回归分析

续本达

清华大学 工程物理系

2023-12-04

复习

变差分解

$$S_T = S_A + S_E$$

$$\nu_T = \nu_A + \nu_E$$

- 偏差平方和与自由度具有可加性

统计量 F

$$\frac{\overline{S_A}}{\overline{S_E}} = F \sim F(\nu_A, \nu_E)$$

假设检验

$H_0 : \mu_1 = \mu_2 = \mu_3, H_1 : \mu_1, \mu_2, \mu_3$不全相等

引子

电子体温计

- 电子体温计使用方便

工作原理：金属氧化物热敏电阻

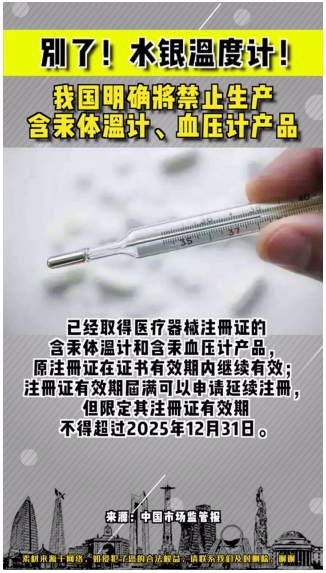
- 负温度系数型：随着温度升高，电阻值变小

电子体温计

- 电子体温计使用方便

工作原理：金属氧化物热敏电阻

- 负温度系数型：随着温度升高，电阻值变小

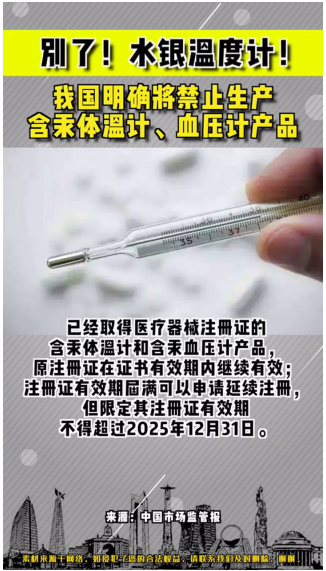


电子体温计

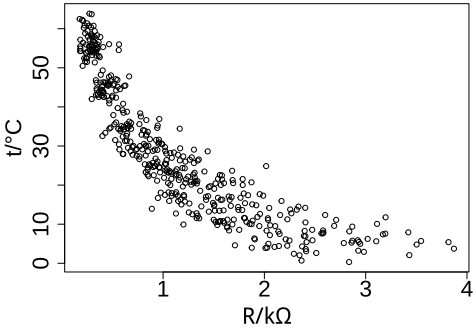
- 电子体温计使用方便

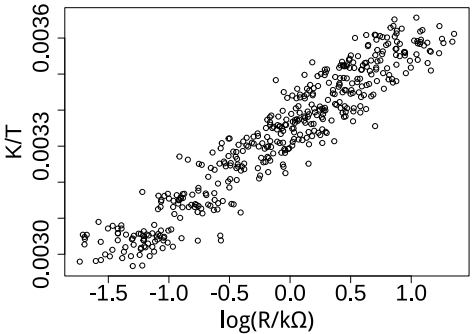
工作原理：金属氧化物热敏电阻

- 负温度系数型：随着温度升高，电阻值变小



热敏电阻的规律

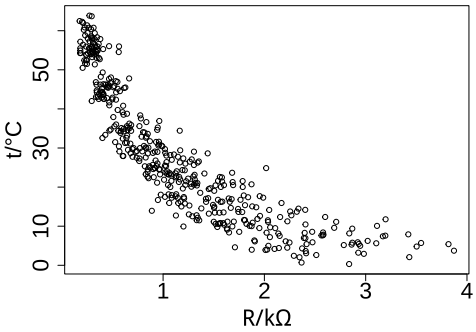


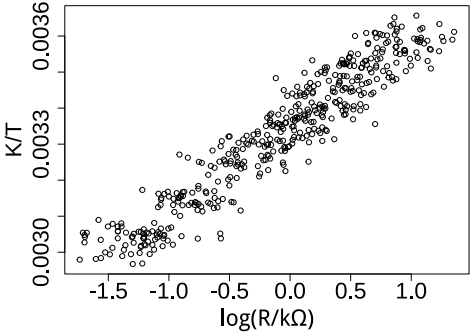


- 温度 t 与电阻 R 有对应关系
- 变成直线 $t \to \dfrac{1}{t + 273.15\,\mathrm{K}}$ 简称为 y，$R \to \log \dfrac{R}{\mathrm{k\Omega}}$ 简称为 x。

