《传播统计学》

回归分析

教师：林志良
邮箱：linzhl@nfu.edu.cn
个人网站：www.zhilianglin.com

目录

回归分析介绍

- 之前介绍的推断性统计分析方法（t检验、单因素方差分析、卡方检验、相关分析）都是分析两变量之间的关系的。

- 回归分析则可以分析一个或以上的自变量与一个因变量的关系。
  - 只有一个自变量的回归分析——**一元回归（简单线性回归）**
  - 有多个自变量的回归分析——**多元回归**

回归分析介绍

- 之前介绍的推断性统计分析方法对变量类型有严格要求（分类型变量/数值型变量）。

- 回归分析则可以适用于多种变量类型的情况。（这里我们仅介绍因变量为数值型变量，自变量为数值型变量+分类型变量的情况。）

回归分析介绍

相关 vs. 回归

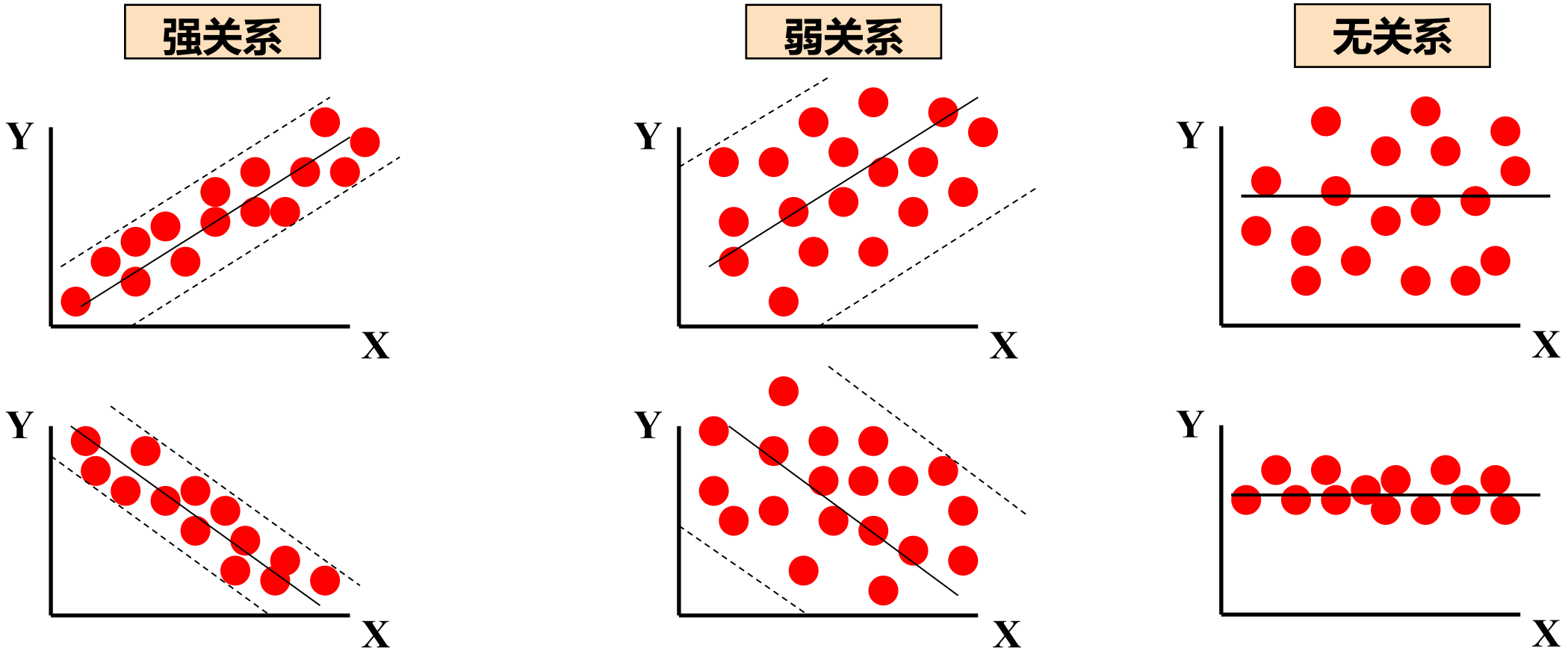
- **相关**（correlation）用于：分析两变量线性关系的强度

  （无所谓自变量和因变量）

- **回归**（regression）用于：
  - 预测——给定自变量的值，预测因变量的结果；
  - 解释——自变量的变化会导致因变量有多大程度的变化。

回归分析介绍

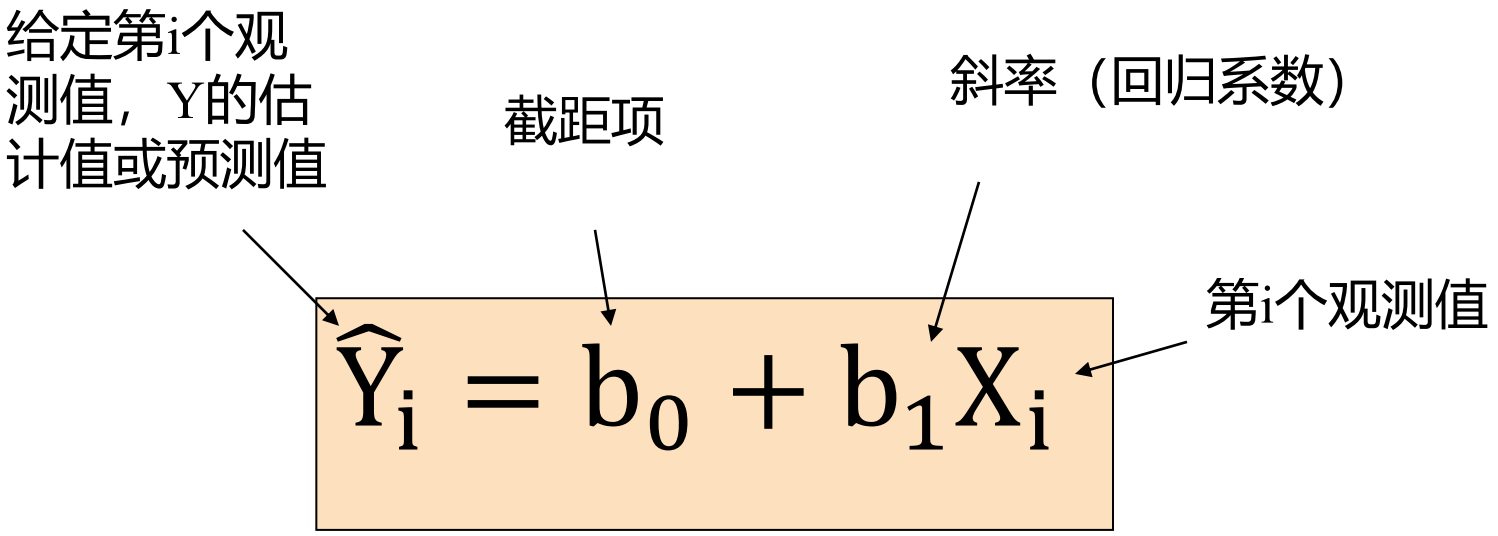
