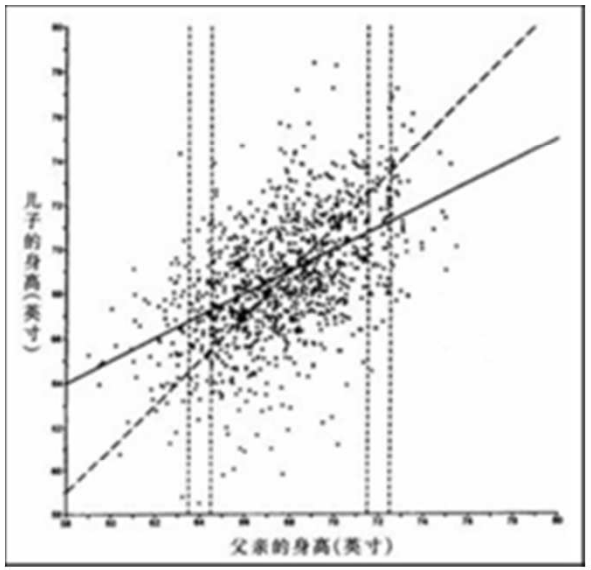
$$\min_{\varphi} \sum_{i=1}^{m} |\delta_i| = \sum_{i=1}^{m} |\varphi(x_i) - y_i|$$

（2）使偏差绝对值最大的最小，即

$$\min_{\varphi} \max_{i} |\delta_i| = |\varphi(x_i) - y_i|$$

（3）使偏差平方和最小，即

$$\min_{\varphi} \sum_{i=1}^{m} \delta_i^2 = \sum_{i=1}^{m} (\varphi(x_i) - y_i)^2$$