Calibration and self-validation of thermistors for high-precision temperature measurements, Measurement Vol 76, 2015

热敏电阻的规律





- 温度 t 与电阻 R 有对应关系
- 变成直线 $t \to \dfrac{1}{t + 273.15\,\mathrm{K}}$ 简称为 y，$R \to \log \dfrac{R}{\mathrm{k\Omega}}$ 简称为 x。

Calibration and self-validation of thermistors for high-precision temperature measurements, Measurement Vol 76, 2015

把非线性关系线性化

- 在许多实际问题中，两个变量之间并不一定是线性关系，而是某种曲线关系，应该用曲线来拟合.
- 可以进行适当的变量代换，把它线性化，这样就把一个非线性回归问题化为线性回归问题而得以解决.

步骤

1. 根据样本数据，在直角坐标系中画出散点图
2. 根据散点图，推测出 Y 与 x 之间的函数关系
3. 选择适当的坐标变换，使之变成线性关系
4. 用线性回归方法求出线性回归方程
5. 返回到原来的函数关系，得到要求的回归方程

把非线性关系线性化

- 在许多实际问题中，两个变量之间并不一定是线性关系，而是某种曲线关系，应该用曲线来拟合.
- 可以进行适当的变量代换，把它线性化，这样就把一个非线性回归问题化为线性回归问题而得以解决.

步骤

1. 根据样本数据，在直角坐标系中画出散点图
2. 根据散点图，推测出 Y 与 x 之间的函数关系
3. 选择适当的坐标变换，使之变成线性关系
4. 用线性回归方法求出线性回归方程
5. 返回到原来的函数关系，得到要求的回归方程

把非线性关系线性化

- 在许多实际问题中，两个变量之间并不一定是线性关系，而是某种曲线关系，应该用曲线来拟合.
- 可以进行适当的变量代换，把它线性化，这样就把一个非线性回归问题化为线性回归问题而得以解决.

步骤

1. 根据样本数据，在直角坐标系中画出散点图
2. 根据散点图，推测出 Y 与 x 之间的函数关系
3. 选择适当的坐标变换，使之变成线性关系
4. 用线性回归方法求出线性回归方程
5. 返回到原来的函数关系，得到要求的回归方程

把非线性关系线性化

- 在许多实际问题中，两个变量之间并不一定是线性关系，而是某种曲线关系，应该用曲线来拟合.
- 可以进行适当的变量代换，把它线性化，这样就把一个非线性回归问题化为线性回归问题而得以解决.

步骤

1. 根据样本数据，在直角坐标系中画出散点图
2. 根据散点图，推测出 Y 与 x 之间的函数关系
3. 选择适当的坐标变换，使之变成线性关系
4. 用线性回归方法求出线性回归方程
5. 返回到原来的函数关系，得到要求的回归方程

把非线性关系线性化

- 在许多实际问题中，两个变量之间并不一定是线性关系，而是某种曲线关系，应该用曲线来拟合.
- 可以进行适当的变量代换，把它线性化，这样就把一个非线性回归问题化为线性回归问题而得以解决.

步骤

1. 根据样本数据，在直角坐标系中画出散点图
2. 根据散点图，推测出 Y 与 x 之间的函数关系
3. 选择适当的坐标变换，使之变成线性关系
4. 用线性回归方法求出线性回归方程
5. 返回到原来的函数关系，得到要求的回归方程

把非线性关系线性化

- 在许多实际问题中，两个变量之间并不一定是线性关系，而是某种曲线关系，应该用曲线来拟合.
- 可以进行适当的变量代换，把它线性化，这样就把一个非线性回归问题化为线性回归问题而得以解决.

步骤

1. 根据样本数据，在直角坐标系中画出散点图
2. 根据散点图，推测出 Y 与 x 之间的函数关系
3. 选择适当的坐标变换，使之变成线性关系
4. 用线性回归方法求出线性回归方程
5. 返回到原来的函数关系，得到要求的回归方程

把非线性关系线性化

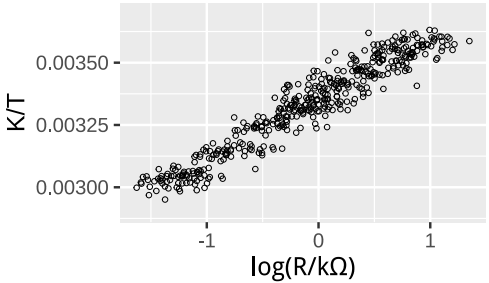
- 在许多实际问题中，两个变量之间并不一定是线性关系，而是某种曲线关系，应该用曲线来拟合.
- 可以进行适当的变量代换，把它线性化，这样就把一个非线性回归问题化为线性回归问题而得以解决.

