《新媒体数据运营与分析》

Excel (9): 统计分析-回归分析（一）

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
- Excel操作
- 实例
- 判定系数：R^2

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。）

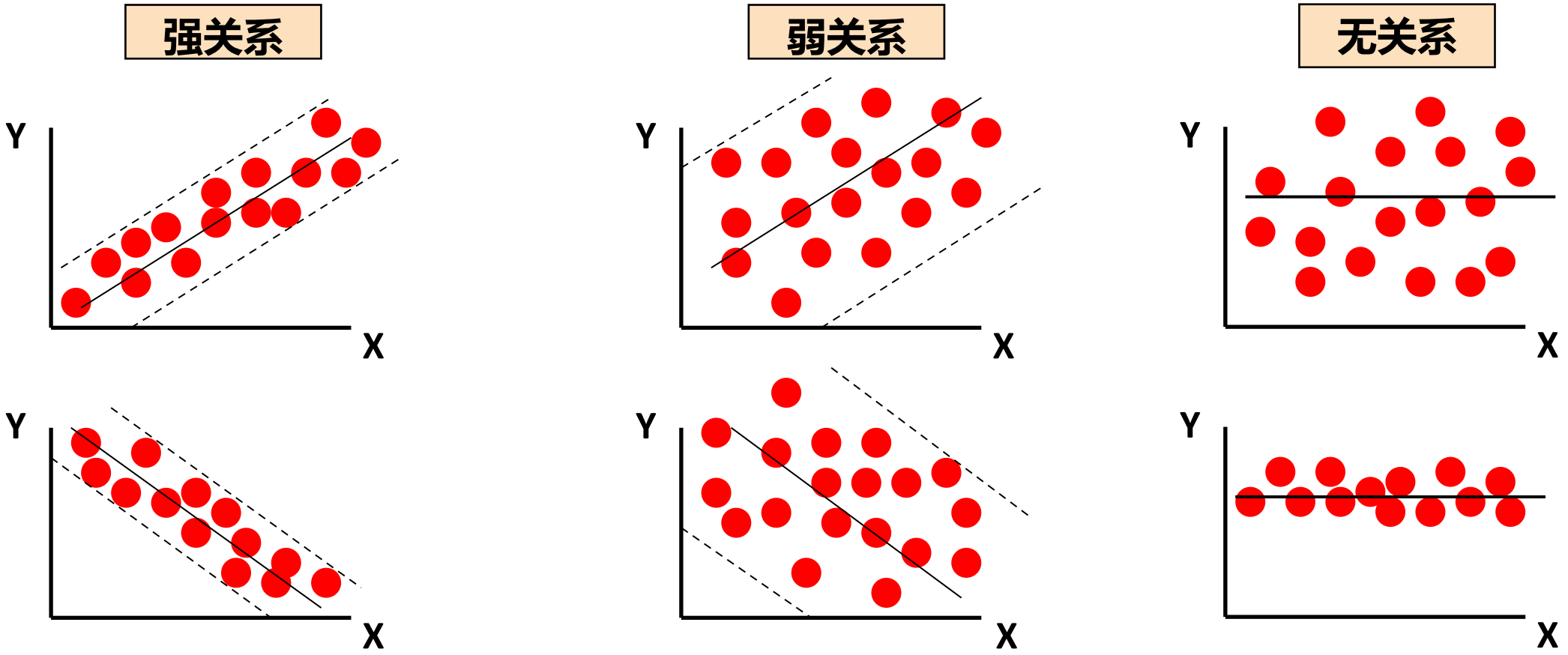
回归分析介绍

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）

给定第i个观测值，Y的估计值或预测值

截距项