线性关系



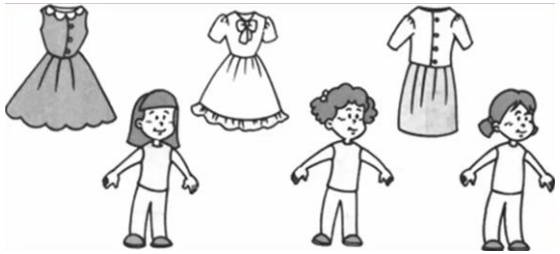
回归分析介绍

回归分析方程

（回归直线）



回归分析介绍



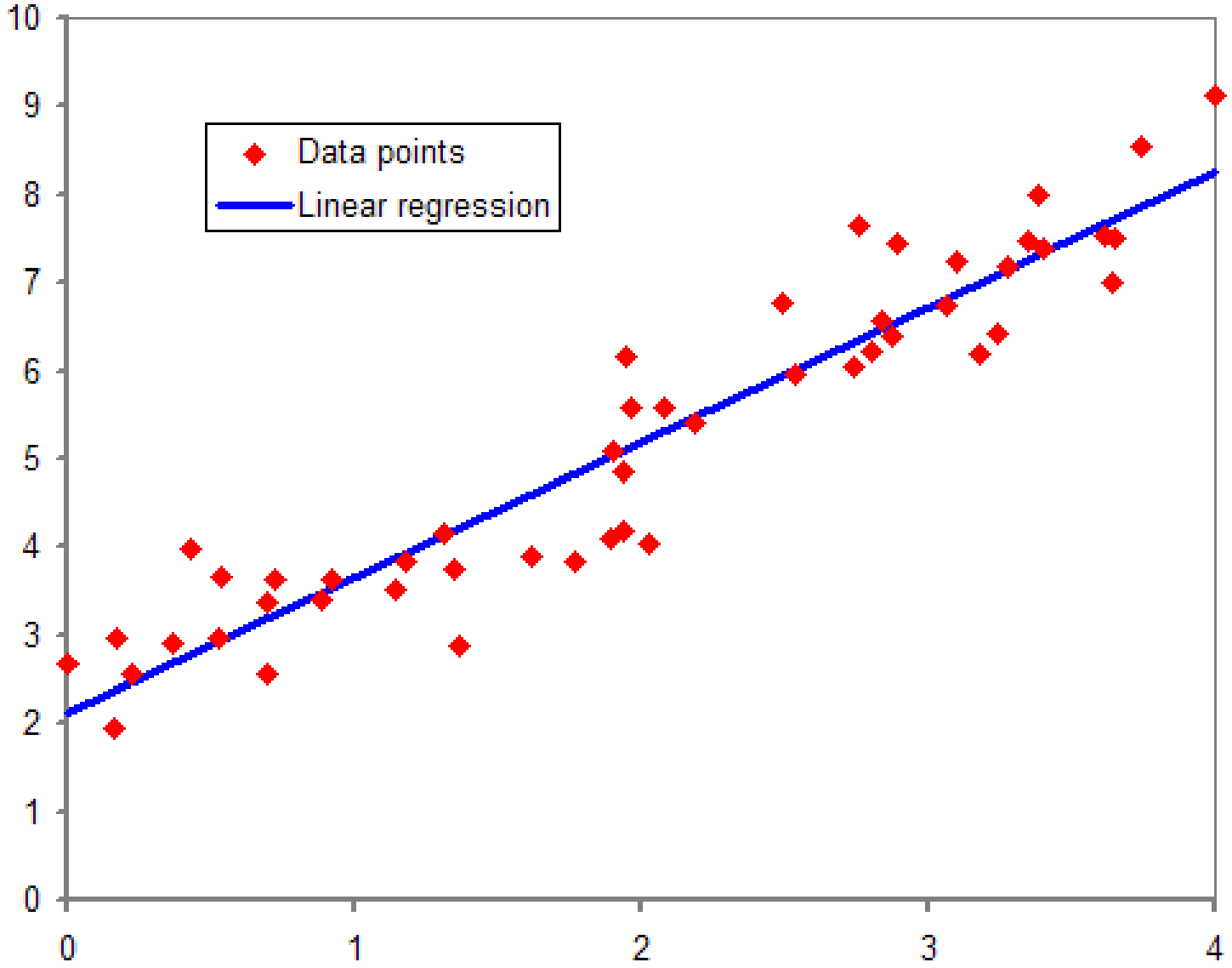
最小二乘法

- 我们通过最小二乘法（Ordinal least squares, OLS）找到**拟合**(fit)程度最好的回归直线

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

  - OLS回归

回归分析介绍



最小二乘法