步骤

1. 根据样本数据，在直角坐标系中画出散点图
2. 根据散点图，推测出 Y 与 x 之间的函数关系
3. 选择适当的坐标变换，使之变成线性关系
4. 用线性回归方法求出线性回归方程
5. 返回到原来的函数关系，得到要求的回归方程

线性回归

直线表征规律



- 变量 x 和 y 于图中呈直线关系。
- 欲求出与各点最匹配的直线，

$$y = a + bx$$

匹配：横向、纵向　　、垂线?

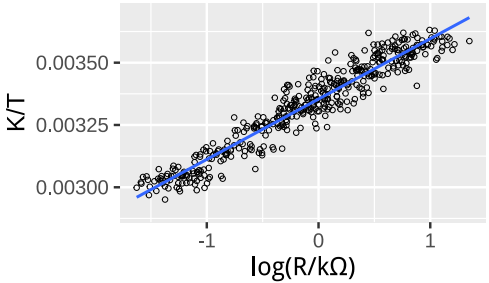
- 记观测值为 $(x_1, y_1), \cdots, (x_N, y_N)$

最小二乘法估计 a, b

取残差平方和 $E^2(a,b) = \sum_{i=1}^{N} [y_i - (a + bx_i)]^2$ 刻画点与线的差异，

以使 E^2 最小时的 a, b 为估计值 $\hat{a}, \hat{b} = \arg\min_{a,b} E^2(a,b)$

直线表征规律



- 变量 x 和 y 于图中呈直线关系。
- 欲求出与各点最匹配的直线，

$$y = a + bx$$

匹配：横向、纵向　　、垂线?

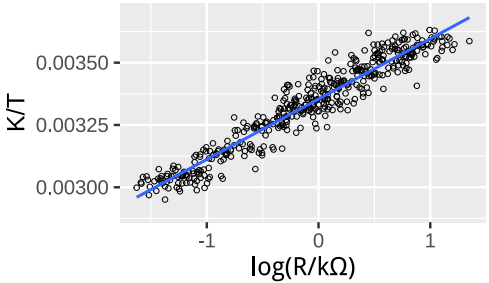
- 记观测值为 $(x_1, y_1), \cdots, (x_N, y_N)$

最小二乘法估计 a, b

取残差平方和 $E^2(a,b) = \sum_{i=1}^{N} [y_i - (a + bx_i)]^2$ 刻画点与线的差异，

以使 E^2 最小时的 a, b 为估计值 $\hat{a}, \hat{b} = \arg\min_{a,b} E^2(a,b)$

直线表征规律



- 变量 x 和 y 于图中呈直线关系。
- 欲求出与各点最匹配的直线，

$$y = a + bx$$

匹配：横向、纵向 ✓ 预测、垂线?

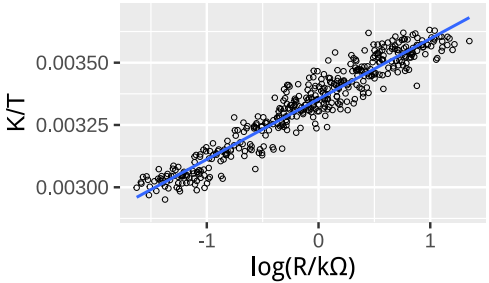
- 记观测值为 $(x_1, y_1), \cdots, (x_N, y_N)$

最小二乘法估计 a, b

取残差平方和 $E^2(a,b) = \sum_{i=1}^{N} [y_i - (a + bx_i)]^2$ 刻画点与线的差异，

以使 E^2 最小时的 a, b 为估计值 $\hat{a}, \hat{b} = \arg\min_{a,b} E^2(a,b)$

直线表征规律



- 变量 x 和 y 于图中呈直线关系。
- 欲求出与各点最匹配的直线，

$$y = a + bx$$

匹配：横向、纵向 ✓ 预测、垂线?