斜率（回归系数）

第i个观测值

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

回归分析介绍

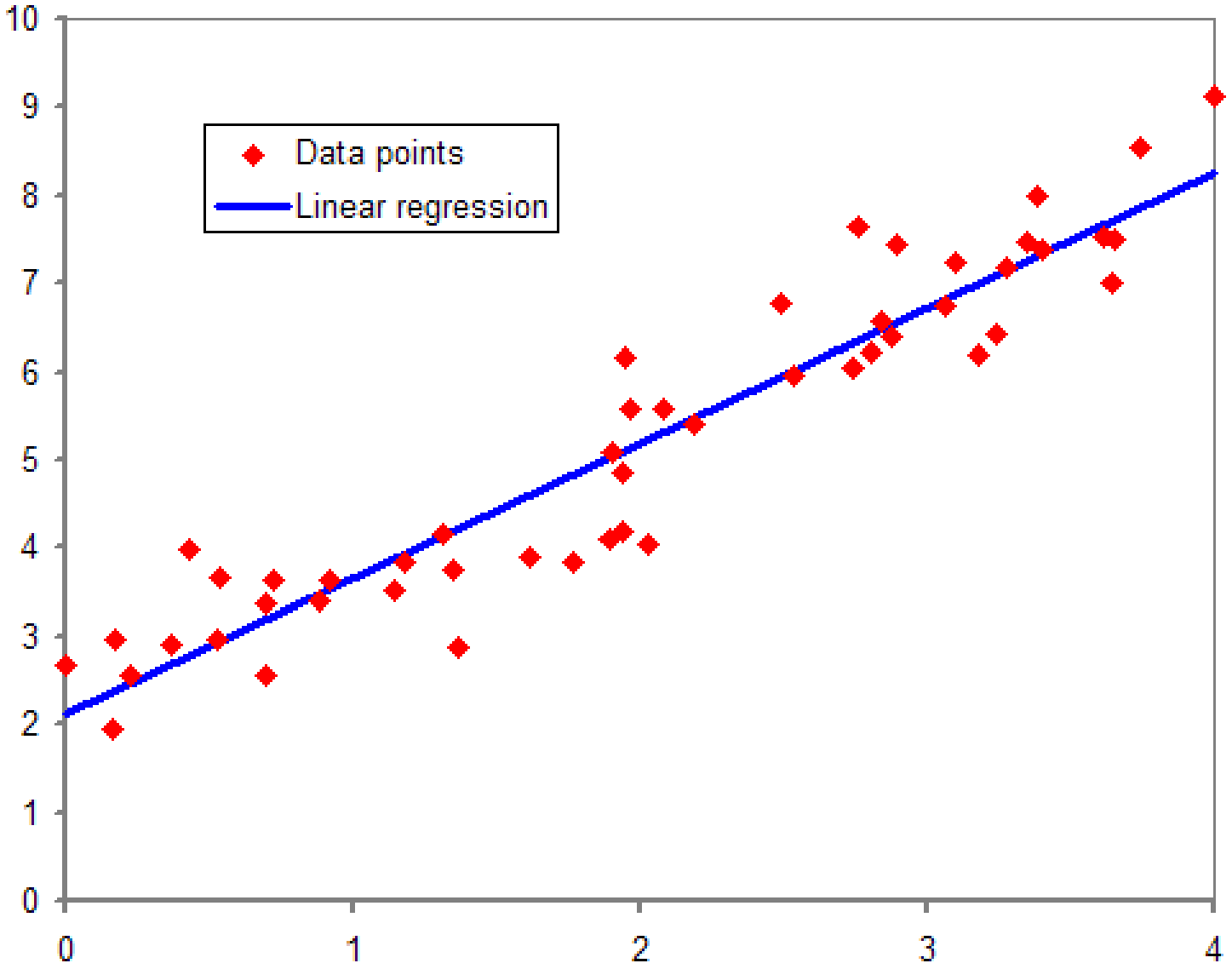
最小二乘法

- 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

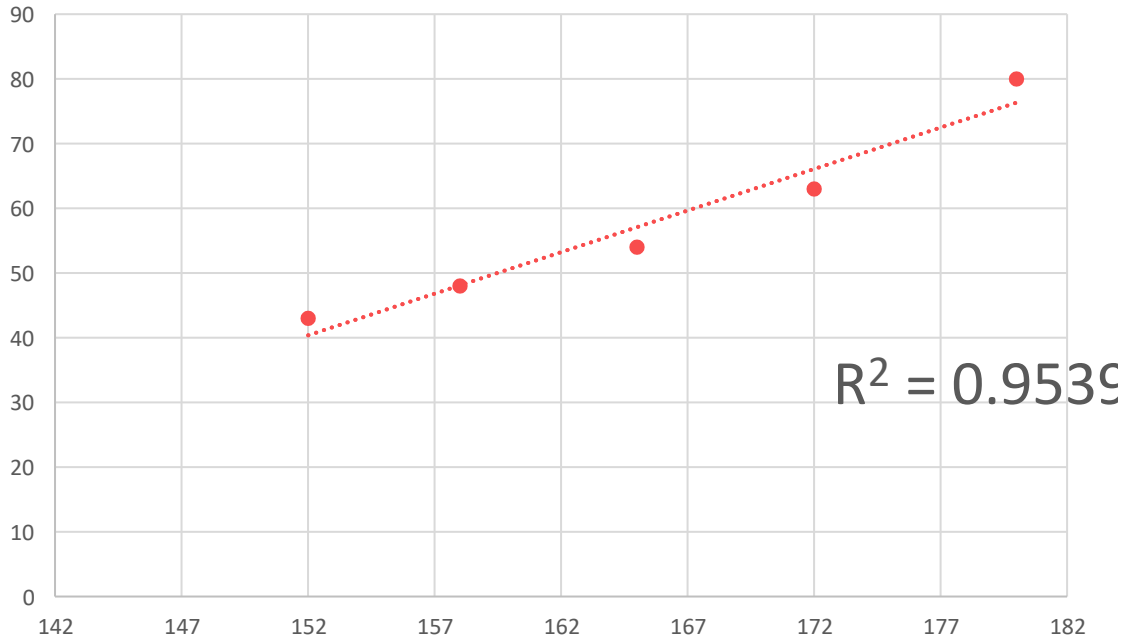
截距项和斜率的解释

- 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

excel操作