$$\hat{Y}_i = b_0 + b_1 X_i$$

回归分析介绍

截距项和斜率的解释

- b_0（截距项）代表 X 为0时，Y的取值（一般没有什么实际用途）

- b_1（斜率）代表 X 每变化一个单位，Y 平均变化的单位

实例

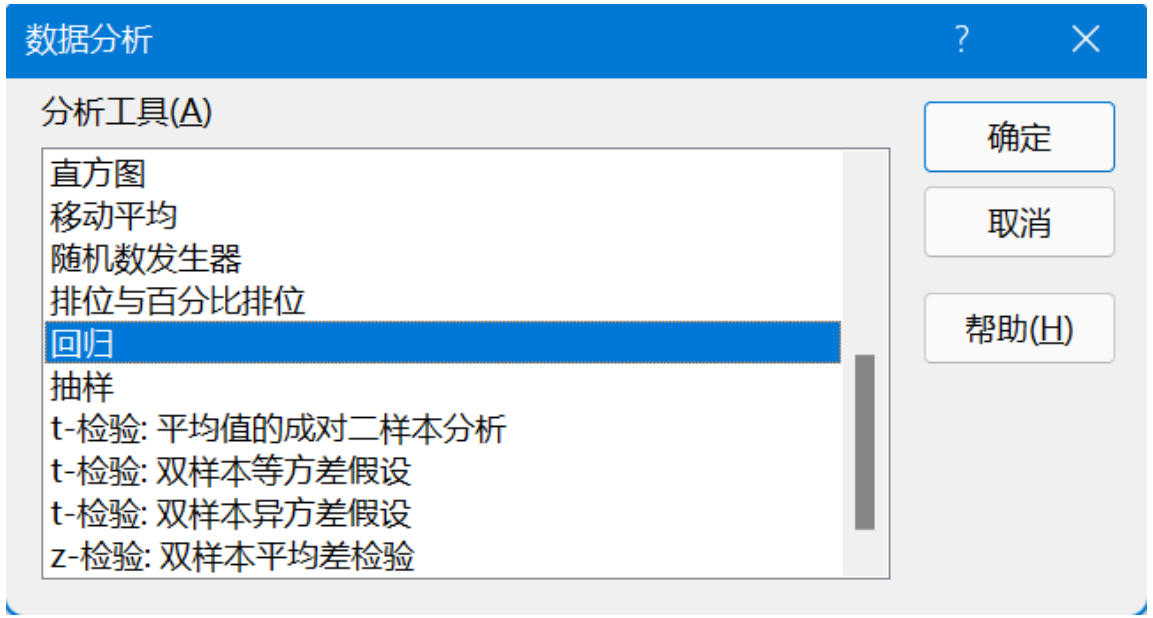
身高与体重

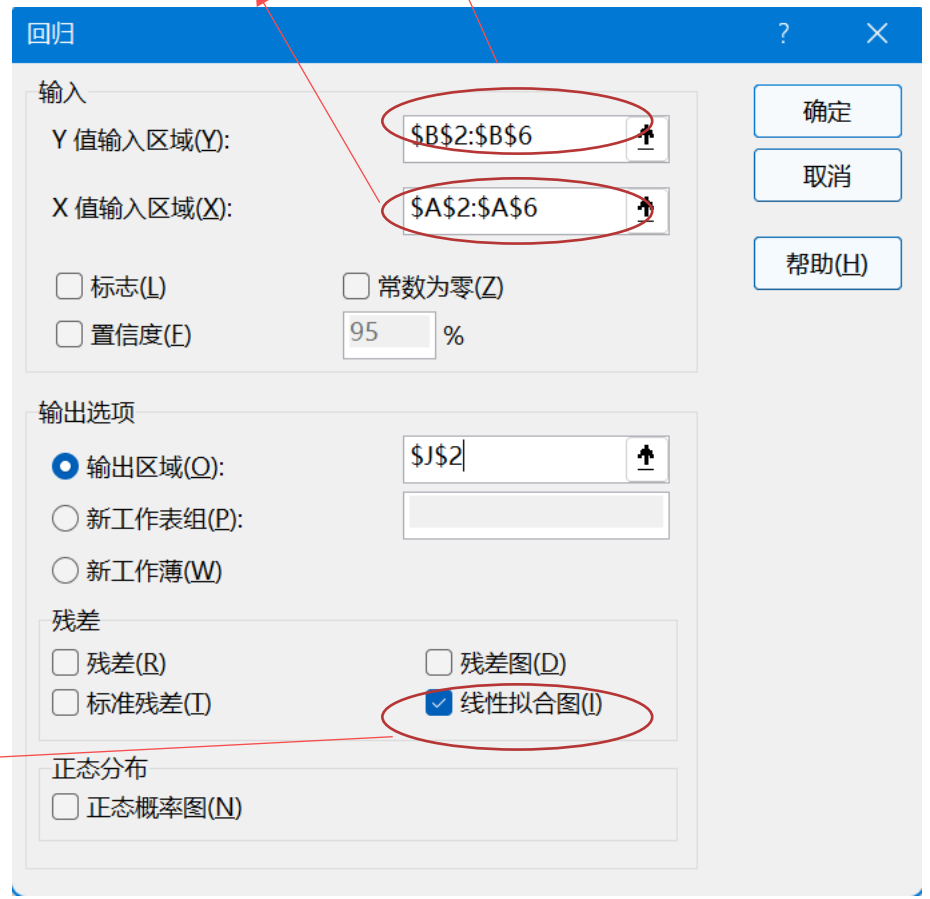
有五位学生，他们的身高分别是152、158、165、172、180厘米；体重分别是43、48、54、63、80公斤。以身高作为自变量、体重作为因变量做回归分析并做显著性检验。