- 记观测值为 $(x_1, y_1), \cdots, (x_N, y_N)$

最小二乘法估计 a, b

取残差平方和 $E^2(a,b) = \sum_{i=1}^{N} [y_i - (a + bx_i)]^2$ 刻画点与线的差异，

以使 E^2 最小时的 a, b 为估计值 $\hat{a}, \hat{b} = \arg\min_{a,b} E^2(a,b)$

最小二乘法的求解

$\hat{a}, \hat{b} = \arg\min_{a,b} E^2(a,b)$ 有多种解法：配平方；求导 $\frac{\partial E^2}{\partial (a,b)} = 0$。

线性代数视角

把残差记为 $\epsilon_i = y_i - (a + bx_i)$，有 $y_i = 1a + x_i b + \epsilon_i$，写成 N 维列向量，

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} a + \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} b + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

最小二乘法的求解

$\hat{a}, \hat{b} = \arg\min_{a,b} E^2(a,b)$ 有多种解法：配平方；求导 $\dfrac{\partial E^2}{\partial (a,b)} = 0$。

线性代数视角

把残差记为 $\epsilon_i = y_i - (a + bx_i)$，有 $y_i = 1a + x_i b + \epsilon_i$，写成 N 维列向量，

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} a + \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} b + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

最小二乘法的求解

$\hat{a}, \hat{b} = \arg\min_{a,b} E^2(a,b)$ 有多种解法：配平方；求导 $\dfrac{\partial E^2}{\partial(a,b)} = 0$。

线性代数视角

把残差记为 $\epsilon_i = y_i - (a + bx_i)$，有 $y_i = 1a + x_i b + \epsilon_i$，写成 N 维列向量，

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} a + \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} b + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

最小二乘法的求解

$\hat{a}, \hat{b} = \arg\min_{a,b} E^2(a,b)$ 有多种解法：配平方；求导 $\dfrac{\partial E^2}{\partial (a,b)} = 0$。

线性代数视角

把残差记为 $\epsilon_i = y_i - (a + bx_i)$，有 $y_i = 1a + x_i b + \epsilon_i$，写成 N 维列向量，

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} a + \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} b + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

$$\vec{Y} = (\vec{X}_1, \vec{X}_2) \begin{pmatrix} a \\ b \end{pmatrix} + \vec{E}$$

- 残差平方和 $E^2 = \sum \epsilon_i^2$ 是 $\vec{E}$ 欧几里得距离的平方 $\|\vec{E}\|^2$

向子空间投影

$\vec{Y} = (\vec{X}_1, \vec{X}_2) \begin{pmatrix} a \\ b \end{pmatrix} + \vec{E}$，当 a, b 为何值能使 $\|\vec{E}\|^2$ 最小?

- N 维线性空间中，当 $a\vec{X}_1 + b\vec{X}_2$ 是 $\vec{Y}$ 向子空间的投影时，$\|\vec{E}\|^2$ 最小。

$\vec{E} = \vec{Y} - (a\vec{X}_1 + b\vec{X}_2)$ 取极小

$\vec{X}_1$ 与 $\vec{X}_2$ 张成 2 维子空间

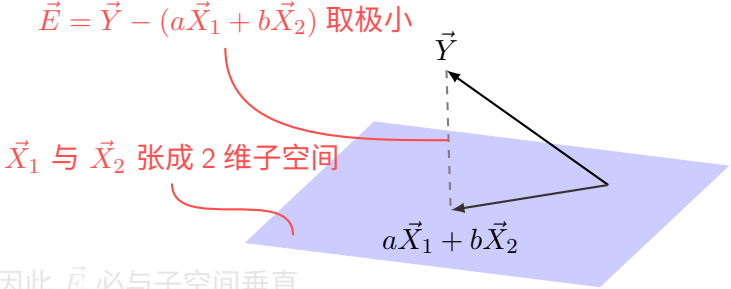
- 因此 $\vec{E}$ 必与子空间垂直

$$\begin{cases} \vec{E} \cdot \vec{X}_1 = 0 \\ \vec{E} \cdot \vec{X}_2 = 0 \end{cases}$$

向子空间投影

$\vec{Y} = (\vec{X}_1, \vec{X}_2) \begin{pmatrix} a \\ b \end{pmatrix} + \vec{E}$，当 a, b 为何值能使 $\|\vec{E}\|^2$ 最小?

- N 维线性空间中，当 $a\vec{X}_1 + b\vec{X}_2$ 是 $\vec{Y}$ 向子空间的投影时，$\|\vec{E}\|^2$ 最小。



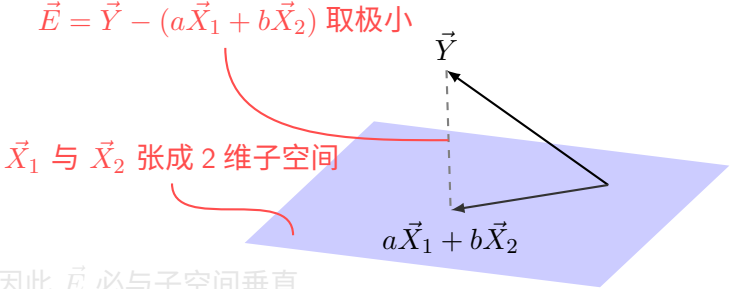
- 因此 $\vec{E}$ 必与子空间垂直

$$\begin{cases} \vec{E} \cdot \vec{X}_1 = 0 \\ \vec{E} \cdot \vec{X}_2 = 0 \end{cases}$$

向子空间投影

$\vec{Y} = (\vec{X}_1, \vec{X}_2)\begin{pmatrix} a \\ b \end{pmatrix} + \vec{E}$，当 a, b 为何值能使 $\|\vec{E}\|^2$ 最小?