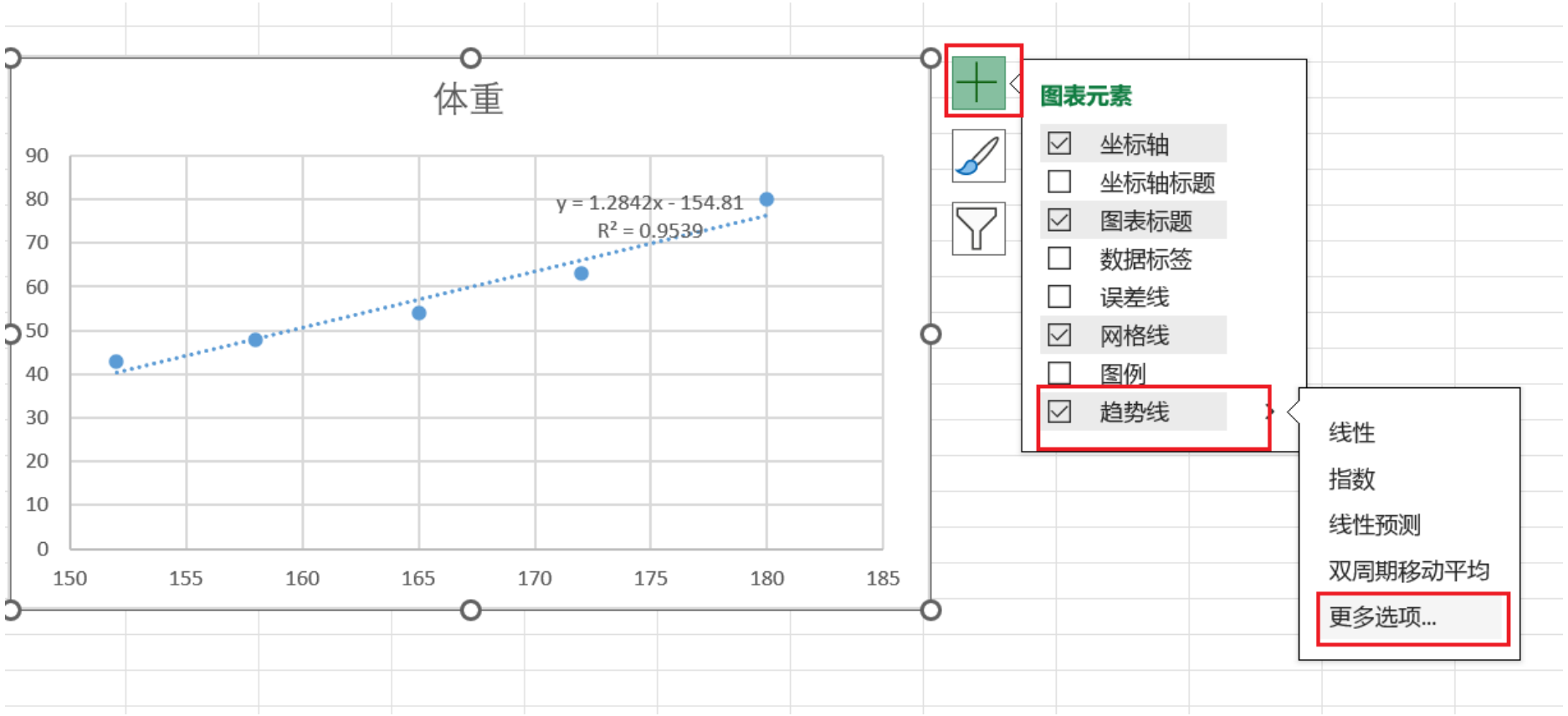
散点图

- 为了可视化两变量的回归关系，我们可以绘制散点图并**添加回归直线（趋势线）**，另外在excel还可以添加**回归方程（公式）和R^2值。**



excel操作

散点图


体重
y = 1.2842x - 154.81
R² = 0.9539
90
80
70
60
50
40
30
20
10
0
150
155
160
165
170
175
180
185
图表元素
坐标轴
坐标轴标题
图表标题
数据标签
误差线
网格线
图例
趋势线
线性
指数
线性预测
双周期移动平均
更多选项...

