线性回归
Linear Regression

一个数值变量的特征

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

One-sample t-test

$$Var(X) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

Chi-square test

两个变量的关系

一个数值变量与一个分类变量的关系	Two-sample t-test One-way ANOVA
一个数值变量与两个分类变量的关系	Two-way ANOVA
两个数值变量的关系	?

协方差 Covariance

方差 vs 协方差

- 方差: 刻画一个数值变量偏离其中心的程度
- 协方差: 刻画两个数值变量共同变化的程度

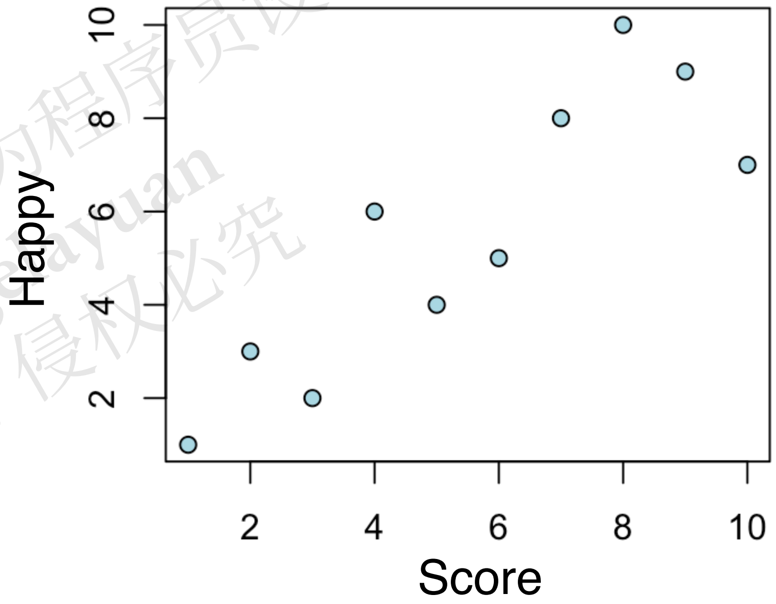
$$Var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

协方差

Score	Happy
1	1
2	3
3	2
4	6
5	4
6	5
7	8
8	10
9	9
10	7

- 散点图: 方向、形状、强度、极端值



协方差

Score	Happy
1	1
2	3
3	2
4	6
5	4
6	5
7	8
8	10
9	9
10	7

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$\longrightarrow$ $\bar{X} = 5.5, \bar{Y} = 5.5$

$\longrightarrow$ $$Cov(X,Y) = \frac{(1-5.5)\times(1-5.5) + \cdots + (10-5.5)\times(7-5.5)}{10-1}$$

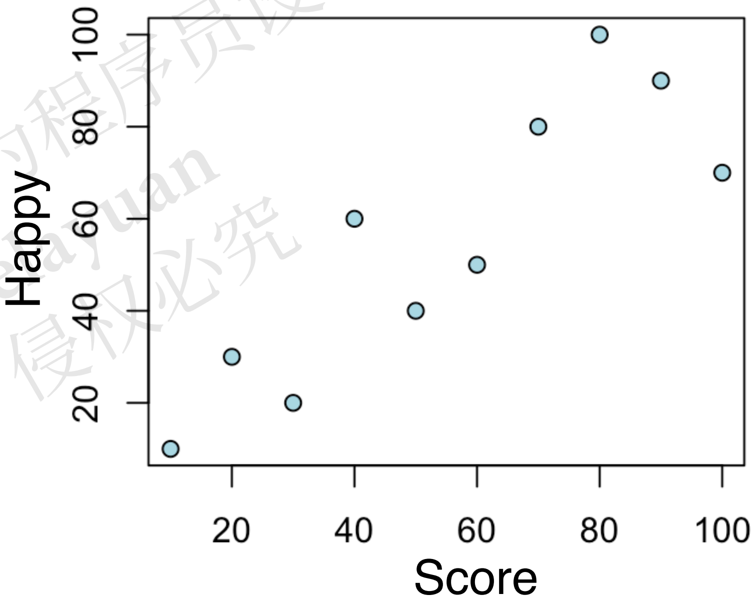
$$= \frac{71.5}{9} = 7.94$$

强？弱？正？负？

协方差

Score	Happy
10	10
20	30
30	20
40	60
50	40
60	50
70	80
80	100
90	90
100	70

- 散点图: 方向、形状、强度、极端值



协方差

Score	Happy
10	10
20	30
30	20
40	60
50	40
60	50
70	80
80	100
90	90
100	70

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$\longrightarrow \bar{X} = 55, \bar{Y} = 55$$

$$\longrightarrow Cov(X,Y) = \frac{(10-55)\times(10-55) + \cdots + (100-55)\times(70-55)}{10-1}$$

$$= \frac{7150}{9} = 794$$

强？ 弱？ 正？ 负？ $794 > 7.94$???

协方差 vs 相关系数

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

S_X 变量X的标准差

S_Y 变量Y的标准差

相关 Correlation

相关系数

Score	Happy
1	1
2	3
3	2
4	6
5	4
6	5
7	8
8	10
9	9
10	7

$$Cov(X,Y) = 7.94$$

$$S_X = 3.03, S_Y = 3.03$$

$$r = \frac{Cov(X,Y)}{S_X S_Y} = \frac{7.94}{3.03\times3.03} = 0.86$$

$$Cov(X,Y) = 794$$

$$S_X = 30.3, S_Y = 30.3$$

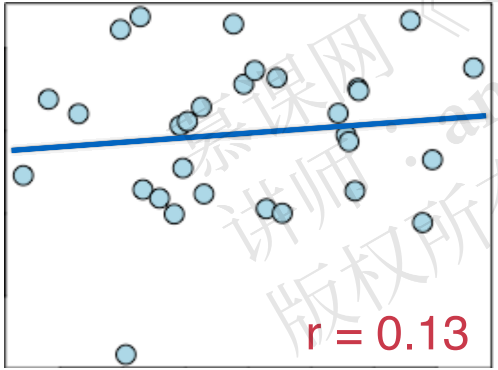
$$r = \frac{Cov(X,Y)}{S_X S_Y} = \frac{794}{30.3\times30.3} = 0.86$$

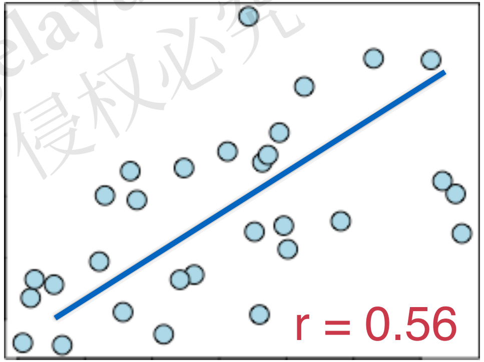
Score	Happy
10	10
20	30
30	20
40	60
50	40
60	50
70	80
80	100
90	90
100	70