身高	152	158	165	172	180
体重	43	48	54	63	80

实例

Excel操作


数据分析
分析工具(A)
直方图
移动平均
随机数发生器
排位与百分比排位
回归
抽样
t-检验: 平均值的成对二样本分析
t-检验: 双样本等方差假设
t-检验: 双样本异方差假设
z-检验: 双样本平均差检验
确定
取消
帮助(H)


因变量
自变量
回归
输入
Y 值输入区域(Y):
$B$2:$B$6
X 值输入区域(X):
$A$2:$A$6
标志(L)
常数为零(Z)
置信度(F)
95
%
确定
取消
帮助(H)
输出选项
输出区域(O):
$J$2
新工作表组(P):
新工作薄(W)
残差
残差(R)
残差图(D)
标准残差(T)
线性拟合图(I)
正态分布
正态概率图(N)
可勾选此项

实例

Excel结果

SUMMARY OUTPUT

回归统计	
Multiple R	0.976691735
R Square	0.953926746
Adjusted R Square	0.938568995
标准误差	3.611343571
观测值	5

方差分析

	df	SS	MS	F	Significance F
回归分析	1	810.0746	810.0746	62.1137	0.00425671
残差	3	39.12541	13.0418		
总计	4	849.2			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	-154.8070033	26.99936	-5.73373	0.010533	-240.73103	-68.882981	-240.73103	-68.882981
X Variable 1	1.284201954	0.162944	7.881224	0.004257	0.76563992	1.80276398	0.76563992	1.80276398

RESIDUAL OUTPUT

观测值	预测 Y	残差
1	40.39169381	2.608306
2	48.09690554	-0.09691
3	57.08631922	-3.08632
4	66.0757329	-3.07573
5	76.34934853	3.650651

回归方程：

$$\widehat{\text{体重}} = -154.81 + 1.28\ (\text{身高})$$

实例

截距项b_0的解释

$$\widehat{\text{体重}} = \boxed{-154.81} + 1.28\ (\text{身高})$$

- 截距项b_0是自变量X=0时Y的观测值
- 截距项b_0一般没有实际含义（不存在身高为0的情况）

实例

斜率b_1的解释

$$\widehat{\text{体重}} = -154.81 + \boxed{1.28}(\text{身高})$$

- 斜率b_1是自变量X每变化一个单位，因变量Y的变化情况
  - 例如，这里的b_1=1.28，意味着身高每提高1厘米，体重平均增加1.28公斤。

实例

用回归分析做预测

如果身高是175厘米，预测其体重：

$$\widehat{\text{体重}} = -154.81 + 1.28\ (\text{身高})$$

$$= -154.81 + 1.28\ (175)$$

$$= 69.19$$

判定系数：R^2

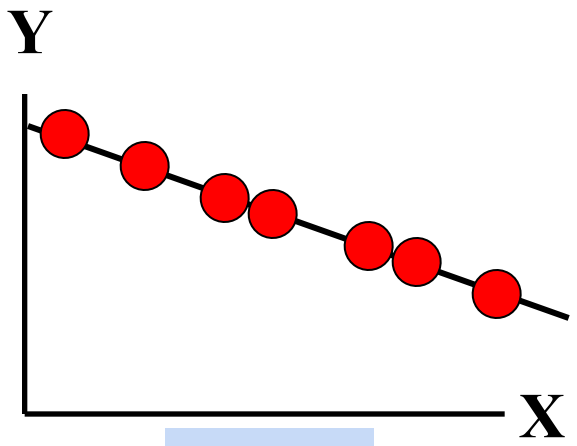
判定系数（coefficient of determination）是回归模型能解释因变量的变异的部分（反映回归直线的拟合程度）