按偏差平方和最小的原则选取拟合曲线的方法，称为最小二乘法。

回归模型的类型

回归模型

一个自变量

两个及两个以上自变量

一元回归

多元回归

线性回归

非线性回归

线性回归

非线性回归

$$y = \varphi(x_1, x_2, x_3, \cdots, x_p) + \varepsilon$$

一元线性回归分析

误差项

理论模型 $y = \alpha + \beta x + \varepsilon$

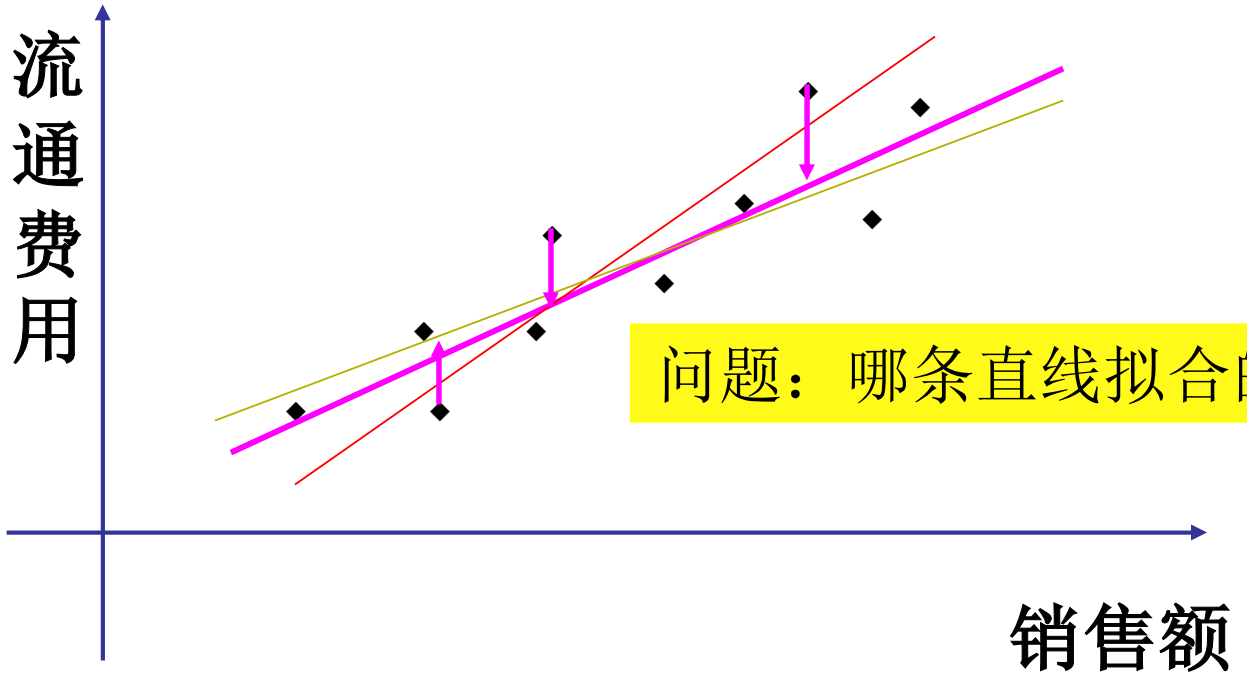
回归参数

估计模型 $y_c = a + bx$

回归参数的估计值

- a是直线的截距，b是直线的斜率
- 应变量y的估计值记为y_c

参数估计（曲线拟合）



思路：<u>离差的平方和最小</u>

最小二乘法

北京大学

设估计模型为： $y_c = a + bx$

$$\min Q = \sum (y - y_c)^2$$

$$\min Q = \sum (y - a - bx)^2$$

$$\frac{\partial Q}{\partial a} = \sum 2(y - a - bx)(-1) = 0$$

$$\frac{\partial Q}{\partial b} = \sum 2(y - a - bx)(-x) = 0$$

整理得：

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

最后解得：

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$a = \bar{y} - b\bar{x}$$

回归系数b与相关系数r

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{\sigma^2{}_{xy}}{\sigma^2{}_x} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = r \cdot \frac{\sigma_y}{\sigma_x}$$