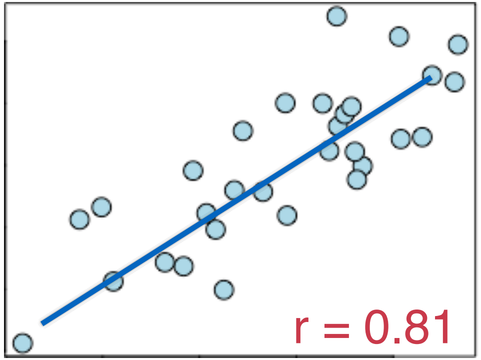
相关系数

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

- 描述两个数值变量线性关系的强度和方向
- 取值范围 $-1 \le r \le 1$
- r的绝对值大小代表线性关系的强弱



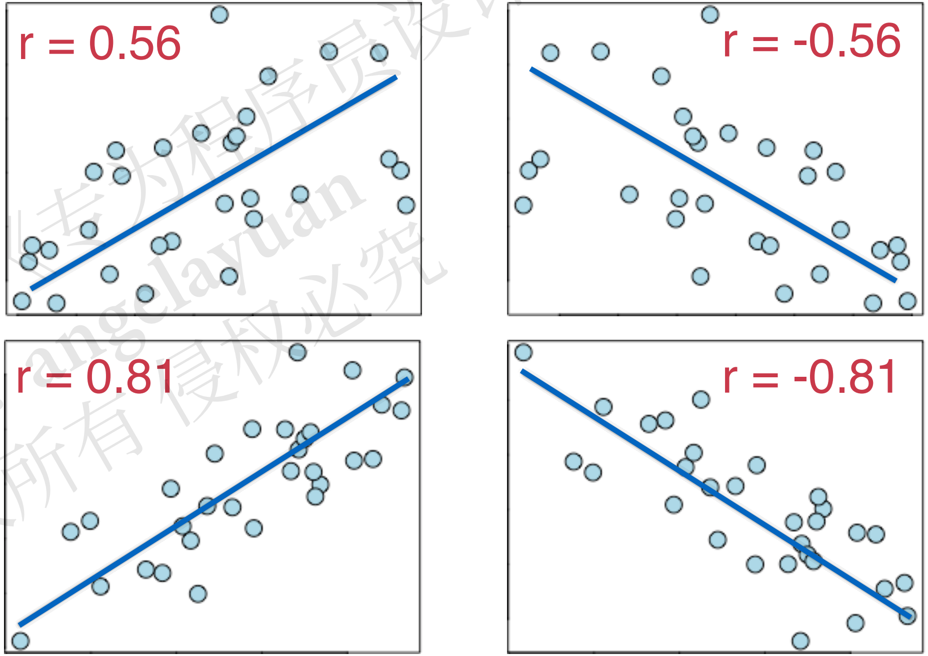




相关系数

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

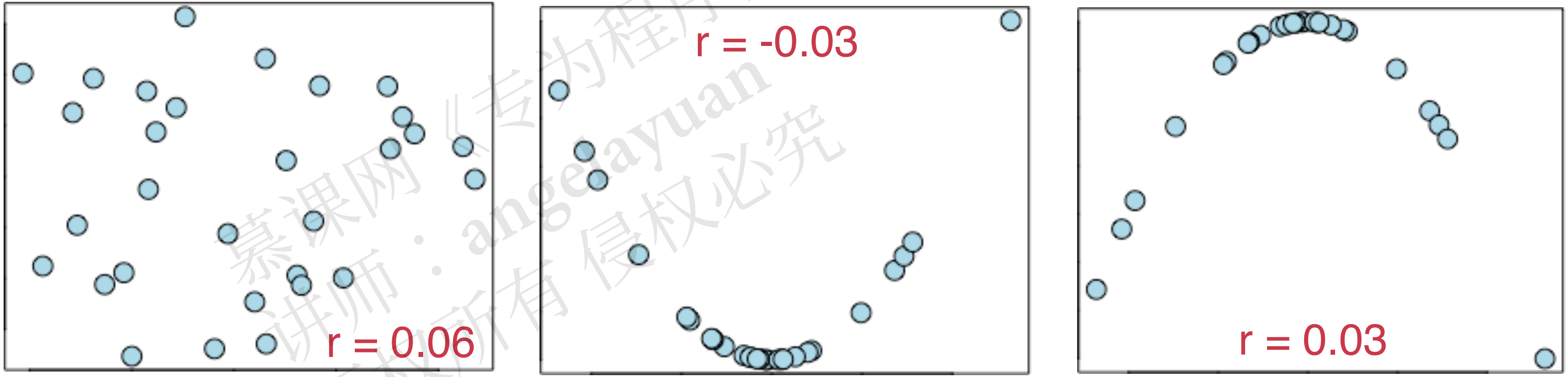
- r的符号代表线性关系的方向



相关系数

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

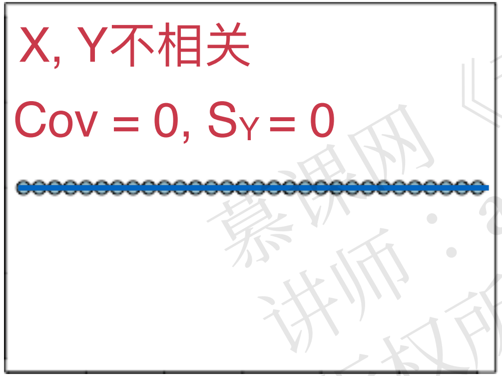
- r = 0: 没有线性关系 ≠ 没有关系

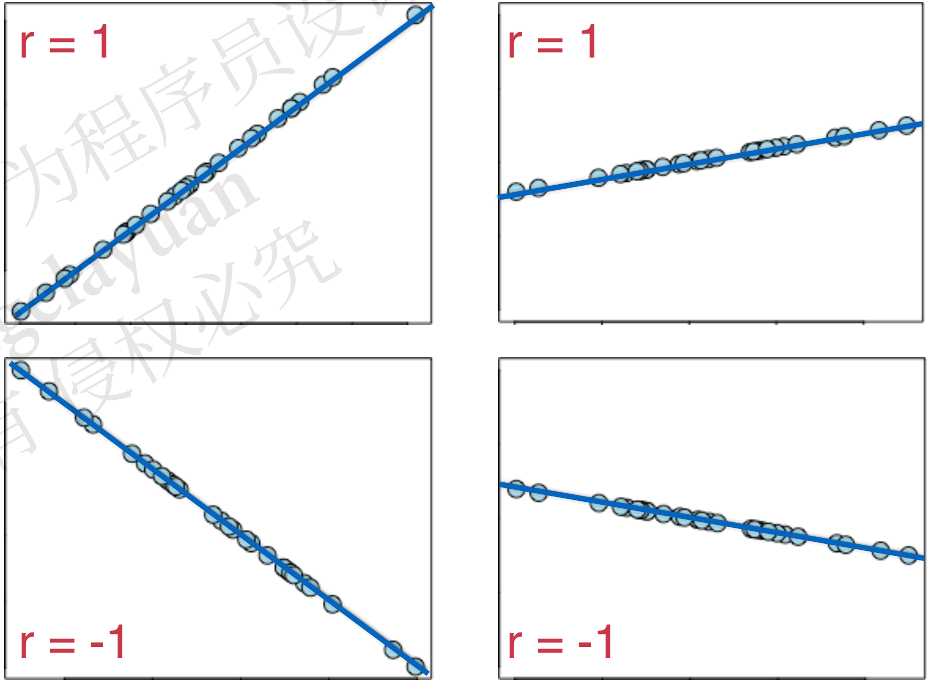


相关系数

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

- r = 1: 完美线性正相关; r = -1: 完美线性负相关

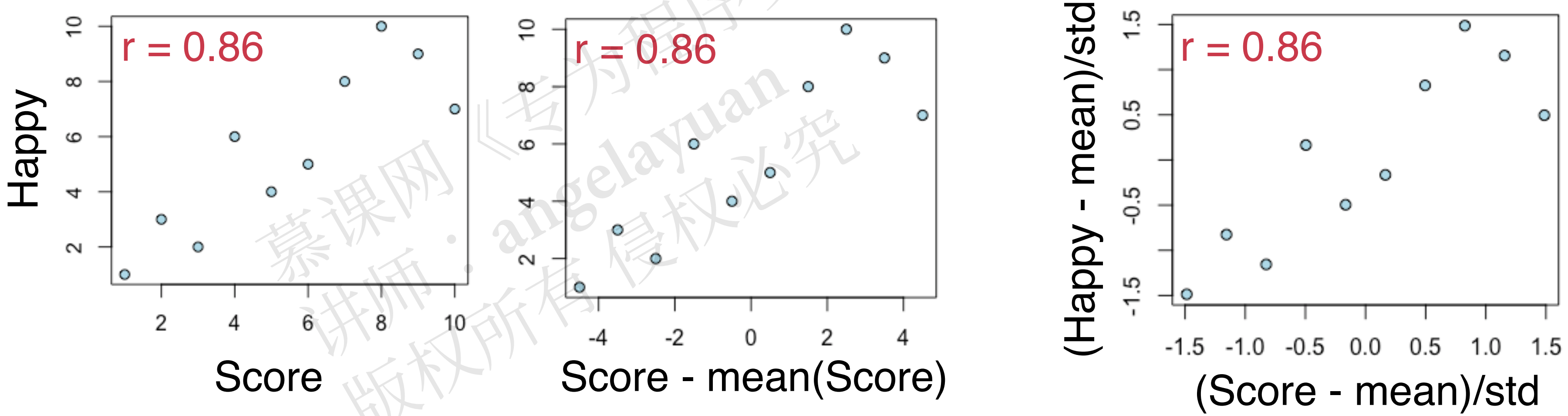




相关系数

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

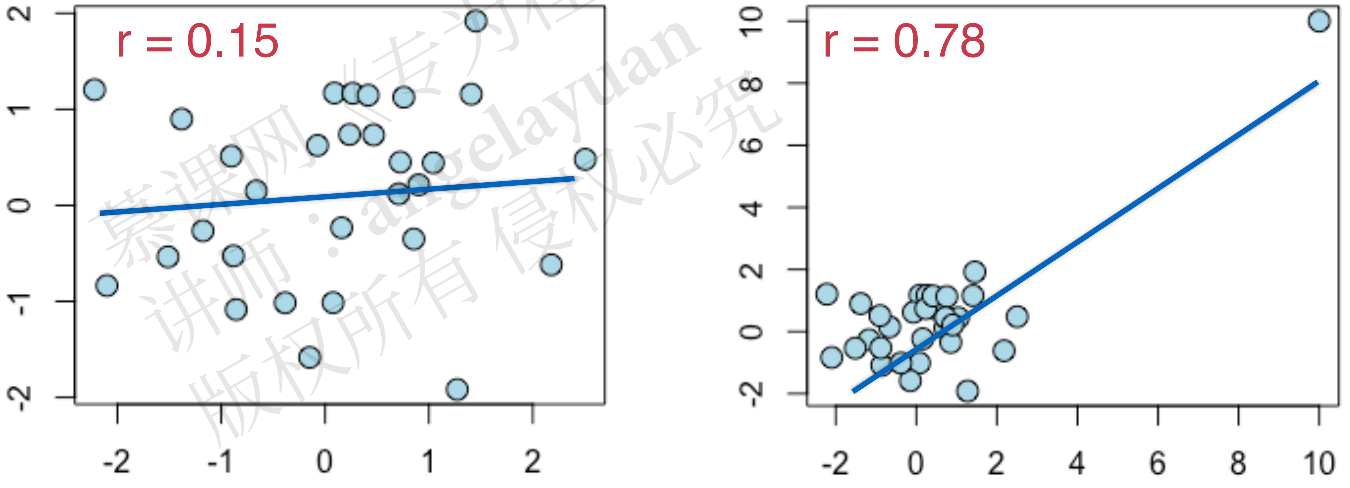
- r没有单位, 不受变量平移伸缩的影响



相关系数

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

- X,Y的相关系数 = Y, X的相关系数
- r受极端值影响大



相关的假设检验

X(Score),Y(Happy)的相关系数$r = 0.86$ → X与Y有较强的线性关系

X与Y没有线性关系的时候，是否可能观察到相关系数0.86?

- $H_0: r = 0$　　　P(当前结果或更极端结果 | H_0为真)
- $H_A: r \neq 0$

- H_0,H_A中的r是参数
- 题干中的相关系数0.86是样本的函数的一个观察值

相关的假设检验

X(Score),Y(Happy)的相关系数r = 0.86 → X与Y有较强的线性关系

X与Y没有线性关系的时候，是否可能观察到相关系数0.86?

- H_0: r = 0
- H_A: r ≠ 0

P(当前结果或更极端结果 | H_0为真)