判定系数的符号表示：R-square（R方），R^2

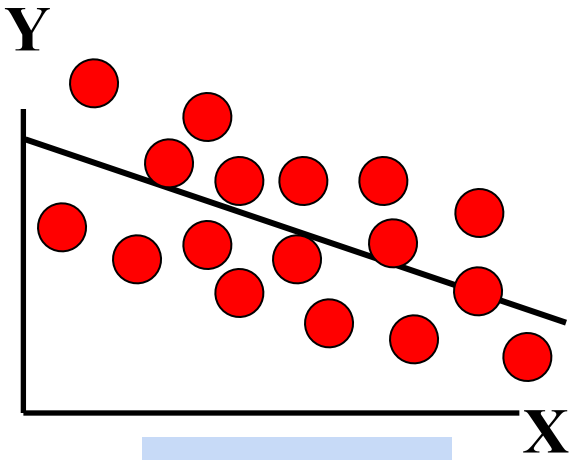
$$R^2 = \frac{SS_{回归}}{SS_{总}}$$

注意：$0 \le R^2 \le 1$

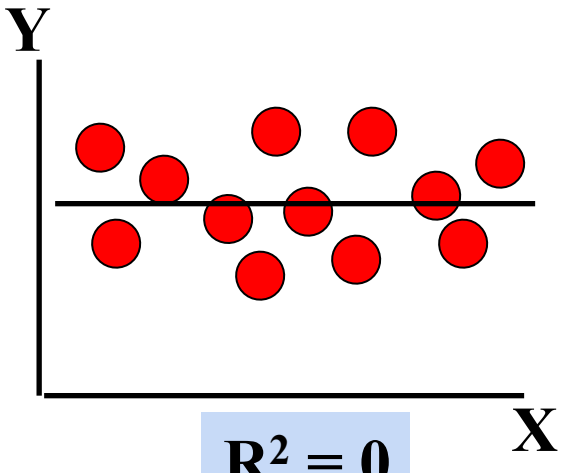
判定系数：R^2



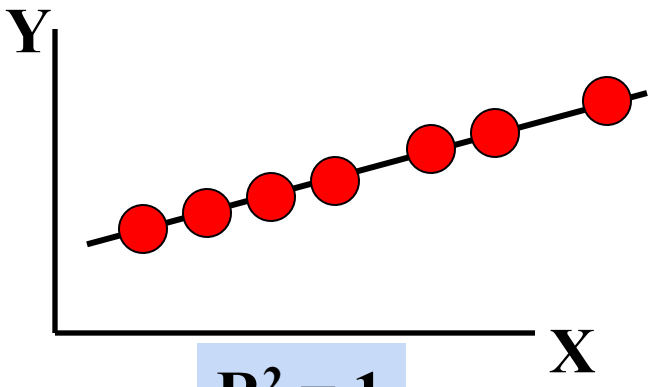
$R^2 = 1$



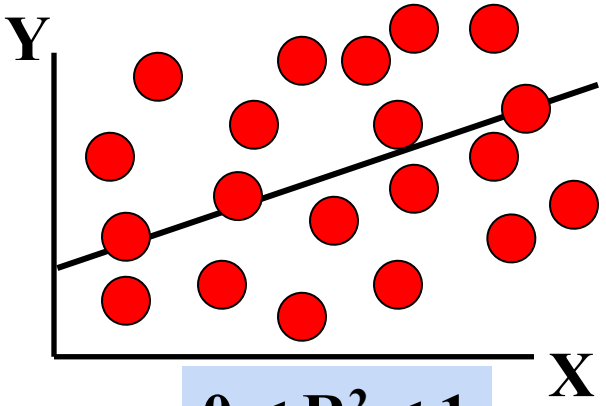
$0 < R^2 < 1$



$R^2 = 0$



$R^2 = 1$



$0 < R^2 < 1$

判定系数：R^2

$$R^2 = \frac{SS_{回归}}{SS_{总}}$$

$$R^2_{adj} = 1 - \left[(1 - R^2)\left(\frac{n-1}{n-k-1}\right)\right]$$

调整后的R^2：因为R^2是有偏估计量，它高估了总体中的R^2，所以需要调整后的R^2得到无偏估计量

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.977[a]	.954	.939	3.61134

a. 预测变量：(常量), 身高

回归统计	
Multiple R	0.976691735
R Square	0.953926746
Adjusted R Square	0.938568995
标准误差	3.611343571
观测值	5

结论：

- 身高可以解释体重95.4%的变异。
- 在考虑了样本量和自变量个数的前提下，身高可以解释体重93.9%的变异。

线性关系检验：F检验

- **零假设与备择假设：**

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (所有回归系数都等于零)