趋势线选项
趋势线选项
指数(X)
线性(L)
对数(O)
多项式(P)
阶数(D)
2
乘幂(W)
移动平均(M)
周期(E)
2
趋势线名称
自动(A)
线性 (体重)
自定义(C)
趋势预测
前推(F)
0.0
周期
后推(B)
0.0
周期
设置截距(S)
0.0
显示公式(E)
显示 R 平方值(R)

excel操作

Excel函数：计算回归系数（Slope）

- 函数: **SLOPE**(known_ys, known_xs)
- 参数:
  - known_ys: 因变量（Y值）的数据范围。
  - known_xs: 自变量（X值）的数据范围

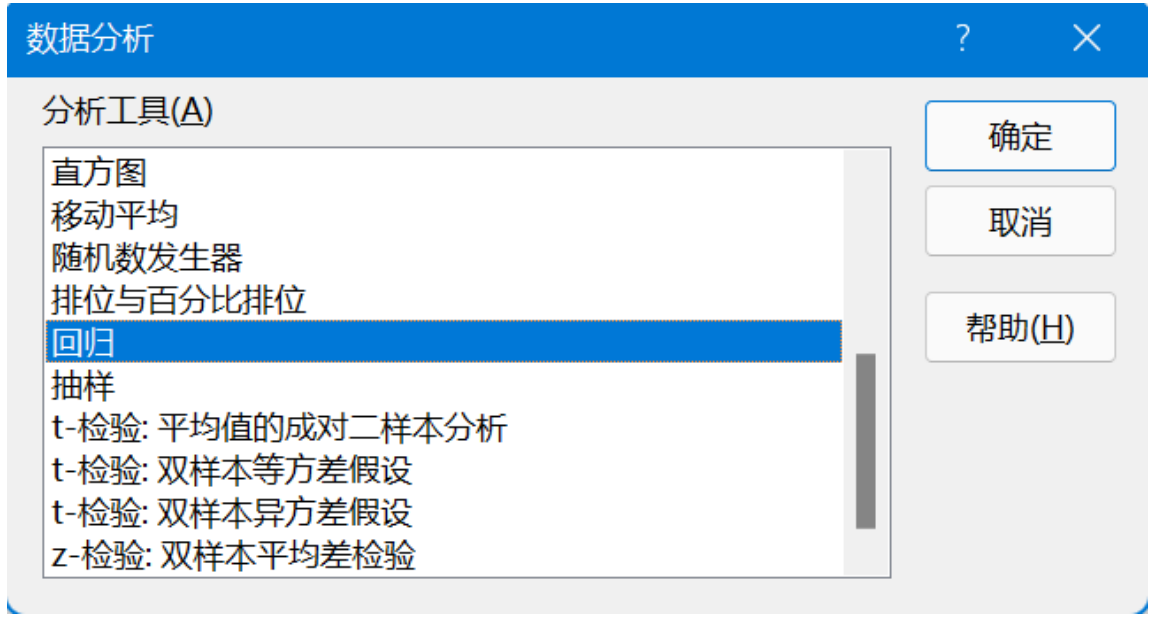
excel操作

Excel函数：计算截距（Intercept）

- 函数: INTERCEPT(known_ys, known_xs)
- 参数:
  - known_ys: 因变量（Y值）的数据范围。
  - known_xs: 自变量（X值）的数据范围。