$$\frac{r}{\sqrt{1-r^2}/\sqrt{n-2}} \sim t(n-2)$$

$$\frac{0.86}{\sqrt{1-0.86^2}/\sqrt{10-2}} = 4.76$$

→ p = 0.0007 x 2 ≈ 0.001

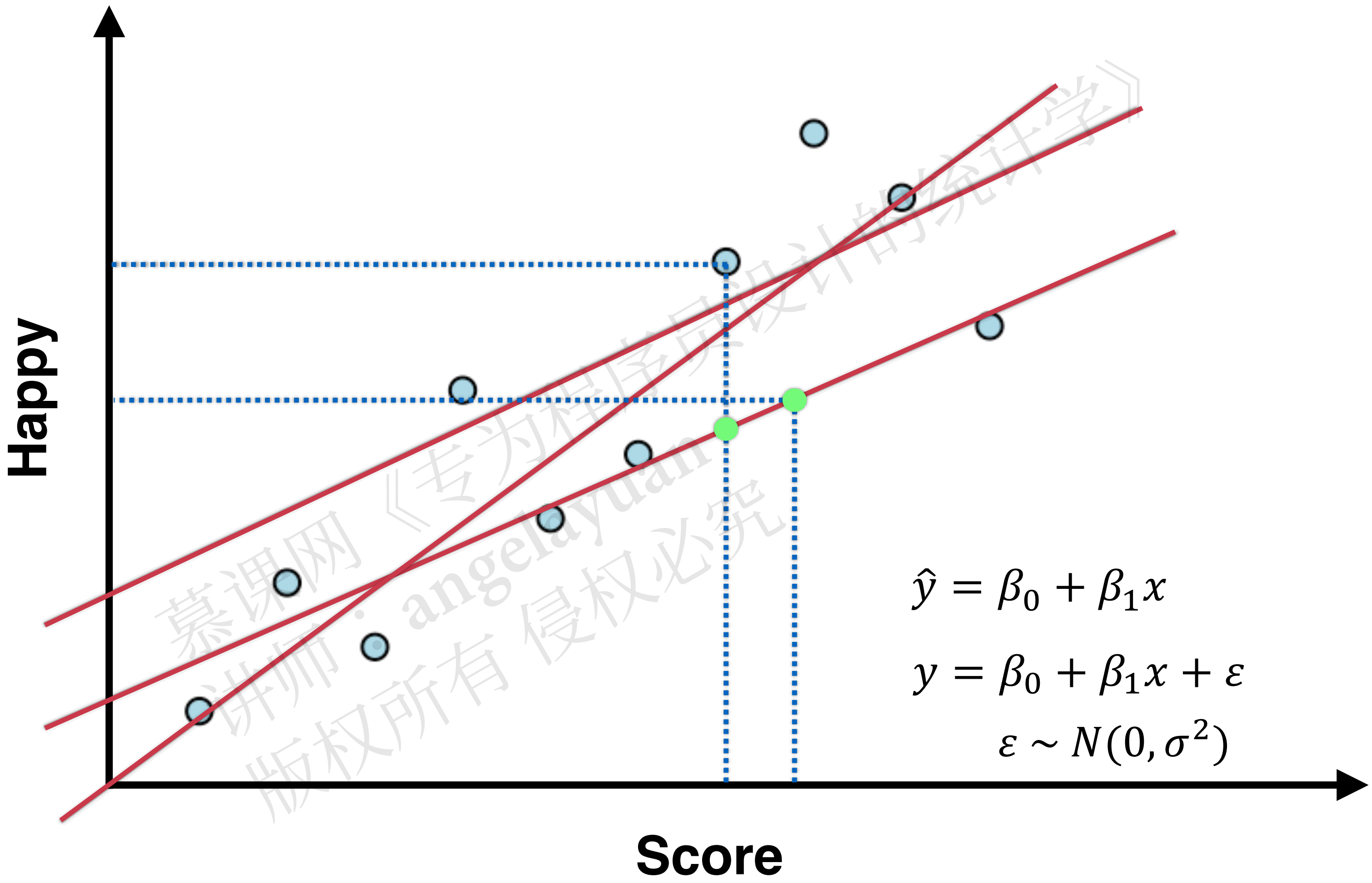
编程理解协方差和相关

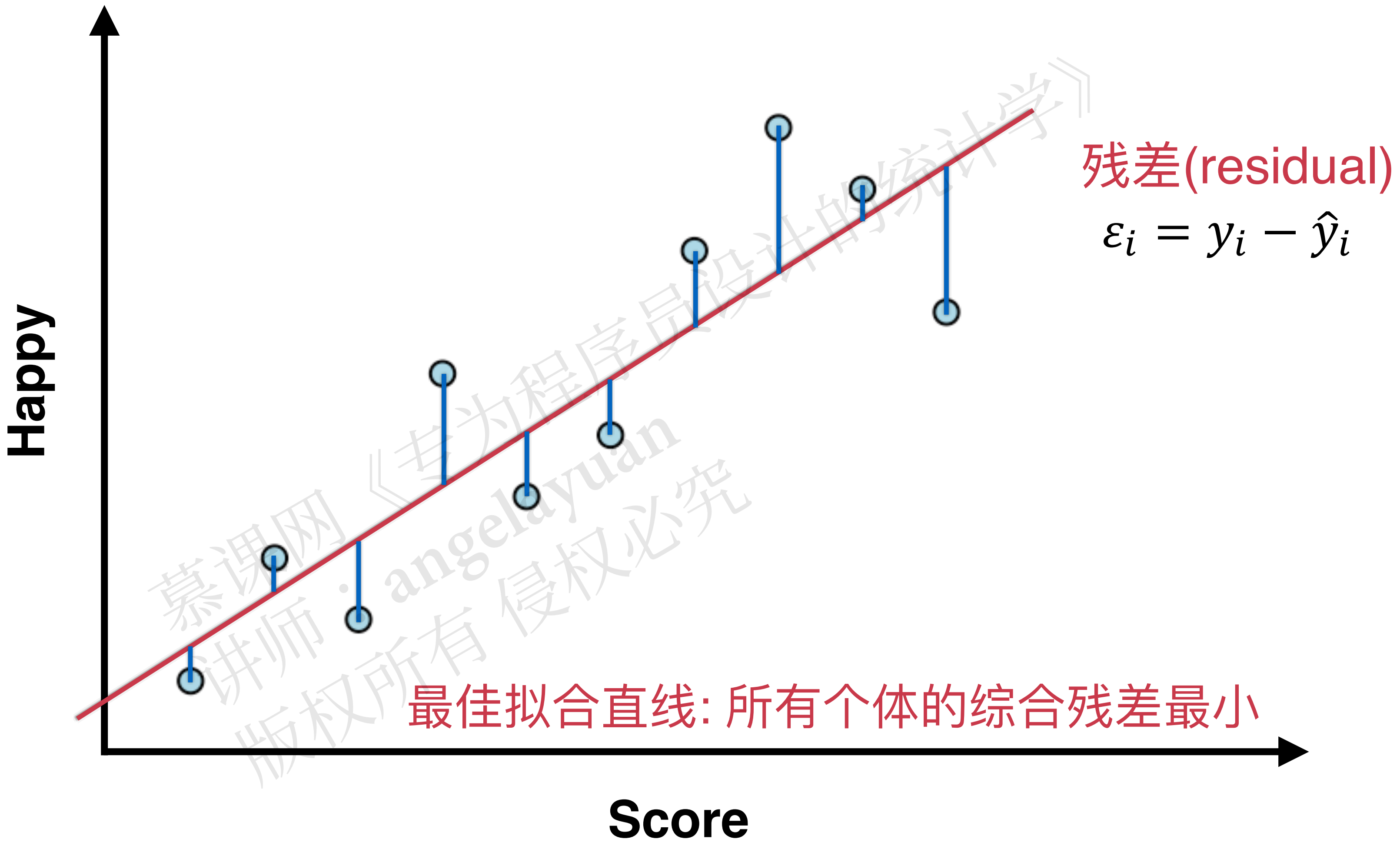
一元线性回归

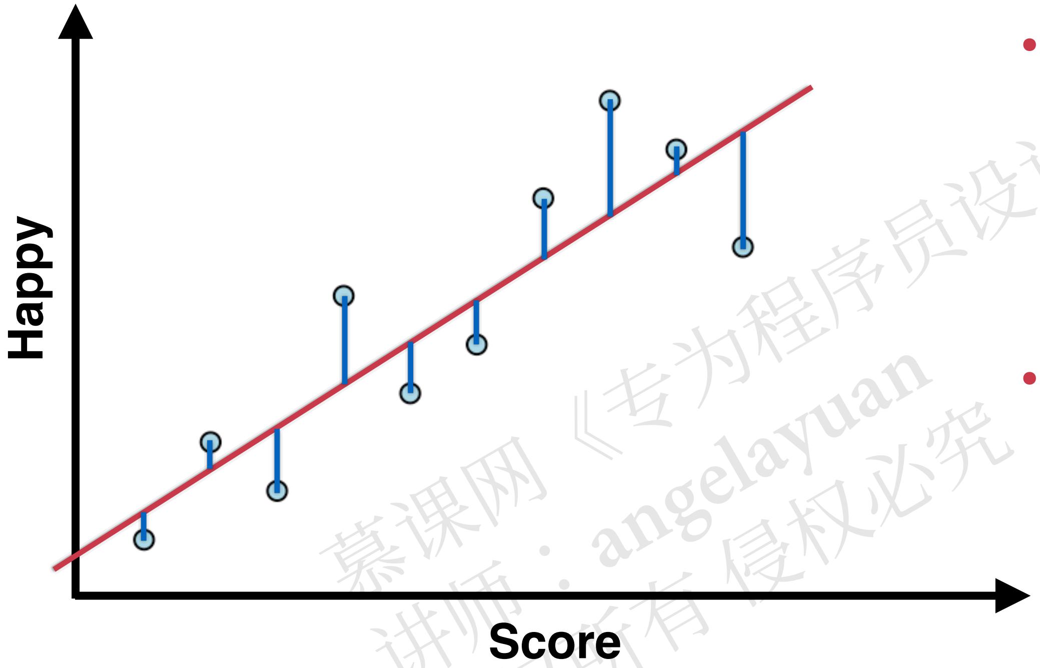
Simple Linear Regression

一元线性回归

- 相关: 考察两个数值变量之间的线性关系及其统计显著性
- 一元线性回归
  - 两个变量之间的线性关系及其统计显著性
  - 两个变量: 一个因变量, 一个自变量
  - 给定自变量的值, 预测因变量的值


Happy
Score
$\hat{y} = \beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 x + \varepsilon$
$\varepsilon \sim N(0, \sigma^2)$


Happy
Score
残差(residual)
$\varepsilon_i = y_i - \hat{y}_i$
最佳拟合直线: 所有个体的综合残差最小



- 最小化所有残差的绝对值的和

$$\sum_{i=1}^{n}|\varepsilon_i| = \sum_{i=1}^{n}|y_i - \hat{y}_i| \quad ✕$$

- 最小化所有残差的平方的和

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

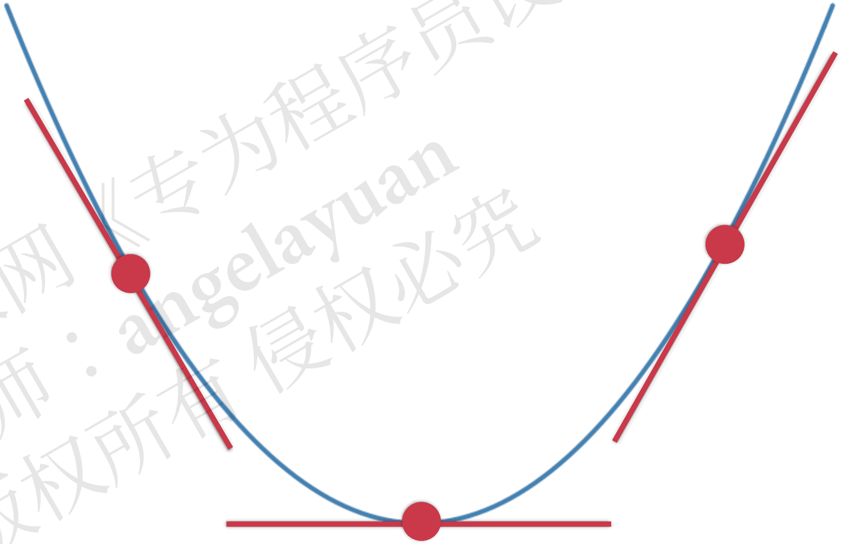
- 最小化所有残差的平方的和

$$\sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$ 典型的最小二乘法问题

目标: 求β_0,β_1, 使得 $J(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ 最小

方法: 取J关于β_0,β_1的偏导数, 并令它们等于0, 求解未知数

方法: 取J关于β_0,β_1的偏导数, 并令它们等于0, 求解未知数



方法: 取J关于β_0,β_1的偏导数, 并令它们等于0, 求解未知数

$$\frac{\partial J}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \longrightarrow n\beta_0 + (\sum_{i=1}^{n} x_i)\beta_1 = \sum_{i=1}^{n} y_i$$

$$\frac{\partial J}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \longrightarrow (\sum_{i=1}^{n} x_i)\beta_0 + (\sum_{i=1}^{n} {x_i}^2)\beta_1 = \sum_{i=1}^{n} x_i y_i$$

正规方程组

方法: 取J关于β_0,β_1的偏导数, 并令它们等于0, 求解未知数

$$n\beta_0 + \left(\sum_{i=1}^{n} x_i\right)\beta_1 = \sum_{i=1}^{n} y_i$$

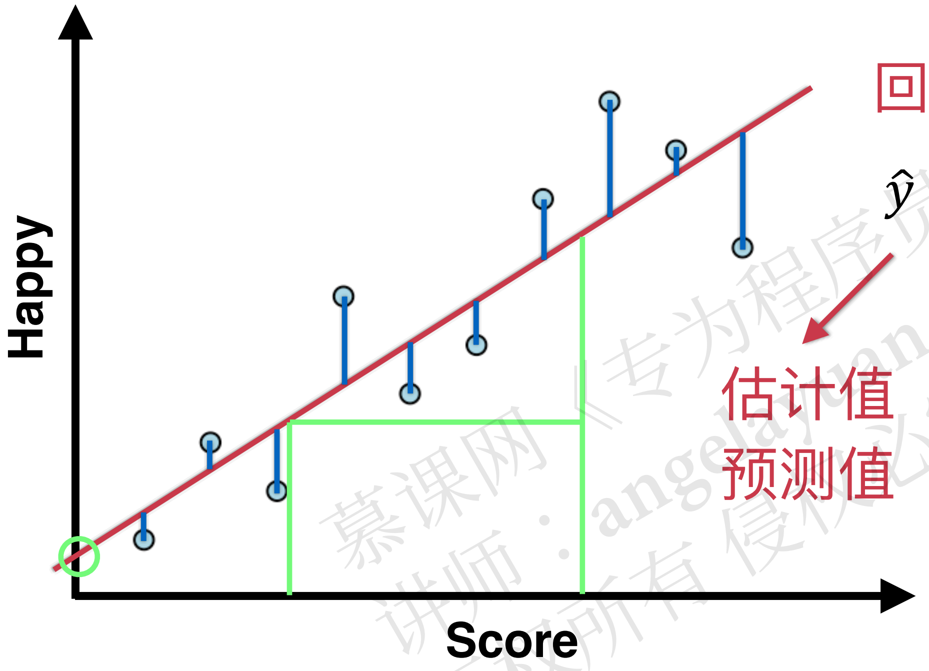
$$\left(\sum_{i=1}^{n} x_i\right)\beta_0 + \left(\sum_{i=1}^{n} {x_i}^2\right)\beta_1 = \sum_{i=1}^{n} x_i y_i$$