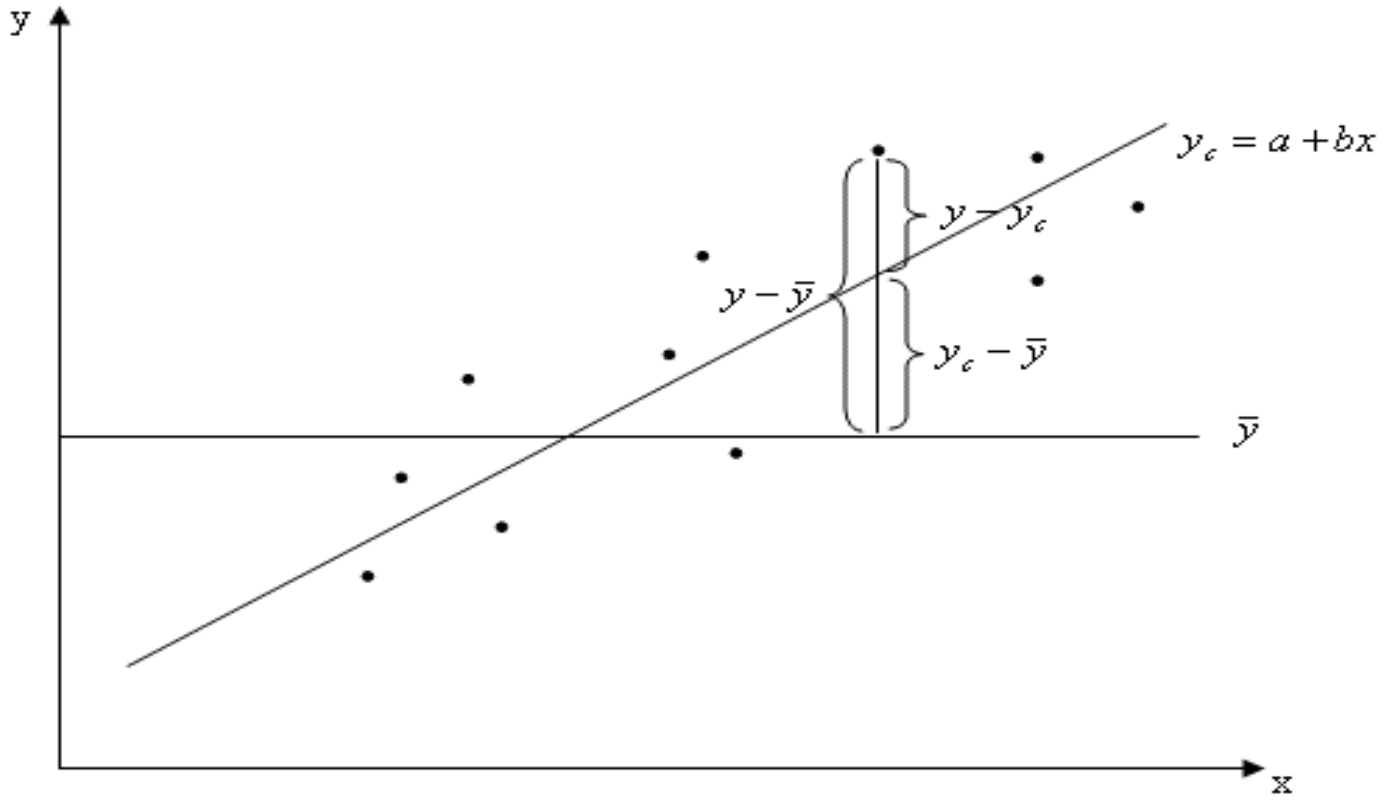
回归系数b和相关系数r的联系

一元线性回归的整体统计检验（单因素方差分析）

- 提出统计假设
  - 原假设H_0:　　$\beta_1 = 0$;
  - 备择假设H_1:　$\beta_1 \neq 0$
- 构建统计量：变差分解
  - 变差：直线回归中，应变量y的取值围绕平均值上下波动，这种波动称为变差
  - 变差的来源：

    (1)自变量x的取值不同造成的

    (2)除x以外的其他因素(如x对y的非线性影响、测量误差等)的影响造成的



$$y - \bar{y} = (y - y_c) + (y_c - \bar{y})$$

两边平方得 $$(y - \bar{y})^2 = (y - y_c)^2 + 2(y - y_c)(y_c - \bar{y}) + (y_c - \bar{y})^2$$

两边求和并化简得 $$\sum (y - \bar{y})^2 = \sum (y - y_c)^2 + \sum (y_c - \bar{y})^2$$

变差的分解

$$\sum(y-\bar{y})^2=\sum(y_c-\bar{y})^2+\sum(y-y_c)^2$$

总变差 (*SST*)　　回归变差 (*SSR*)　　剩余变差或残差 (*SSE*)

$$SST = SSR + SSE$$

反映由于 x 与 y 之间的线性关系引起的 y 的取值变化，也称可解释的变差

因为 $\sum(y_c-\bar{y})^2=\sum(a+bx-a-b\bar{x})^2=b^2\sum(x-\bar{x})^2$

F检验

- F--统计量：

$$F = \frac{SSR/df_R}{SSE/df_E} = \frac{SSR/1}{SSE/n-2}$$

- 统计决策：
  - 若$F > F_\alpha(1, n-2)$, 则拒绝$H_0: \beta_1 = 0$，接受$H_1: \beta_1 \neq 0$
  - 说明用函数$y_i = \beta_0 + \beta_1 x_i$来描述应变量$y$与自变量$x$的关系是合适的
  - 即回归模型是显著性的
- 方差分析表

方差来源	平方和	自由度	均方	F值
回归	*SSR*	1	*MSR= SSR/ 1*	*F=MSR/MSE*
残差	*SSE*	*n*−2	*MSE= SSE/n*−2	
总计	*SST*	*n*−1		

判定系数

- 回归变差占总变差的比例，称为判定系数

$$R^2 = \frac{SSR}{SST} = \frac{\sum (y_c - \bar{y})^2}{\sum (y - \bar{y})^2} = 1 - \frac{\sum (y_c - y)^2}{\sum (y - \bar{y})^2}$$

- 反映回归直线的拟合优度的统计指标，取值范围为 [0 , 1]
- $R^2 \to 1$，说明回归方程拟合得越好
- $R^2 \to 0$，说明回归方程拟合越差。
- 在一元线性模型中，判定系数等于相关系数的平方 ($R^2 = r^2$)