excel操作

数据分析工具



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。

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

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：

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

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

$$= 69.19$$

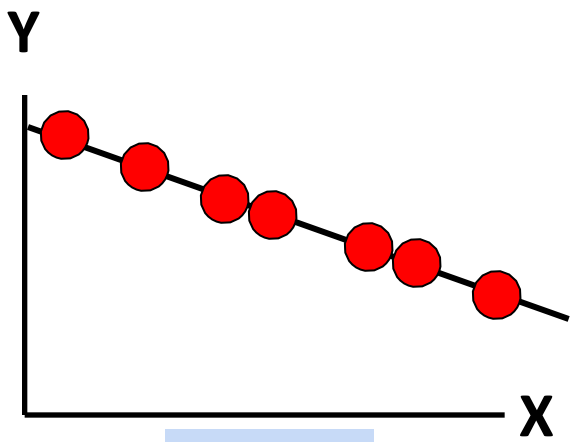
判定系数：R^2

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）

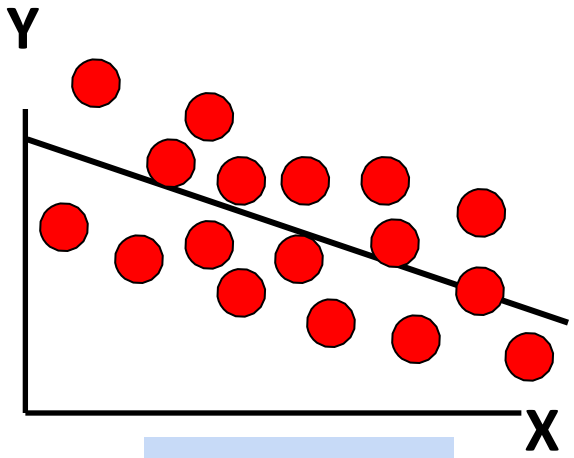
判定系数的符号表示：R-square（R方），R^2

注意: $\boxed{0 \le R^2 \le 1}$

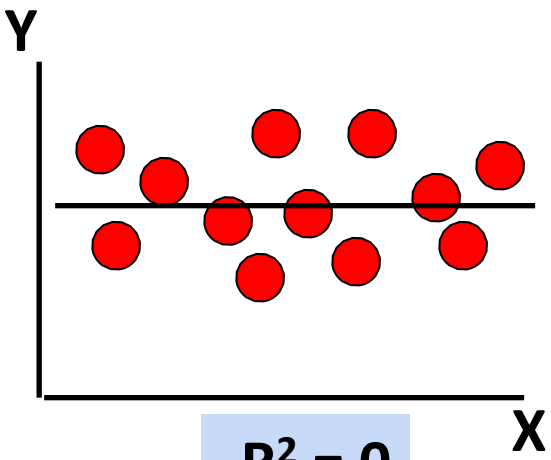
判定系数：R^2



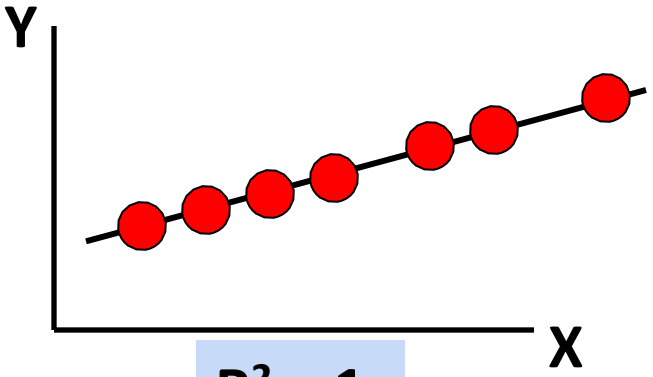
$R^2 = 1$



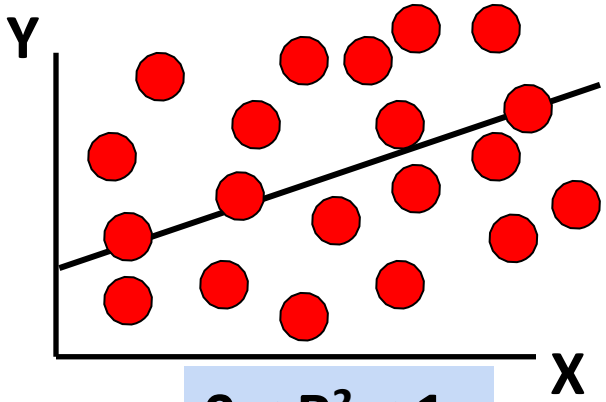
$0 < R^2 < 1$



$R^2 = 0$



$R^2 = 1$



$0 < R^2 < 1$

判定系数：R^2

Excel函数：计算R^2（判定系数）

- 函数: RSQ (known_ys, known_xs)
- 参数:
  - known_ys: 因变量（Y值）的数据范围。
  - known_xs: 自变量（X值）的数据范围。

判定系数：R^2

调整后的R^2: 样本量和自变量个数会影响R^2，为了消除样本量和自变量个数的影响

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。

参考资料

- Lizongzhang的个人空间-合集 ·Excel 数据分析实战

谢谢！