正规方程组

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$



回归直线

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

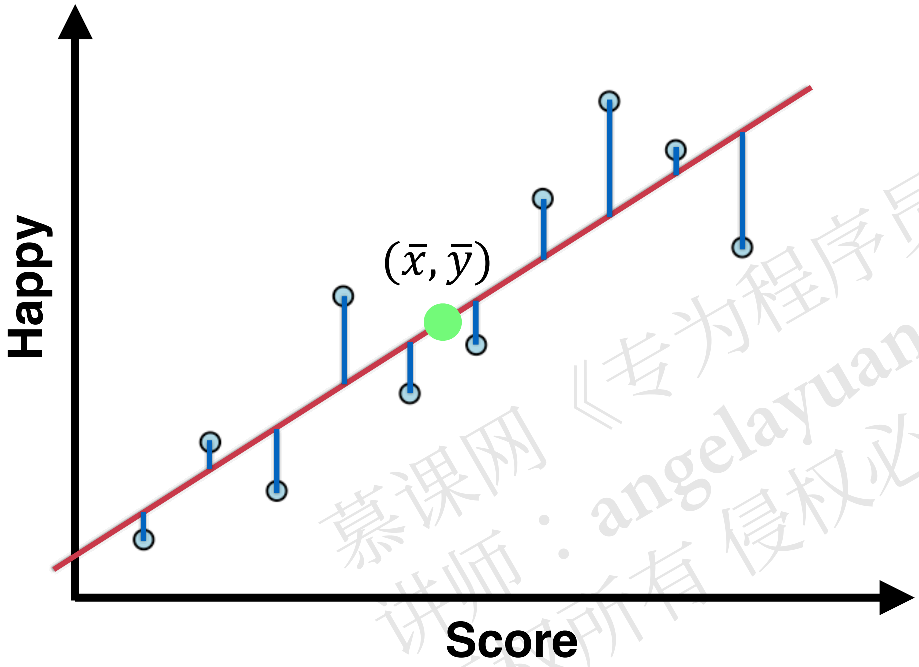
Y关于X的经验回归方程

估计值
预测值

截距 斜率

回归系数

自变量/解释变量

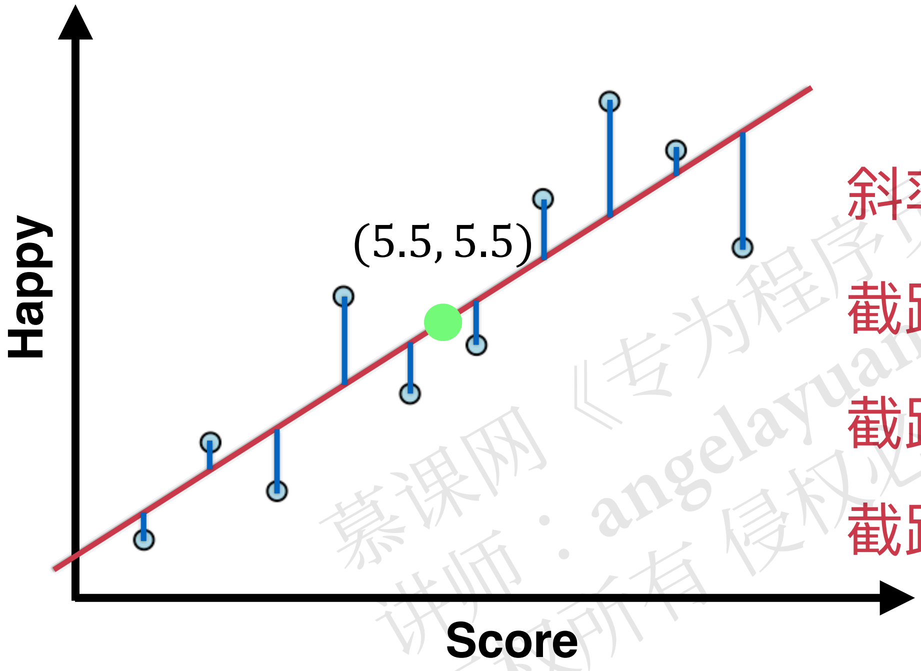


$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x \longleftarrow \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

$$\downarrow$$

$$\hat{y} = \bar{y} + \widehat{\beta_1}(x - \bar{x})$$

回归直线一定经过点(x均值, y均值)



$$\widehat{happy} = 0.73 + 0.87score$$

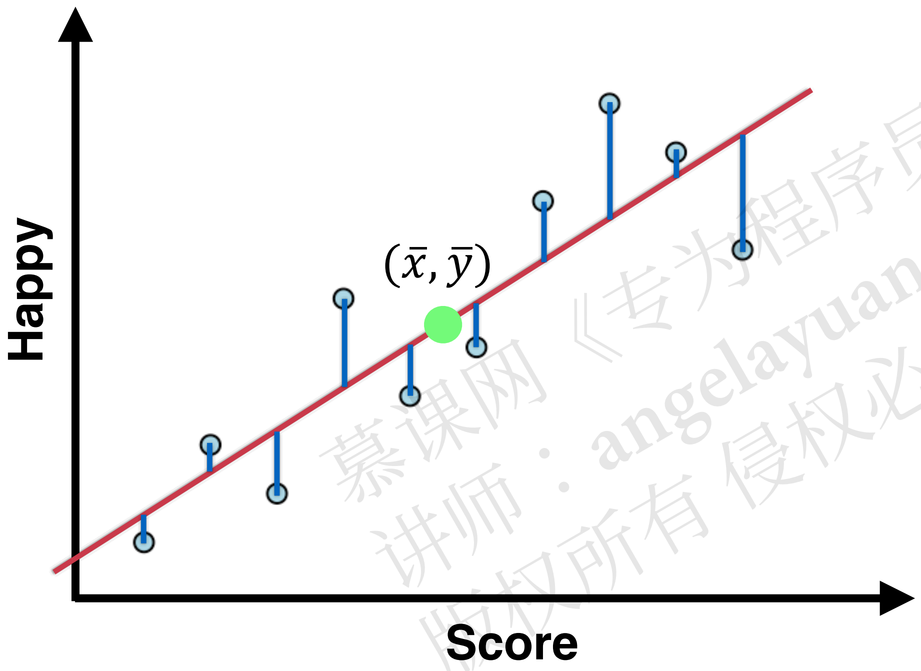
斜率: 成绩每增加1分, 满意度增加0.87分

截距: 成绩为0分时, 满意度为0.73分

截距是否有意义与数据的含义有关

截距的作用只是调整回归直线的高度

相关系数与回归系数的关系



$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

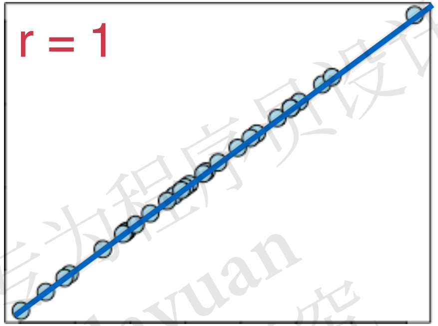
$$= \frac{(n-1)Cov(x,y)}{(n-1)Var(x)} = \frac{Cov(x,y)}{Var(x)}$$

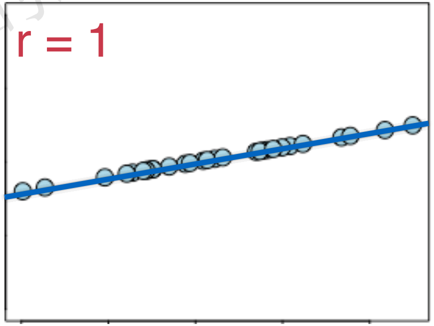
$$r = \frac{Cov(x,y)}{S_x S_y}$$

$$\widehat{\beta_1} = \frac{Cov(x,y)}{S_x S_y} \times \frac{S_y}{S_x} = \frac{S_y}{S_x} r$$