H_1: 不是所有回归系数都等于零

其中，k = 自变量个数

线性关系检验：F检验

检验统计量F:

$$F = \frac{MS_{回归}}{MS_{残差}}$$

其中：

$$MS_{回归} = \frac{SS_{回归}}{k}$$

$$MS_{残差} = \frac{SS_{残差}}{n-k-1}$$

k = 自变量个数
n = 样本量

F统计量服从$df_1 = k, df_2 = n - k - 1$的F分布

线性关系检验：F检验

$df_{回归} = k = 1$

(k为自变量个数)

$df_{残差} = n - k - 1 = 5 - 1 - 1 = 3$

(n为总样本量)

$df_{总} = n - 1 = 5 - 1 = 4$

$SS_{回归}$

$MS_{回归}$

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	810.075	1	810.075	62.114	.004[b]
	残差	39.125	3	13.042		
	总计	849.200	4			

a. 因变量：体重

b. 预测变量：(常量), 身高

p值

$SS_{残差}$

$SS_{总}$

$MS_{残差}$

$$F = \frac{MS_{回归}}{MS_{残差}} = \frac{SS_{回归}/df_{回归}}{SS_{残差}/df_{残差}} = \frac{810.075}{13.042} = 62.114$$

线性关系检验：F检验

$df_{回归} = k = 1$

(k为自变量个数)

$df_{残差} = n - k - 1 = 5 - 1 - 1 = 3$

(n为总样本量)

$SS_{回归}$

$MS_{回归}$

p值

方差分析					
	df	SS	MS	F	Significance F
回归分析	1	810.0745928	810.0745928	62.1137	0.004256714
残差	3	39.12540717	13.04180239		
总计	4	849.2			

$df_{残差} = n - k - 1 = 5 - 1 - 1 = 3$

$SS_{残差}$

$SS_{总}$

$MS_{残差}$

$df_{总} = n - 1 = 5 - 1 = 4$

$$F = \frac{MS_{回归}}{MS_{残差}} = \frac{SS_{回归}/df_{回归}}{SS_{残差}/df_{残差}}$$

$$= \frac{810.075}{13.042} = 62.1137$$

线性关系检验：F检验

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\alpha = 0.05$

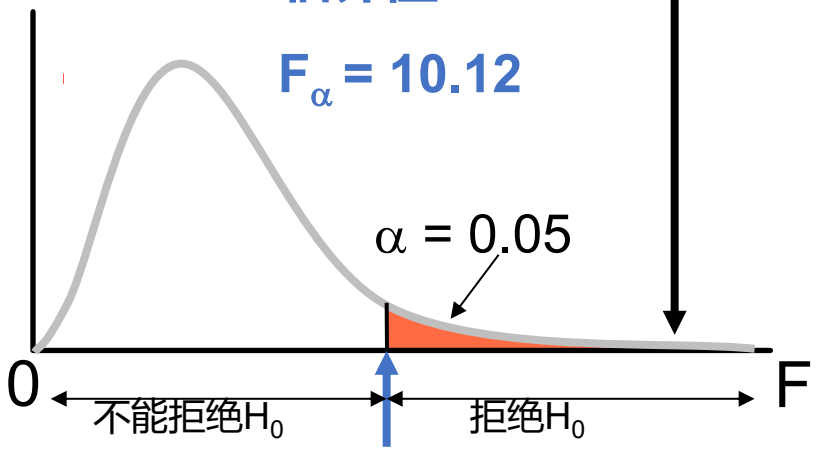
$df_1 = 1 \quad df_2 = 3$

检验统计量：

$$F = \frac{MS_{回归}}{MS_{残差}} = 62.12$$

临界值：

$F_\alpha = 10.12$



结论：拒绝 H_0

回归模型成立（具有统计学显著意义）。/至少有一个自变量显著影响因变量。

回归系数检验：t检验

- **零假设与备择假设：**

$H_0: \beta_1 = 0$ (存在线性关系)