- N 维线性空间中， 当 $a\vec{X}_1 + b\vec{X}_2$ 是 $\vec{Y}$ 向子空间的投影时，$\|\vec{E}\|^2$ 最小。



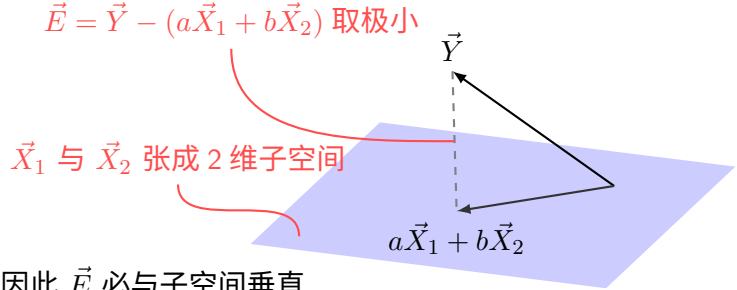
- 因此 $\vec{E}$ 必与子空间垂直

$$\begin{cases} \vec{E} \cdot \vec{X}_1 = 0 \\ \vec{E} \cdot \vec{X}_2 = 0 \end{cases}$$

向子空间投影

$\vec{Y} = (\vec{X}_1, \vec{X}_2) \begin{pmatrix} a \\ b \end{pmatrix} + \vec{E}$，当 a, b 为何值能使 $\|\vec{E}\|^2$ 最小?

- N 维线性空间中， 当 $a\vec{X}_1 + b\vec{X}_2$ 是 $\vec{Y}$ 向子空间的投影时，$\|\vec{E}\|^2$ 最小。



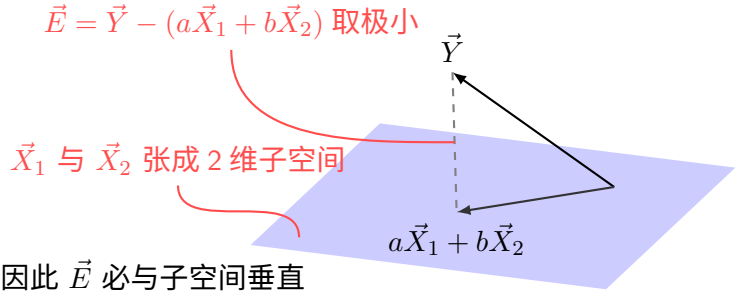
- 因此 $\vec{E}$ 必与子空间垂直

$$\begin{cases} \vec{E} \cdot \vec{X}_1 = 0 \\ \vec{E} \cdot \vec{X}_2 = 0 \end{cases}$$

向子空间投影

$\vec{Y} = (\vec{X}_1, \vec{X}_2)\begin{pmatrix} a \\ b \end{pmatrix} + \vec{E}$，当 a, b 为何值能使 $\|\vec{E}\|^2$ 最小?

- N 维线性空间中， 当 $a\vec{X}_1 + b\vec{X}_2$ 是 $\vec{Y}$ 向子空间的投影时，$\|\vec{E}\|^2$ 最小。



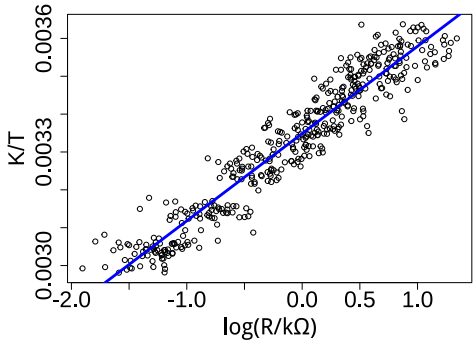
- 因此 $\vec{E}$ 必与子空间垂直

$$\begin{cases} \vec{E} \cdot \vec{X}_1 = \sum_i [y_i - (a + bx_i)]1 = \dfrac{\partial E^2}{\partial a} = 0 \\ \vec{E} \cdot \vec{X}_2 = \sum_i [y_i - (a + bx_i)]x_i = \dfrac{\partial E^2}{\partial b} = 0 \end{cases}$$