相关系数与回归系数的关系

$$\widehat{\beta_1} = \frac{S_y}{S_x} r$$

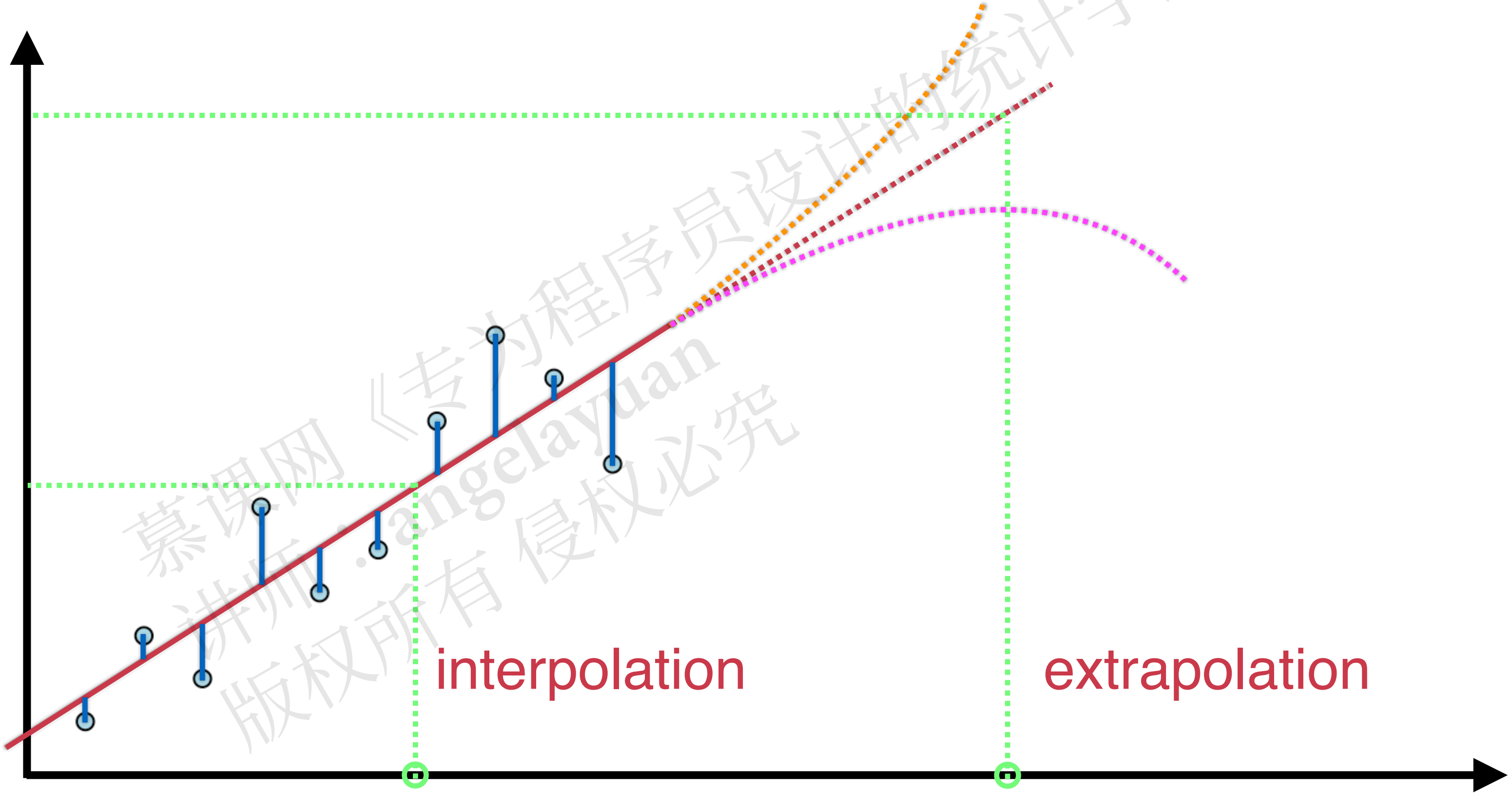




线性关系的方向和强度一样: 回归直线拟合数据的好坏

左侧斜率大(S_y/S_x大), 右侧斜率小(S_y/S_x小): x变化一个单位, y变化的大小

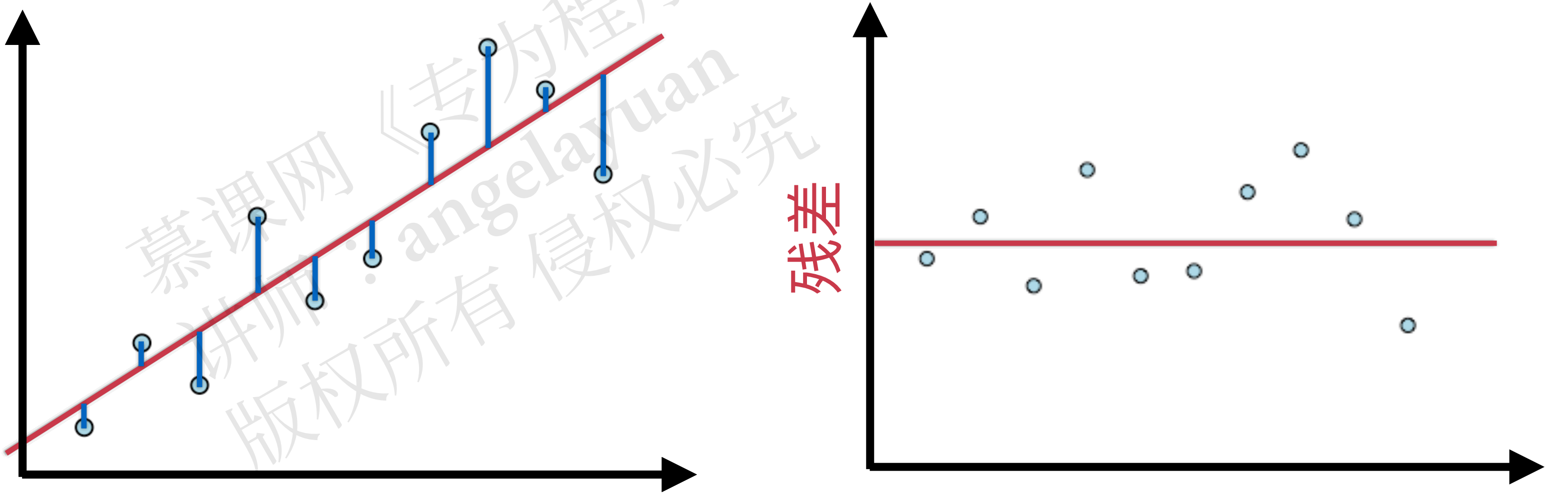
内插(interpolation)与外推(extrapolation)



一元线性回归的前提条件

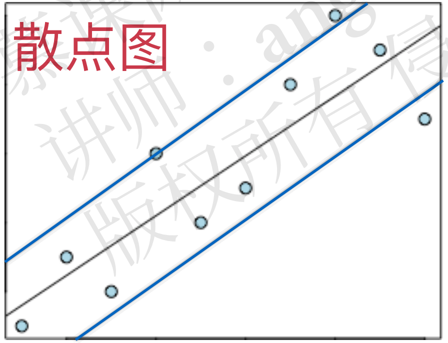
一元线性回归的前提条件

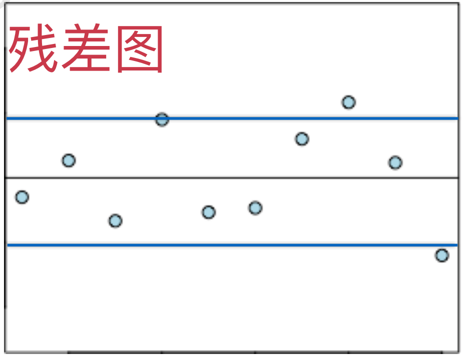
- 线性(linearity): 自变量和因变量之间的关系是线性的
  - 使用散点图或残差图来检验



一元线性回归的前提条件

- 残差(近似)服从均值为0的正态分布
  - 使用频率直方图来检验
- 数据点围绕回归直线的变化程度基本不变(variability constant)
  - 残差围绕直线y=0的变化程度基本不变

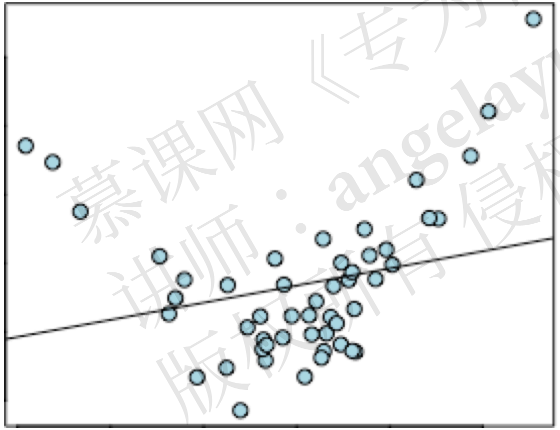




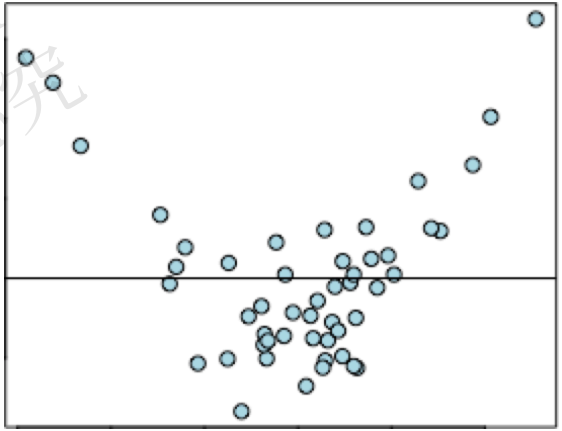
一元线性回归的前提条件

- 数据点围绕回归直线的变化程度基本不变(variability constant)
  - 残差围绕直线y=0的变化程度基本不变

散点图



残差图



回归模型的评价指标

RMSE

$$y = \underline{\beta_0 + \beta_1 x} + \varepsilon$$

数据 模型 误差(残差)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Root Mean Squared Error

R^2

$$y = \beta_0 + \beta_1 x + \varepsilon$$

数据 模型 误差(残差)

因变量的变化

模型可以解释的变化

模型无法解释的变化

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

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

R^2

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

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

使用我们的模型预测产生的错误

使用 $y = \bar{y}$ 预测产生的错误

Baseline model $0 \leq R^2 \leq 1$

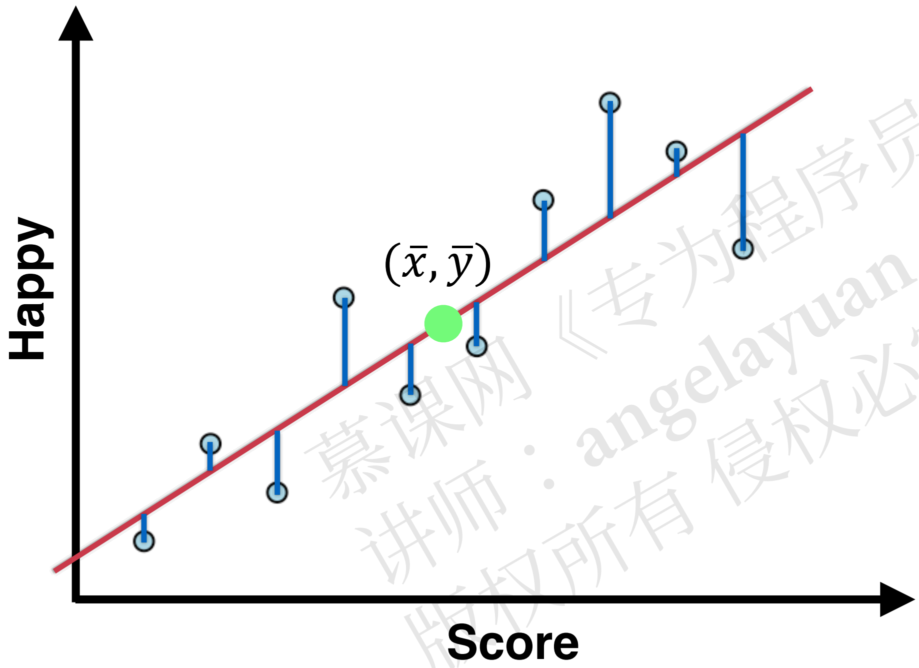
$$= \frac{SS_{model}}{SS_{total}}$$

模型能够解释的变化占总变化的百分比

$$= r^2$$

假设检验

模型误差的方差的点估计



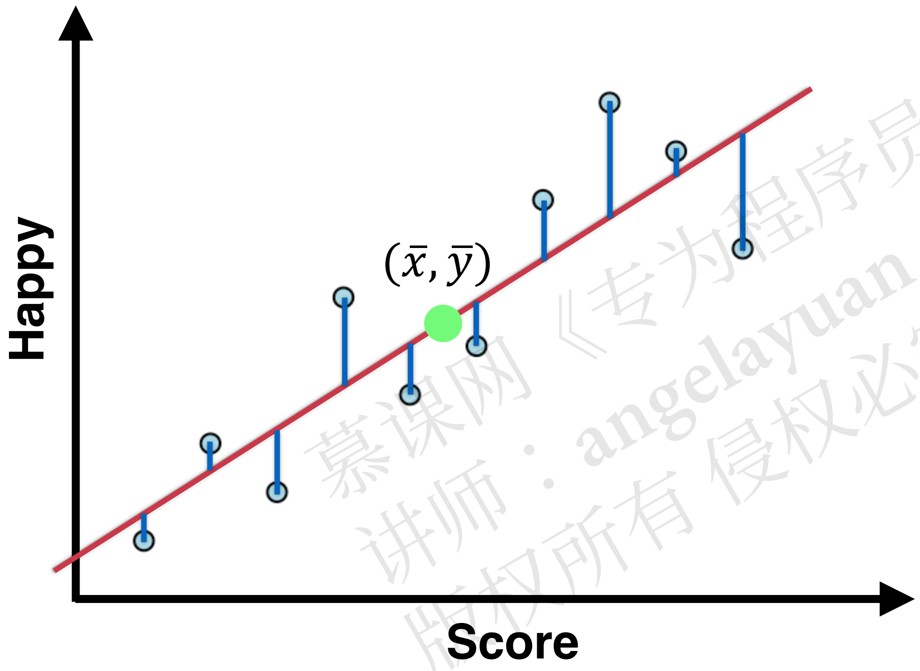
$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0,\sigma^2)$$