回归方程的统计检验：参数显著性的检验

- 判断每个自变量对于回归模型是否必要的
- 一元线性回归模型截距和斜率的显著性检验（*t*检验）

$$t=\frac{a}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^{2}}{\sum(x-\bar{x})^{2}}}}\propto t(n-2)$$

$$t=\frac{b}{\hat{\sigma}\sqrt{\frac{1}{\sum(x-\bar{x})^{2}}}}\propto t(n-2)$$

$$\text{其中 } \hat{\sigma}=\sqrt{\frac{\sum(y-y_c)^2}{n-2}}=S_{yx}$$

回归分析的模型检查

- 前提假设：

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i \qquad i=1\ ,\ 2,\ \ldots\ ,\ n$$

  - 其中$\varepsilon_i \sim N(0\ ,\sigma^2)$
  - ε_i 是相互独立且它们的方差相同

- 模型检查：检查对模型所做的假设是否成立
  - ε_i 是相互独立的随机变量序列的检查
  - ε_i 是方差齐性的检查

残差图

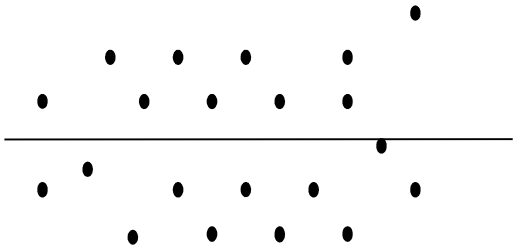
- 残差：观测值与理论值的差

$$\gamma_i = y_i - \hat{y}_i$$

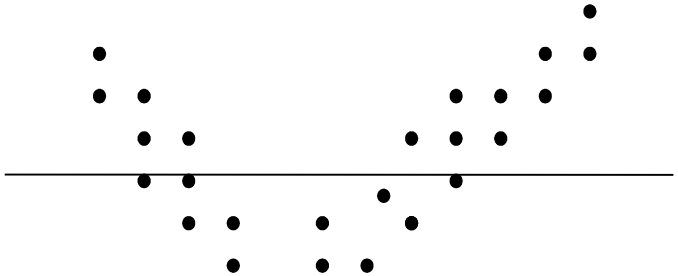
- 标准化残差：

$$e_i = \gamma_i \Big/ \sqrt{D(\gamma_i)}$$

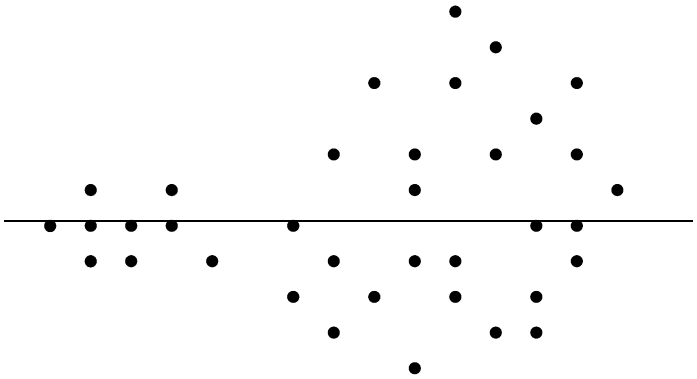
- 残差图：
  - 以x为坐标横轴，残差e为坐标纵轴，由所有点(x_i, e_i)构成
- 可检验随机变量序列ε_i的独立性，正态性和方差整齐性
  - 理论上可证明$e_1, e_2, \cdots, e_n$应相互独立且近似的服从N(0, 1)
  - 正常残差图：残差图中的点随机分布在-2到+2之间的子带里