求解回归方程

$$\begin{cases} \hat{a} = 3.353(3) \times 10^{-3} \\ \hat{b} = 2.364(35) \times 10^{-4} \end{cases}$$

$$y = \hat{a} + \hat{b}x$$



$$\frac{1}{t/^\circ\mathrm{C} + 273.15} = \hat{a} + \hat{b}\log\frac{R}{\mathrm{k}\Omega}$$

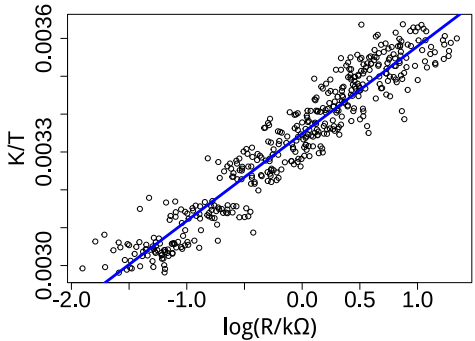
$$\Longrightarrow \frac{t}{^\circ\mathrm{C}} = \left(\hat{a} + \hat{b}\log\frac{R}{\mathrm{k}\Omega}\right)^{-1} - 273.15$$

- 通过电路测得 R，即可通过公式转化得到温度 t 的读数。

求解回归方程

$$\begin{cases} \hat{a} = 3.353(3) \times 10^{-3} \\ \hat{b} = 2.364(35) \times 10^{-4} \end{cases}$$

$$y = \hat{a} + \hat{b}x$$



$$\frac{1}{t/\text{℃} + 273.15} = \hat{a} + \hat{b}\log\frac{R}{\text{k}\Omega}$$

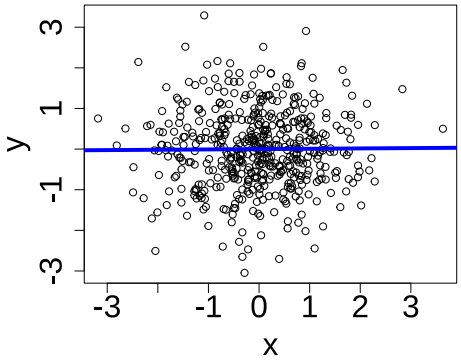
$$\Longrightarrow \frac{t}{\text{℃}} = \left(\hat{a} + \hat{b}\log\frac{R}{\text{k}\Omega}\right)^{-1} - 273.15$$

- 通过电路测得 R，即可通过公式转化得到温度 t 的读数。

显著性检验

显著性检验的必要性

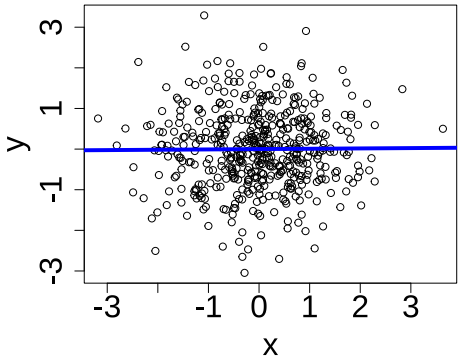
即使两个量毫无关联，仍可（流于形式地）做最小二乘法。



- 必须对回归结果进行假设检验，才能理解回归系数的含义。
- 回归所得的 b ，与 0 有显著差异吗?

显著性检验的必要性

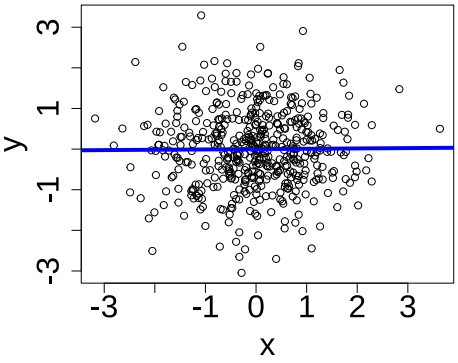
即使两个量毫无关联，仍可（流于形式地）做最小二乘法。

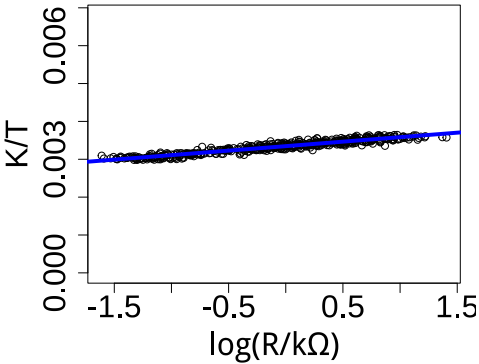


- 必须对回归结果进行假设检验，才能理解回归系数的含义。
- 回归所得的 b ，与 0 有显著差异吗?