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - 2}$$

自由度 = 样本容量 — 模型参数的个数

模型误差的方差的点估计

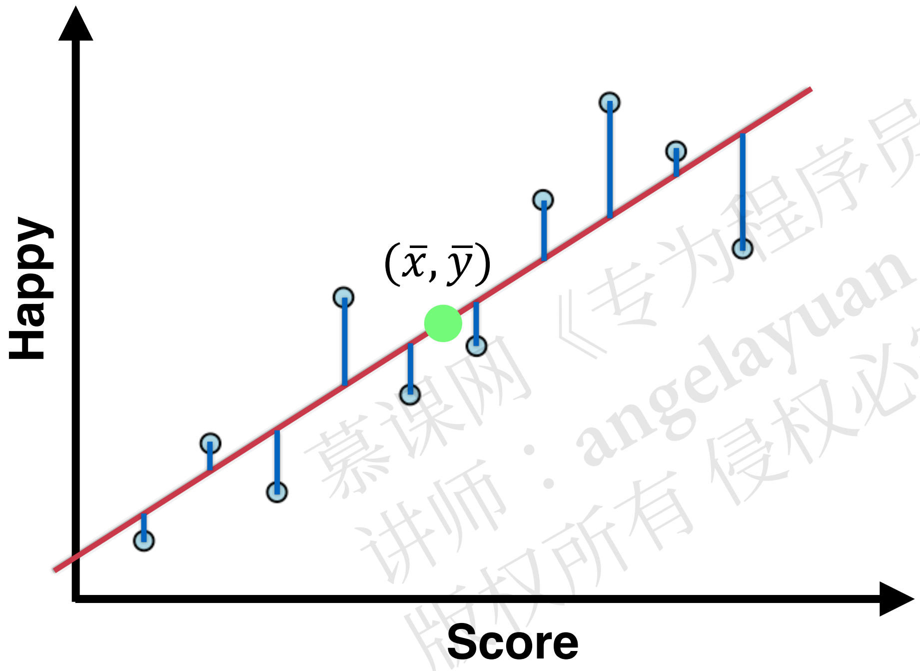


$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - 2}$$

$$SE_{model} = \sqrt{\frac{SS_{residual}}{n-p}}$$

再看相关系数的检验

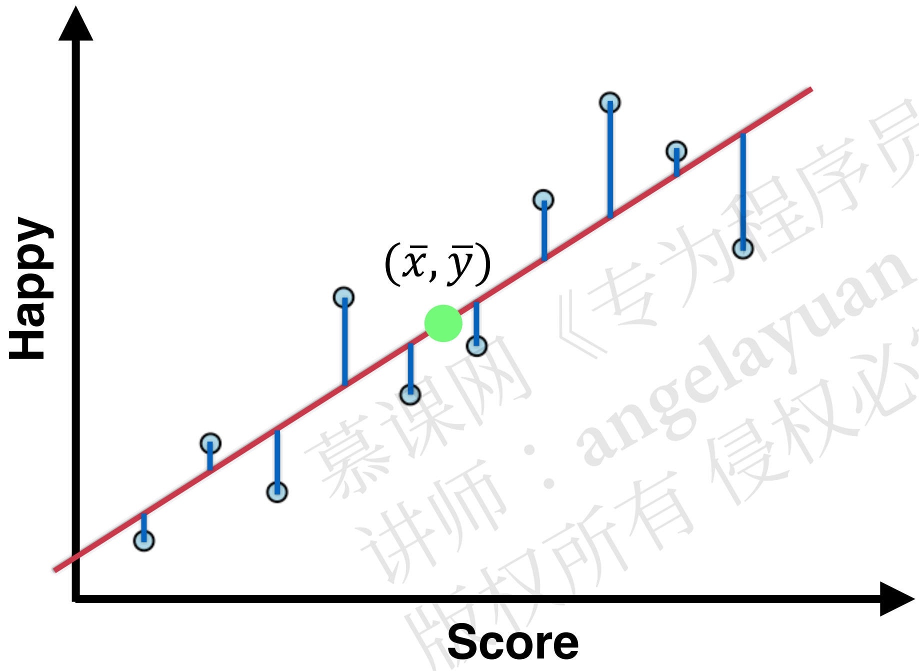


$$\frac{r}{\sqrt{1-r^2}/\sqrt{n-2}} \sim t(n-2)$$

$$\sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{SS_{residual}/SS_{total}}{n-2}}$$

$$= \frac{SE_{model}}{\sqrt{SS_{total}}} = \frac{SE_{model}}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

线性关系的假设检验

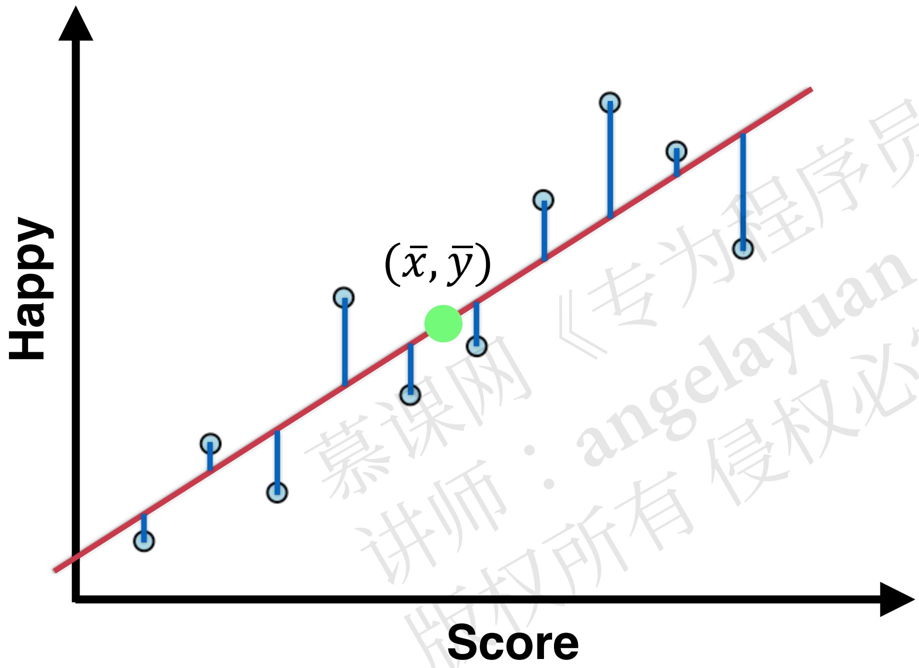


$\hat{y} = \beta_0 + \beta_1 x$ 两个参数

$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$ 两个点估计

通过假设检验来判断参数β_1是否为0, 从而得到线性回归方程是否具有实用价值

线性关系的假设检验



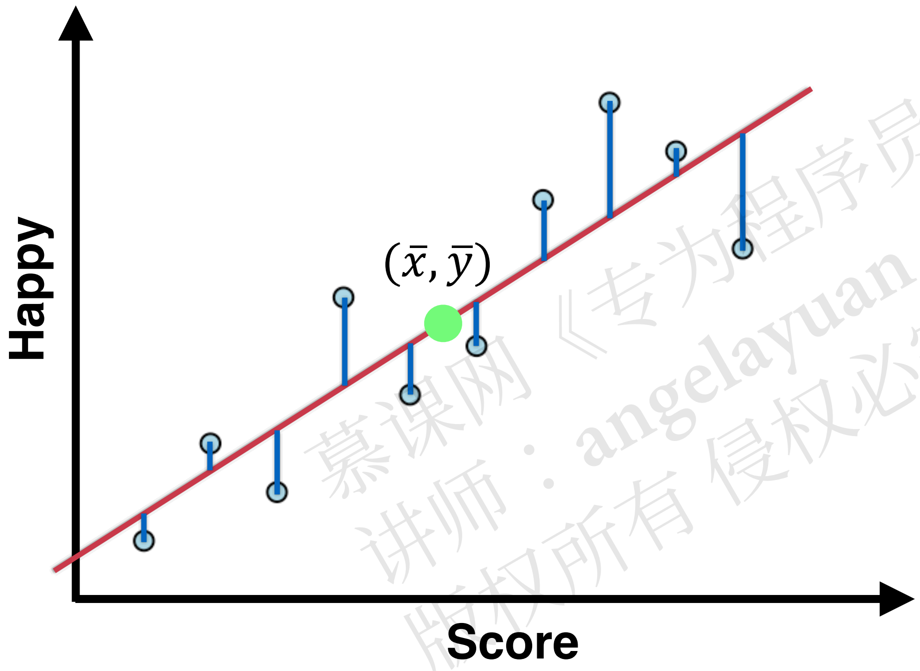
$H_0: \beta_1 = 0$; $H_A: \beta_1 \neq 0$

$$\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{(n-1)Var(x)}\right)$$

$$\widehat{\beta_1} \sim N\left(\beta_1, \left(\boxed{\frac{SE_{model}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}\right)^2\right)$$

$$SE_{\widehat{\beta_1}}$$

线性关系的假设检验



$H_0: \beta_1 = 0$; $H_A: \beta_1 \neq 0$

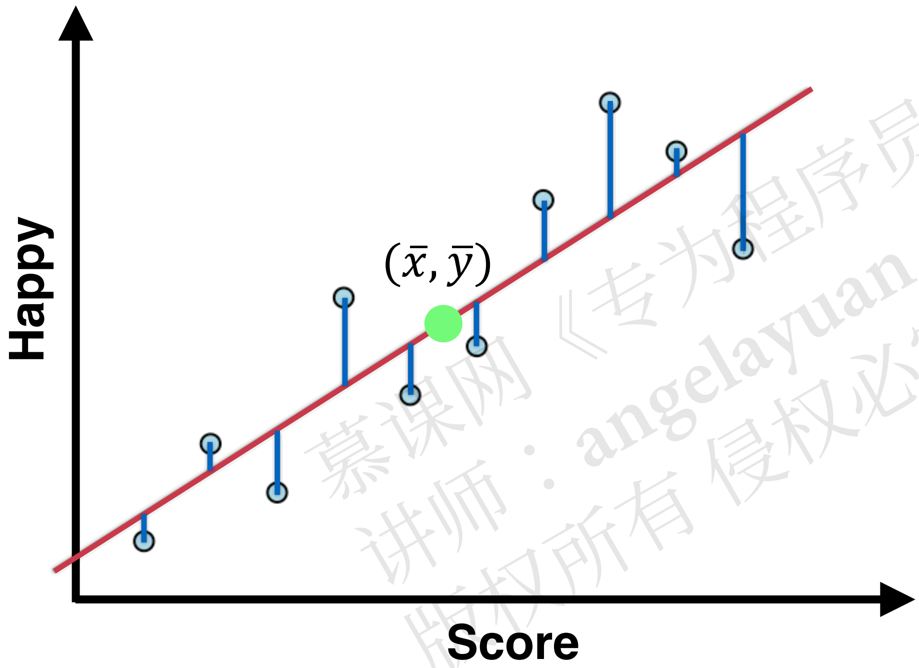
$$\frac{\widehat{\beta_1} - 0}{SE_{\widehat{\beta_1}}} \sim t(n-2)$$

$$SE_{\hat{r}} = \frac{SE_{model}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\widehat{\beta_1} = \frac{S_y}{S_x} r$$

$$SE_{\widehat{\beta_1}} = \frac{SE_{model}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

线性关系的假设检验



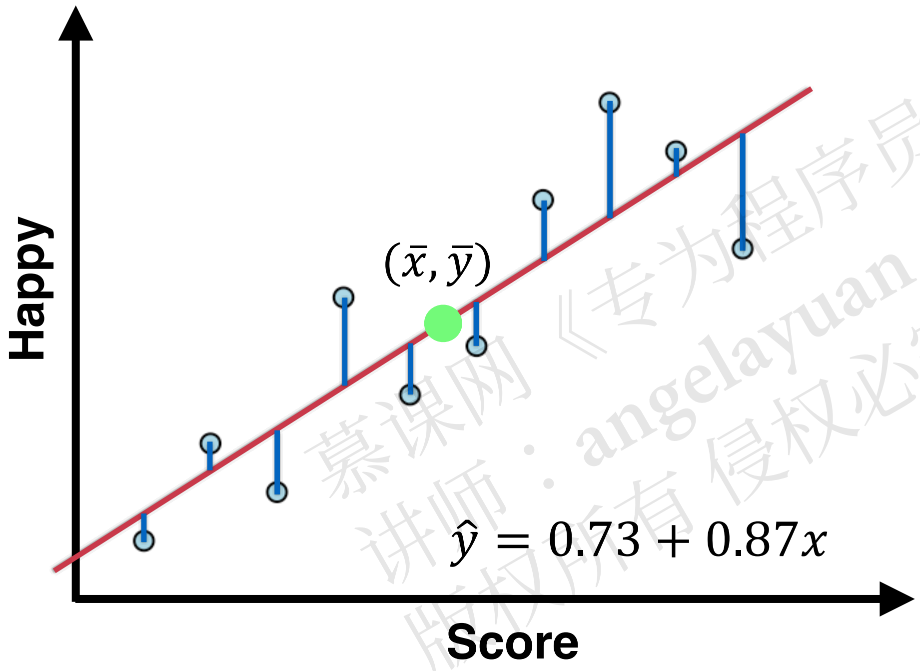
$H_0: \beta_1 = 0; \quad H_A: \beta_1 \neq 0$

$$\frac{r}{\sqrt{1-r^2}/\sqrt{n-2}} \sim t(n-2)$$

$$\frac{\frac{S_y}{S_x} \times r - 0}{\frac{S_y}{S_x} \times SE_{model}/\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \sim t(n-2)$$

$$\frac{\widehat{\beta_1} - 0}{SE_{model}/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim t(n-2)$$