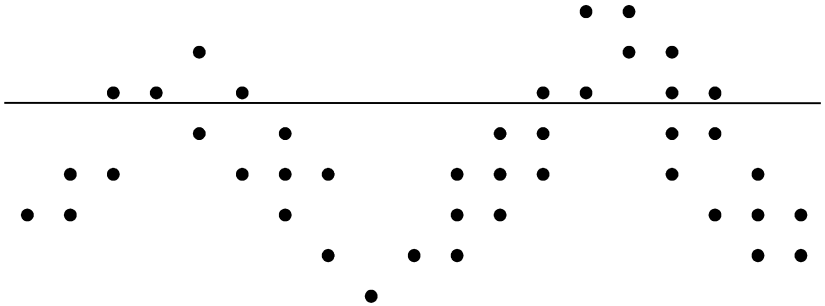
1．正常的残差图



2．直线回归模型不合适



3.方差齐性不成立



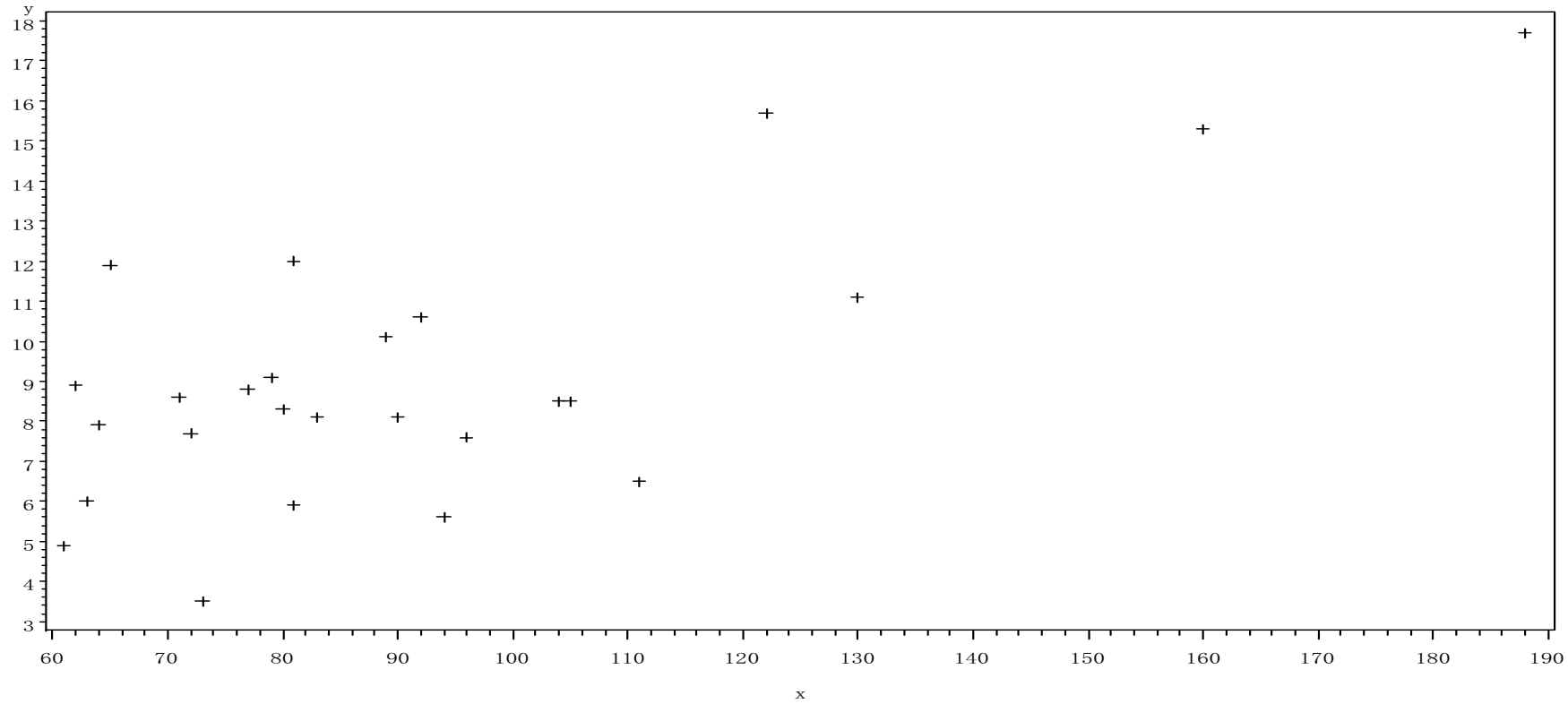
4.误差项不独立

方差齐性的诊断及修正方法

- 误差方差非齐性时, 残差图不正常
- 可通过对应变量作适当的变换
  - 令$z=f(y)$使得回归分析中误差的方差接近于齐性
  - 变换后重新做回归及残差图, 如残残差图有改善或已属正常, 则该变换是合适的; 否则改变变换函数计算直到找到合适的变换为止
- 常用变换有:

$$Z=\ln(Y) \qquad Z=\sqrt{Y}\,(Y>0) \qquad Z=1/Y\,(Y\neq 0)$$

采样数据表														
编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x	77	64	62	72	71	83	79	94	104	96	61	90	81	122
y	8.8	7.9	8.9	7.7	8.6	8.1	9.1	5.6	8.5	7.6	4.9	8.1	12.0	15.7
编号	15	16	17	18	19	20	21	22	23	24	25	26		
x	65	130	111	160	188	81	92	80	63	105	89	73		
y	11.9	11.1	6.5	15.3	17.7	5.9	10.6	8.3	6.0	8.5	10.1	3.5		



1．建立回归方程

由所给的数据得

$$\sum x_i = 2396 \qquad \bar{x} = 92.15 \qquad \sum y_i = 236.9 \qquad \bar{y} = 9.11$$

$$\sum x_i^2 = 243990 \quad \sum y_i^2 = 2435.23 \quad \sum x_i y_i = 23618.9$$

于是得

$$b = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n} = \frac{23618.9 - 2396 \times 236.9/26}{243990 - (2396)^2/26} = 0.0771$$

$$a = \bar{y} - b\bar{x} = 9.11 - 0.0771 \times 92.15 = 2.01$$

可得回归方程为

$$\hat{y} = 2.01 + 0.0771\,x$$

2. 回归方程显著性检验

总平方和 $SS_T = \sum (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2 / n$

$= 2435.23 - (236.9)^2 / 26 = 276.71$

回归平方和

$SS_R = \sum (\hat{y}_i - \bar{y})^2 = b^2 \sum (x_i - \bar{x})^2 = b^2 [\sum x_i^2 - (\sum x_i)^2 / n]$

$= (0.0771)^2 \times [243990 - (2396)^2 / 26] = 137.81$

剩余平方和 $SS_E = SS_T - SS_R = 276.71 - 139.52 = 138.90$

回归方程的方差分析				
变异来源	平方和(SS)	自由度(df)	均方(MS)	F值
回 归	1	137.81	137.81	23.81**
剩 余	24	138.90	5.79	
总变异	25	276.71		

P<0.0001， 表明回归方程是显著的

多元线性回归

- 设影响应变量y的自变量有p个,并分别记为$x_1, x_2, \ldots, x_p$
- 多元线性回归模型是指这些自变量对应变量的影响是线性的，即

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

其中$f(x_1, x_2, \cdots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

称为p个自变量 $x_1, x_2, \ldots, x_p$ 的线性回归函数.

- 记n组样本为$(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i)\ (i = 1, 2, \cdots, n)$，由上式可得到

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1,2,\cdots n$$

其中$\varepsilon_i \sim N(0, \sigma^2)$, $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$

- 主要问题：
  - 基于模型对未知参数$\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_p$进行估计、检验
  - 并利用得到的回归模型进行预测或控制

回归模型的矩阵表示

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\
&\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n
\end{aligned}
$$

$$
\underset{n\times 1}{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}} = \underset{n\times(p+1)}{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}} \underset{(p+1)\times 1}{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}
$$

$$
Y = X\beta + \varepsilon, \text{其中} Y = (y_1, y_2, \cdots, y_n)^T
$$

$$
\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)^T, \quad \varepsilon = (\varepsilon_1, \varepsilon_2, \cdots \varepsilon_n)^T, \quad \varepsilon \sim N(0, \sigma^2 I_n)
$$

回归系数的最小二乘估计

- 求解一组参数$\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_p$，使得如下定义的平方和Q达到最小：

$$Q(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots \beta_p x_{ip})^2$$

- 由多元函数的极值理论,分别求Q关于各个参数的偏导数,并令其等于零

$$\begin{cases} -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots \hat{\beta}_p x_{ip}) = 0 \\ -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots \hat{\beta}_p x_{ip}) x_{i1} = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots \hat{\beta}_p x_{ip}) x_{ip} = 0 \end{cases}$$