显著性检验的必要性

即使两个量毫无关联，仍可（流于形式地）做最小二乘法。





- 必须对回归结果进行假设检验，才能理解回归系数的含义。
- 回归所得的 b ，与 0 有显著差异吗?

平方和分解

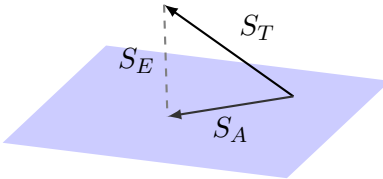
假设：残差服从正态分布

$$\epsilon_i = y_i - (a + bx_i) \sim N(0, \sigma^2)$$

经过最小二乘法获得的最佳 a, b 记为 $\hat{a}, \hat{b}$

记 $\hat{y}_i = \hat{a} + \hat{b}x_i$，称为回归的预测值：

$$\underbrace{\sum_{i=1}^{N}(y_i - \overline{y})^2}_{S_T\text{偏差平方和}} = \underbrace{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}_{S_E\text{误差平方和}} + \underbrace{\sum_{i=1}^{N}(\overline{y} - \hat{y}_i)^2}_{S_A\text{效应平方和}}$$



R^2 值

$$R^2 = \frac{S_A}{S_T}$$

越接近于 1 ，越说明 S_A 主导 S_T，相关性越强。

假设检验问题

- 对任意两个变量的一组观察值 $\{(x_i, y_i)\}$ 都可以用最小二乘法形式上求得 y 对 x 的回归方程, 如果 y 与 x 没有线性相关关系, 这种形式的回归方程就没有意义
- 因此需要考察 y 与 x 间是否确有线性相关关系, 这就是回归效果的 检验问题 .

R^2 值

$$R^2 = \frac{S_A}{S_T}$$

越接近于 1 ，越说明 S_A 主导 S_T，相关性越强。

假设检验问题

- 对任意两个变量的一组观察值 $\{(x_i, y_i)\}$ 都可以用最小二乘法形式上求得 y 对 x 的回归方程, 如果 y 与 x 没有线性相关关系, 这种形式的回归方程就没有意义
- 因此需要考察 y 与 x 间是否确有线性相关关系, 这就是回归效果的 检验问题 .

R^2 值

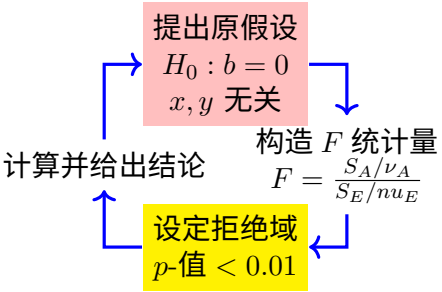
$$R^2 = \frac{S_A}{S_T}$$

越接近于 1 ，越说明 S_A 主导 S_T，相关性越强。

假设检验问题

- 对任意两个变量的一组观察值 $\{(x_i, y_i)\}$ 都可以用最小二乘法形式上求得 y 对 x 的回归方程, 如果 y 与 x 没有线性相关关系, 这种形式的回归方程就没有意义
- 因此需要考察 y 与 x 间是否确有线性相关关系, 这就是回归效果的 检验问题 .

F 假设检验



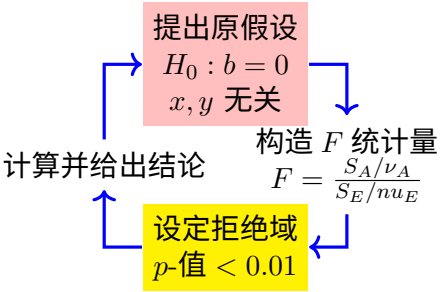
自由度

- $\nu_T = N - 1$
- $\nu_A = p - 1$
- $\nu_E = N - p$
- p 是回归参数个数，此处共有两个参数 a, b， 故 $p = 2$

	自由度	平方和	平方和/自由度	F	p-值
电阻	1	1.348×10^{-5}	1.348×10^{-5}	4545	$< 2 \times 10^{-16}$
残差	438	1.300×10^{-6}	3.000×10^{-9}		
总计	439	1.478×10^{-5}			

- p-值 < 0.01 成立，拒绝原假设，$b \neq 0$，回归结果有意义。

F 假设检验



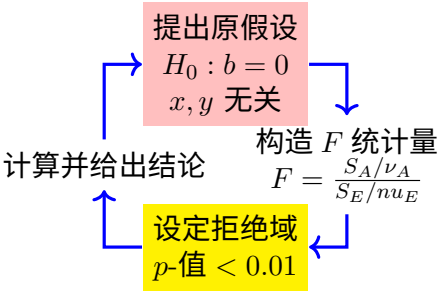
自由度

- $\nu_T = N - 1$
- $\nu_A = p - 1$
- $\nu_E = N - p$
- p 是回归参数个数，此处共有两个参数 a, b， 故 $p = 2$