$H_1: \beta_1 \neq 0$ (不存在线性关系)

t检验统计量：

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$d.f. = n - 2$$

其中：

b_1 = 回归系数（斜率）

β_1 = 检验的斜率

S_{b1} = 斜率的标准误

注：S_{b_1}在此不做讨论

回归系数检验：t检验

回归方程式：

$$\widehat{\text{体重}} = -154.81 + 1.28\ (\text{身高})$$

回归模型的斜率是1.28，那么身高和体重存在关系吗？

回归系数检验：t检验

斜率（回归系数）b_1

斜率b_1的标准误S_{b1}

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{1.284 - 0}{0.163} = 7.881$$

系数[a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	-154.807	26.999		-5.734	.011
	身高	1.284	.163	.977	7.881	.004

a. 因变量：体重

p值

标准化斜率（回归系数）

(主要目的是消除自变量X 单位大小的影响)

回归系数检验：t检验

斜率（回归系数）b_1

斜率b_1的标准误S_{b1}

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{1.284 - 0}{0.163} = 7.881$$

	Coefficient	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	-154.81	26.9993629	-5.733728008	0.010533	-240.731026	-68.8829805	-240.73103	-68.882980
X Variable	1.2842	0.162944471	7.881224484	0.004257	0.765639925	1.802763984	0.76563992	1.802763

p值

斜率（回归系数）b_1 95%的置信区间

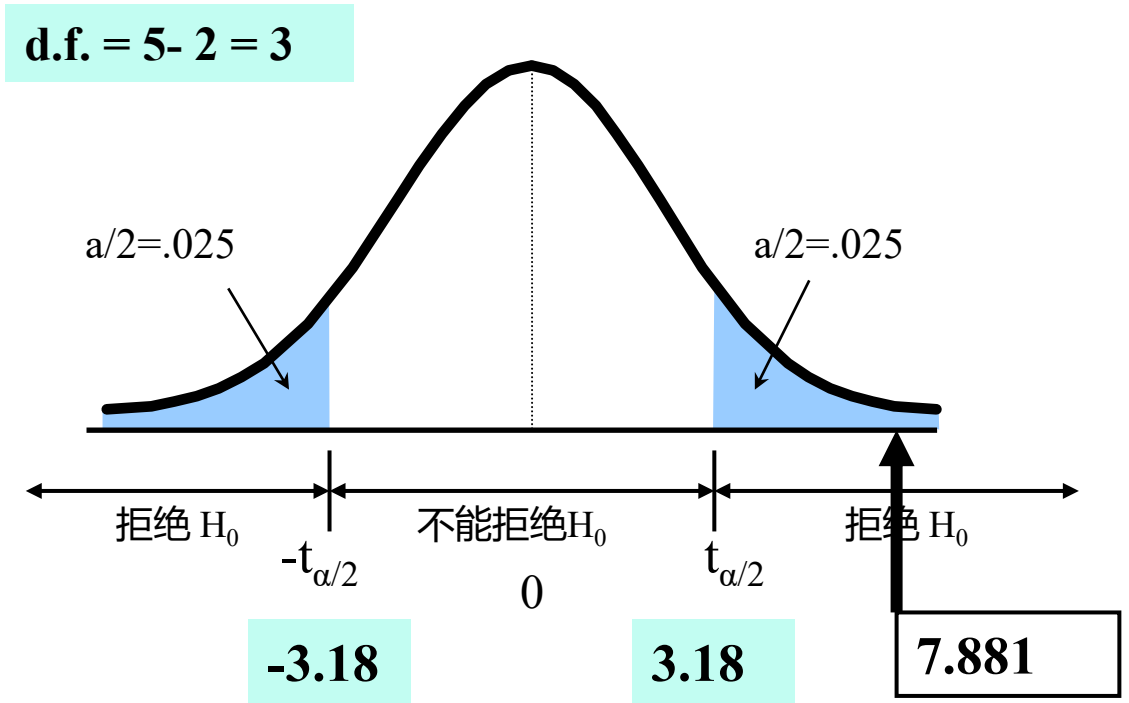
斜率（回归系数）b_1 95%的置信区间

回归系数检验：t检验

检验统计量：**t = 7.881**

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$



结论：拒绝 H_0

在α=0.05的水平下，身高对体重有显著影响。

习题

在样本大小为50人的情况下，回归模型$\hat{Y}=12.40+1.30\ X_1$的方差分析表如下，请填空**并计算判定系数R^2**

	Sum of Squares 平方和	df	Mean Square	F
Regression 回归	195.604	???	???	???
Error 误差	???	???	???	
Total 总和	547.636	???		


谢谢！