- 用矩阵表示为$(X^T X)\hat{\beta} = X^T Y$
- 若$X^T X$可逆,则方程组的解为$\hat{\beta} = (X^T X)^{-1} X^T Y$

回归方程的显著性检验(方差分析)

- 统计假设：H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$；　　H_1 : $\beta_i \neq 0$
- 变差平方和与自由度分解

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

自由度：　$df_T = df_E + df_R$，　其中 $df_T = n-1$，$df_R = p$，$df_E = (n-1)-p$

- F ——统计量：　$F = \dfrac{SSR/p}{SSE/(n-p-1)} \sim F(p, n-p-1)$

拒绝域为：$F > F_\alpha(p, n-p-1)$

- 方差分析表

变异来源	平方和	自由度	均方	F值
回归	SS_R	p	$MS_R = SS_R / p$	$F = MS_R / MS_E$
残差	SS_E	n-p-1	$MS_E = SS_E / n$-p-1	
总变异	SS_T	$n-1$		

回归系数的显著性检验

- 对回归系数是否为0进行逐个检验
- 统计假设：

$$H_0^{(i)}: \beta_i = 0 \; ; \; H_1^{(i)}: \beta_i \neq 0 \; (i = 1,2,\cdots,p)$$

- t ——统计量

$$t_i = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}} \sim t(n-p-1)$$

$$S_{\hat{\beta}_i} = \sqrt{\frac{{S_y}^2}{\sum (x_i - \bar{x}_i)^2}}$$

拒绝域为：$t > t_\alpha(n-p-1)$

自变量的选择

- 动机
  - 若漏掉显著的自变量，实际预测时会产生较大的偏差
  - 若包括了不显著的自变量，也会影响到预测的精度
- 最优回归方程
  - 对y的作用有统计学意义的自变量，全部选入回归方程
  - 对y的作用没有统计学意义的自变量，一个也不引入
- 问题描述：
  - 从自变量集$\{x_1, x_2, \cdots, x_p\}$中选出子集$A = \{x_{i_1}, x_{i_2}, \cdots, x_{i_l}\}$
  - 使得建立应变量 y 与子集的回归方程“最优”

自变量选择：评价标准

- 均方误差S^2最小：

$$S^2(A) = SSE(A)/(n-l-1)$$

其中$SSE(A)$是残差平方和, l是子集中自变量的个数

- 赤池信息准则（Akaike information criterion, AIC）

$$\text{AIC}=n\ \ln(\text{SSE})+2l$$

- 贝叶斯信息准则(Bayesian Information Criterion, BIC)

$$\text{BIC}=n\ \ln(\text{SSE})+l\ln(n)$$

- 修正的R^2准则

$$R'^2 = 1 - \frac{n-i}{n-l}(1-R^2)$$

自变量选择：选择方法

- 全局择优法：
  - 从所有可能的回归模型选取最优者
- 向后剔除法（backward selection）
- 向前引入法（forward selection）
- 逐步回归法（stepwise regression）

向后剔除 (backward elimination)

1. 先构建包括所有自变量的回归模型
2. 然后选择剔除一个自变量，使得剔除后的评价标准最优
3. 如此反复进行，将自变量从模型中逐个剔除，直至剔除一个自变量无法优化评价准则为止

向前选择 (forward selection)

1. 对每个自变量分别拟合一元线性回归模型，找出并保留使评价标准最优的一个模型
2. 分别拟合引入剩余的k-1个自变量的线性回归模型，找出使评价标准最优的一个模型
3. 如此反复进行，直至模型外的自变量均无统计显著性为止

逐步回归 (stepwise regression)

1. 将向前选择和向后剔除两种方法结合起来筛选自变量
2. 在增加了一个自变量后，它会对模型中所有的变量进行考察，看看有没有可能剔除某个自变量。如果在增加了一个自变量后，前面增加的某个自变量对模型的贡献变得不显著，这个变量就会被剔除
3. 按照以上方法不停地增加变量并考虑剔除以前增加的变量的可能性，直至增加变量使评价标准更优
4. 在前面步骤中增加的自变量在后面的步骤中有可能被剔除，而在前面步骤中剔除的自变量在后面的步骤中也可能重新进入到模型中

例：由于环境作用对光合速率的影响很大,要得到能反映环境对光合作用影响的数据,必须在不同的天气下测定光合作用各种指标. 应变量 y —光合速率; x_1—气孔导度; x_2—胞间二氧化碳浓度; x_3—蒸腾速率; x_4—叶片水汽压亏损; x_5—叶片的温度; x_6—相对湿度.