线性关系的假设检验



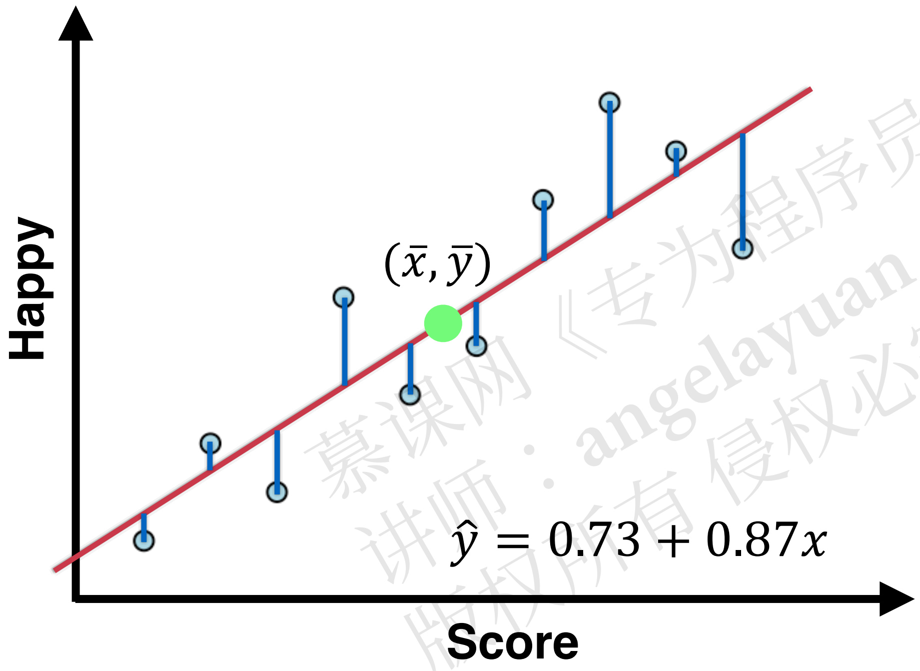
$H_0: \beta_1 = 0$; $H_A: \beta_1 \neq 0$

$$\frac{\widehat{\beta_1} - 0}{SE_{model} / \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \sim t(n-2)$$

T = 4.76

p = 0.0007 x 2 ≈ 0.001

线性关系的假设检验



$H_0: \beta_1 = 0; \quad H_A: \beta_1 \neq 0$

等价于

$H_0: r = 0; \quad H_A: r \neq 0$

另一种视角: ANOVA

$$SS_{model} = SS_{total} - SS_{residual}$$

df = 自变量个数

因变量的变化

模型可以解释的变化

模型无法解释的变化

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

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

df = n - 参数个数

另一种视角: ANOVA

$$\frac{SS_{model}/1}{SS_{residual}/(n-2)} \sim \mathrm{F}(1, \mathrm{n}-2)$$

$$\frac{61.96/1}{20.53/(10-2)} = 24.14$$

p = 0.001

编程实现一元线性回归

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

$$\frac{\widehat{\beta_1} - 0}{SE_{\widehat{\beta_1}}} \sim t(n-2)$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$SE_{\widehat{\beta_1}} = \frac{SE_{model}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

多元线性回归
Multiple Linear Regression

$$\text{房价} = \beta_0 + \beta_1\text{面积}$$

$$\text{房价} = \beta_0 + \beta_1\text{学区房}$$

一元

$$\text{房价} = \beta_0 + \beta_1\text{面积} + \beta_2\text{学区房}$$

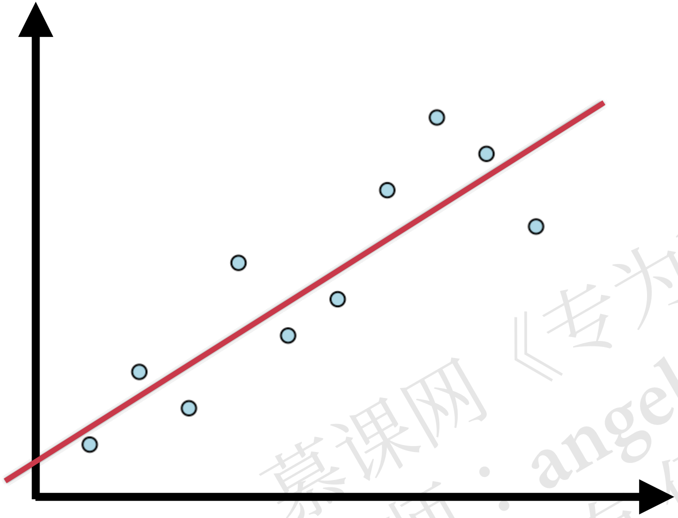
多元

$$\text{房价} = \beta_0 + \beta_1\text{面积} + \beta_2\text{学区房} + \beta_3\text{房间数}$$

$$\text{房价} = \beta_0 + \beta_1\text{面积} + \beta_2\text{学区房} + \beta_3\text{房间数} + \beta_4\text{楼层}$$

数值变量　数值变量　分类变量　数值变量　数值变量

房价 = $\beta_0 + \beta_1$面积

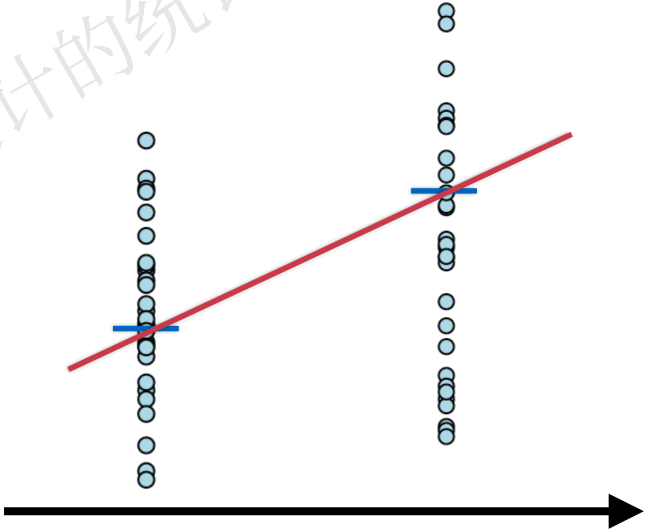


面积增加1, 房价增加β_1

房价 = $\beta_0 + \beta_1$学区房

0: 否

1: 是



β_0: 非学区房的房价均值

$\beta_0 + \beta_1$: 学区房的房价均值

多元线性回归的系数

房价 = $\beta_0 + \beta_1$面积 + β_2学区房

目标: 求β_0,β_1,β_2, 使得 $\sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ 最小

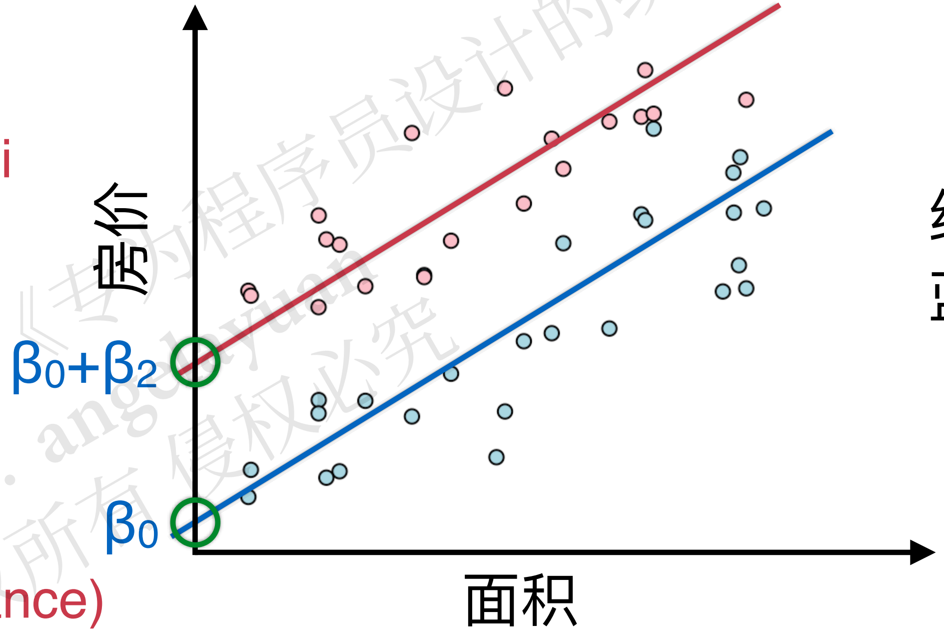
方法: 取J关于β_0,β_1,β_2的偏导数, 并令它们等于0, 求解未知数

房价 = β_0 + β_1面积 + β_2学区房

0: 否
1: 是

在其他自变量的影响不变的情况下，自变量i增加1，因变量增加β_i

β_i: 偏回归系数

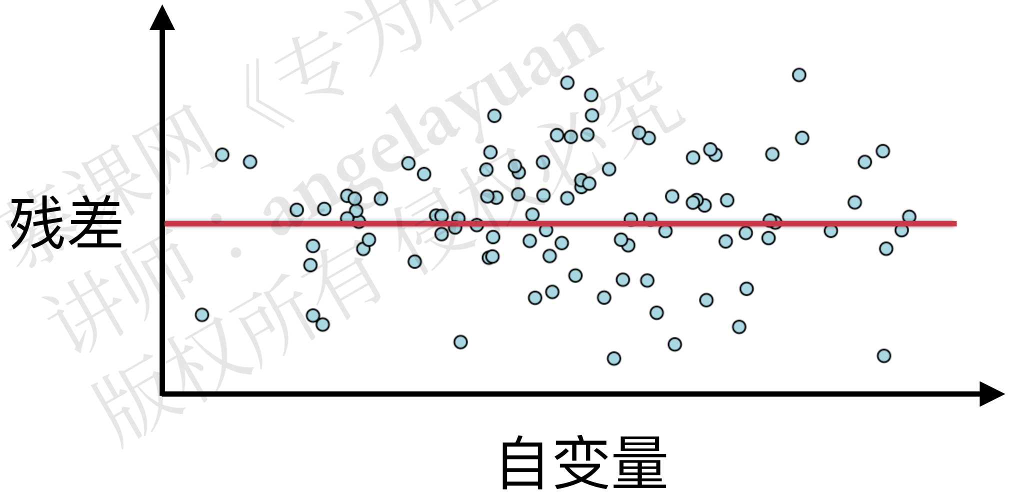


红: 学区房
蓝: 非学区房

One-way ANCOVA
(ANalysis of COVAriance)