环境对光合作用影响数据表

观测号	y	x_1	x_2	x_3	x_4	x_5	x_6
1	8.37	0.0996	204	2.80	2.78	34.81	1063
2	8.19	0.0987	202	2.79	2.79	35.06	1069
3	8.03	0.1030	208	3.11	2.99	35.81	1114
4	8.32	0.1040	199	3.44	3.27	36.76	1162
5	8.38	0.0990	192	3.48	3.45	37.46	1219
6	8.16	0.1010	200	3.78	3.65	37.87	1231
7	7.44	0.0979	208	3.88	3.88	38.39	1288
8	7.28	0.0965	208	3.90	3.95	38.72	1300
9	6.50	0.0893	205	3.85	4.20	39.61	1295
10	7.85	0.0988	203	3.45	3.44	46.68	1193

Variable Selection

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
1	0.7297	0.6959	73.2633	-20.0712	0.33555	x1
1	0.6419	0.5971	98.9888	-17.2605	0.38618	x4
2	0.8723	0.8358	33.4332	-25.5734	0.24651	x3 x4
2	0.8601	0.8202	37.0035	-24.6624	0.25800	x1 x2
3	0.9778	0.9667	4.5158	-41.0567	0.11109	x2 x3 x4
3	0.9574	0.9360	10.5042	-34.5384	0.15389	x1 x2 x4
4	0.9894	**0.9809**	**3.1073**	**-46.4615**	**0.08404**	**x1 x2 x3 x4**
4	0.9811	0.9659	5.5548	-40.6524	0.11236	x2 x3 x4 x6
5	0.9897	0.9769	5.0075	-44.7880	0.09243	x1 x2 x3 x4 x6
5	0.9894	0.9762	5.0987	-44.4892	0.09383	x1 x2 x3 x4 x5
6	0.9898	0.9693	7.0000	-42.8129	0.10660	x1 x2 x3 x4 x5 x6

含定性变量的回归分析

- 设置虚拟变量：
  - 若某个自变量有k个不同的水平,则设置k-1个虚拟的变量. 令

$$u_j = \begin{cases} 1, & \text{当该自变量取第个}j\text{水平时,} \\ 0, & \text{当该自变量取其他水平时} \end{cases} \qquad j = 1, 2, \cdots, k-1$$

- 把 u_1 , u_2 , …, $u_{k\text{-}1}$与其他定量的变量一起建立线性回归方程

非线性回归

- 可化为线性的回归模型

1、指数曲线模型 $y_c = ab^x \Rightarrow \ln y_c = \ln a + x \ln b$

2、对数曲线模型 $y_c = a + b\ln x, 令 x' = \ln x$

3、双曲线模型 $\frac{1}{y_c} = a + b\frac{1}{x}, 令 y_c' = \frac{1}{y_c}, x' = \frac{1}{x}$

4、幂函数曲线模型 $y_c = ax^b \Rightarrow \lg y_c = \lg a + b\lg x$

5、抛物线模型 $y_c = a + bx + cx^2, 令 x_1 = x, x_2 = x^2$

非线性回归分析

- 不可转换成线性的趋势模型
  - 无论用什么方式变换都不能实现线性化的模型
- 常用处理方法：
  - 高斯--牛顿迭代法
  - 借助于泰勒级数展开式进行逐次的线性近似估计

回归分析：模型选择

- 动机：
  - 同一问题可以有多个不同的拟合曲线
  - 不同曲线拟合效果不同
- 模型选择方法：
  - 通过观察选择多个函数进行计算、分析、比较
  - 选择使评价标准最优的数学模型
    - 评价标准可用变量选择中的各种标准