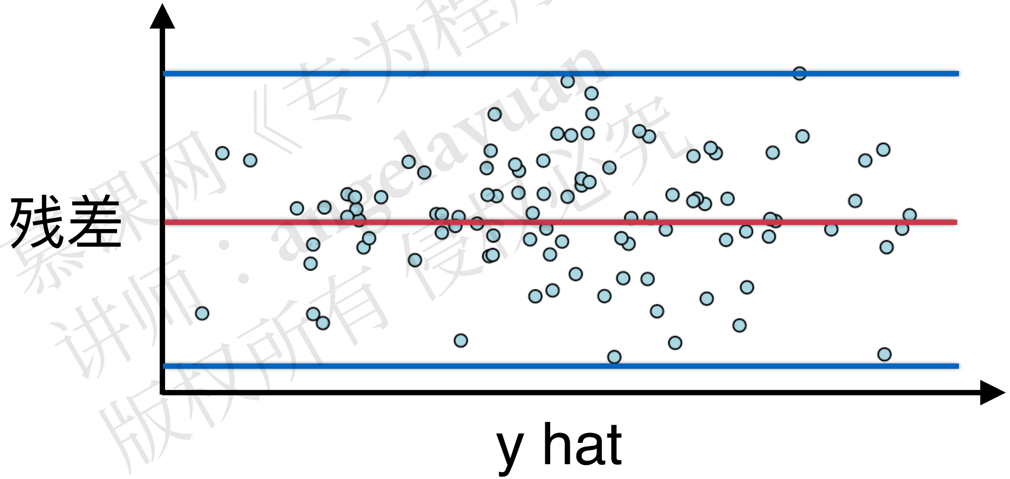
多元线性回归的前提条件

- 线性(linearity): 每个自变量(数值变量)和因变量之间的关系是线性的；使用残差图(残差 vs 自变量)来检验



多元线性回归的前提条件

- 残差围绕直线y=0的变化程度不随因变量估计值的变化而变化；使用残差图(残差 vs y hat)来检验



多元线性回归的前提条件

- 残差(近似)服从均值为0的正态分布; 使用频率直方图来检验
- 残差独立
  - 观测独立
  - 如果怀疑有时序结构, 可以检查残差图(残差 vs 数据收集顺序)

共线性 (collinearity)

- 如果两个自变量相关, 则称这两个自变量共线
- 自变量(independent variables)之间应该彼此独立
- 模型中存在高度共线的自变量会使得模型难以估计准确
- 避免加入与模型中已有的自变量高度共线的自变量，因为该自变量的加入并不能提供有效的新信息, 并且会影响模型的估计

简约性 (parsimony)

- 最简单且有效的模型是最好的
- 奥卡姆剃刀(Occam’s Razor)
    - 由哲学家奥卡姆提出的一个解决问题的法则
    - 对于同一个问题有多种理论可以做出同样准确的预测，那么应该挑选其中使用假设最少的理论

评价多元线性回归模型

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

往模型中添加任意自变量都会使R^2增加

评价多元线性回归模型

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

往模型中添加任意自变量都会使R^2增加

$$R^2_{adjusted} = 1 - \frac{SS_{residual}}{SS_{total}} \times \frac{n-1}{n-k-1}$$

k = 自变量个数

R^2 > adjusted R^2

如果往模型中添加的自变量没有提供有效信息，则adjusted R^2不会增加

假设检验

- 对模型的假设检验

  $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$

  H_A: 至少有一个β不等于0

$$\frac{SS_{model}/k}{SS_{residual}/(n-k-1)} \sim \mathrm{F}(\mathrm{k}, \mathrm{n}-\mathrm{k}-1)$$

- F检验的结果显著: 不代表模型对数据的拟合好, 只代表至少有一个自变量的系数不为0
- F检验的结果不显著: 不代表模型中的自变量不能预测因变量, 只代表这些自变量的组合不是一个好模型

假设检验

- 对(偏)回归系数的假设检验

H_0: $\beta_i = 0$, 当其他自变量被包括在模型中时

H_A: $\beta_i \neq 0$, 当其他自变量被包括在模型中时

$$\frac{\widehat{\beta}_i}{SE_{\widehat{\beta}_i}} \sim t(n-k-1)$$

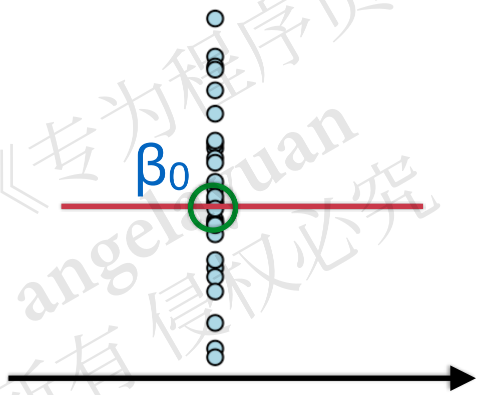
单个自变量的贡献　　　　还可以考察自变量的组合的贡献

常用的检验都是回归的一种特殊形式

One sample t-test	↔	Regression model with intercept only
Two sample t-test	↔	Regression with a categorical explanatory variable
Correlation test	↔	Regression with a numerical explanatory variable
One-way ANOVA	↔	Regression with categorical explanatory variables

One sample t-test ⟷ Regression model with intercept only

$$\text{房价} = \beta_0$$

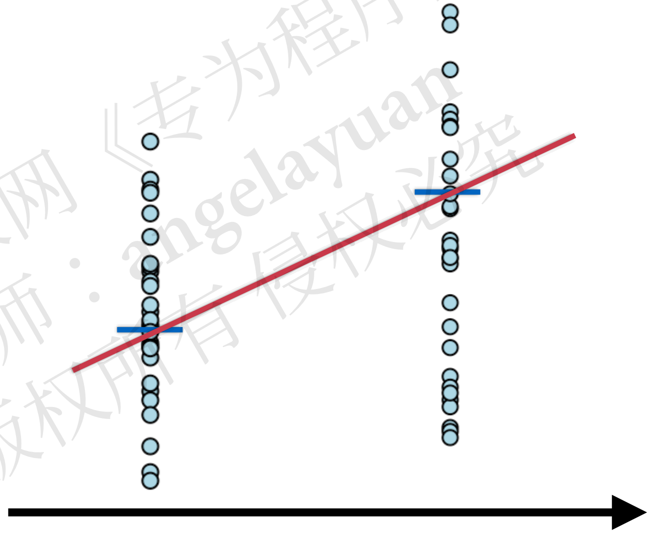


Two sample t-test ⟷ Regression with a categorical explanatory variable

$$\text{房价} = \beta_0 + \beta_1\text{学区房}$$

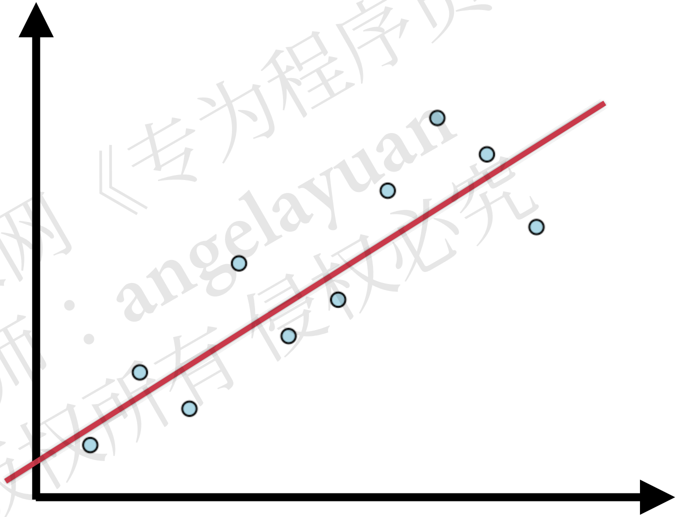
0: 否

1: 是



Correlation test ⟷ Regression with a numerical explanatory variable

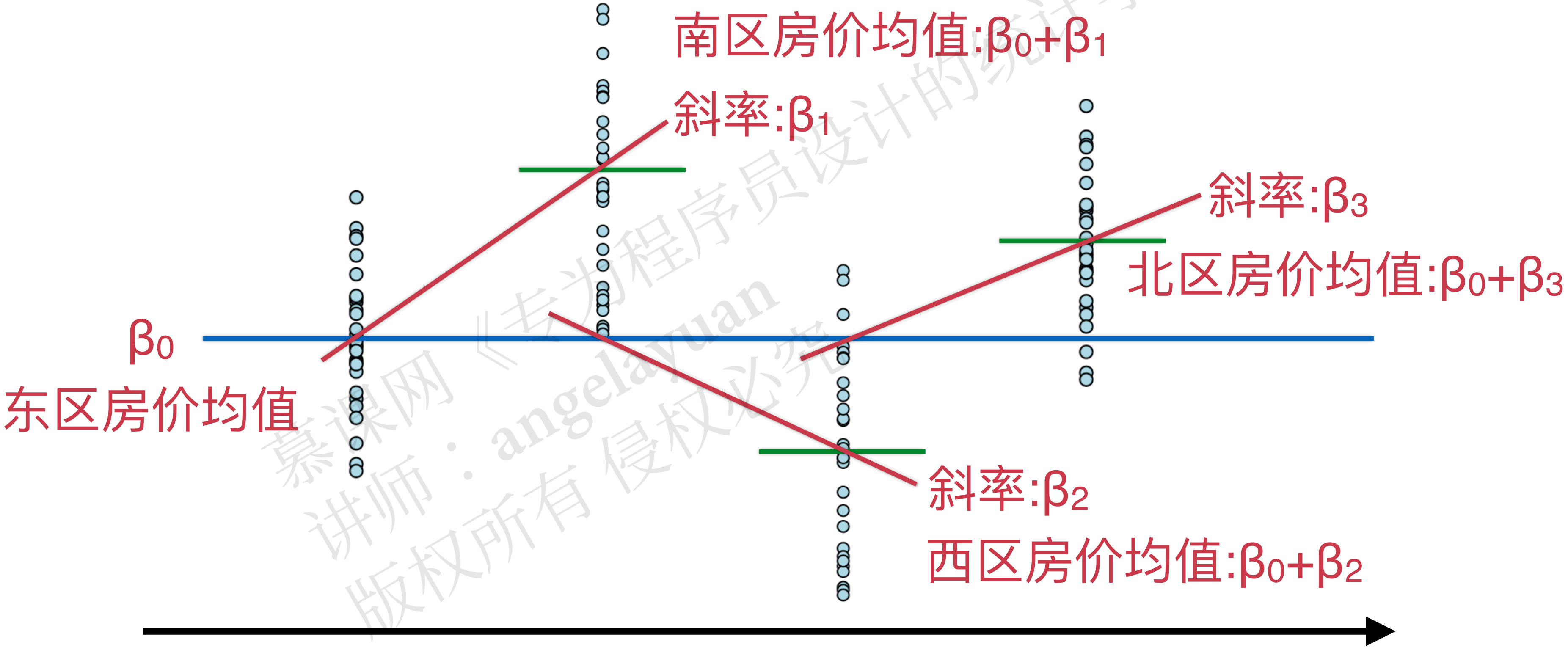
$$\text{房价} = \beta_0 + \beta_1\text{面积}$$



One-way ANOVA ⟷ Regression with categorical explanatory variables

- 考察房价与区域(东,南,西,北)的关系
- 把分类变量“区域”转化成三个哑变量(dummy variables)
  - x_1: 0 = 非南, 1 = 南
  - x_2: 0 = 非西, 1 = 西
  - x_3: 0 = 非北, 1 = 北

$$\text{房价} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$



统计中的回归与机器学习中的回归

相同点

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\hat{y} = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \cdots + \beta_k\, x_k$$

最小化 $\sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

求偏导数

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

$$R^2_{adjusted} = 1 - \frac{SS_{residual}}{SS_{total}} \times \frac{n-1}{n-k-1}$$

不同点

统计学中的线性回归

- 注重可解释性(hypothesis driven)
- 聚焦当前数据，从数据中挖掘本质和机制
- 前提条件; 假设检验
- 考虑影响模型估计准确性的因素

机器学习中的线性回归

- 注重可应用性(data driven)
- 预测; cross-validation

特征工程

本章小结


线性回归
协方差
相关及假设检验
一元线性回归
回归方程
最小二乘法
偏导数
系数的含义
前提条件
评价指标
假设检验
编程实现
多元线性回归
回归方程
最小二乘法
偏导数
系数的含义
前提条件
共线性
评价指标
假设检验
回归统一各种检验
统计学vs机器学习