	自由度	平方和	平方和/自由度	F	p-值
电阻	1	1.348×10^{-5}	1.348×10^{-5}	4545	$< 2 \times 10^{-16}$
残差	438	1.300×10^{-6}	3.000×10^{-9}		
总计	439	1.478×10^{-5}			

- p-值 < 0.01 成立，拒绝原假设，$b \neq 0$，回归结果有意义。

F 假设检验



自由度

- $\nu_T = N - 1$
- $\nu_A = p - 1$
- $\nu_E = N - p$
- p 是回归参数个数，此处共有两个参数 a, b， 故 $p = 2$

	自由度	平方和	平方和/自由度	F	p-值
电阻	1	1.348×10^{-5}	1.348×10^{-5}	4545	$< 2 \times 10^{-16}$
残差	438	1.300×10^{-6}	3.000×10^{-9}		
总计	439	1.478×10^{-5}			

- p-值 < 0.01 成立，拒绝原假设，$b \neq 0$，回归结果有意义。

σ^2 的估计

$$\hat{y} = \hat{a} + \hat{b}x$$

一元线性回归模型

$$y = a + bx + \epsilon, \epsilon \sim N(0, \sigma^2)$$

- 因随机因素引起的误差称为 残差平方和

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

$$\frac{Q_e}{\sigma^2} \sim \chi^2(n-2)$$

$$\Longrightarrow \hat{\sigma^2} = \frac{Q_e}{n-2}$$

t 检验

检验假设 $H_0 : b = 0, H_1 : b \neq 0$

- $\hat{b}$ 满足 $\hat{b} \sim N\left(b, \dfrac{\sigma^2}{L_{xx}}\right)$ 其中 $L_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}$
- 因此如果原假设成立，$\hat{b}\sqrt{L_{xx}}/\sigma \sim N(0,1)$，$Q_e/\sigma^2 \sim \chi^2(n-2)$

$$\frac{\hat{b}\sqrt{L_{xx}}/\sigma}{\sqrt{\frac{Q_e/\sigma^2}{n-2}}} = \frac{\hat{b}\sqrt{L_{xx}}}{\sqrt{\frac{Q_e}{n-2}}} = t \sim t(n-2)$$

- 选取拒绝域 $|t| \geq t_{\alpha/2}(n-2)$ 进行 t 检验。

区间估计

类似 $\hat{b}$， $\hat{a}$ 以及 $\hat{Y} = \hat{a} + \hat{b}x$ 都可经线性变换构造 t 分布统计量，进行区间估计。

t 检验

检验假设 $H_0 : b = 0, H_1 : b \neq 0$

- $\hat{b}$ 满足 $\hat{b} \sim N\left(b, \dfrac{\sigma^2}{L_{xx}}\right)$ 其中 $L_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}$
- 因此如果原假设成立，$\hat{b}\sqrt{L_{xx}}/\sigma \sim N(0,1)$，$Q_e/\sigma^2 \sim \chi^2(n-2)$

$$\frac{\hat{b}\sqrt{L_{xx}}/\sigma}{\sqrt{\frac{Q_e/\sigma^2}{n-2}}} = \frac{\hat{b}\sqrt{L_{xx}}}{\sqrt{\frac{Q_e}{n-2}}} = t \sim t(n-2)$$

- 选取拒绝域 $|t| \geq t_{\alpha/2}(n-2)$ 进行 t 检验。

区间估计

类似 $\hat{b}$， $\hat{a}$ 以及 $\hat{Y} = \hat{a} + \hat{b}x$ 都可经线性变换构造 t 分布统计量，进行区间估计。

t 检验

检验假设 $H_0 : b = 0, H_1 : b \neq 0$

- $\hat{b}$ 满足 $\hat{b} \sim N\left(b, \dfrac{\sigma^2}{L_{xx}}\right)$ 其中 $L_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}$
- 因此如果原假设成立，$\hat{b}\sqrt{L_{xx}}/\sigma \sim N(0,1)$，$Q_e/\sigma^2 \sim \chi^2(n-2)$

$$\frac{\hat{b}\sqrt{L_{xx}}/\sigma}{\sqrt{\frac{Q_e/\sigma^2}{n-2}}} = \frac{\hat{b}\sqrt{L_{xx}}}{\sqrt{\frac{Q_e}{n-2}}} = t \sim t(n-2)$$

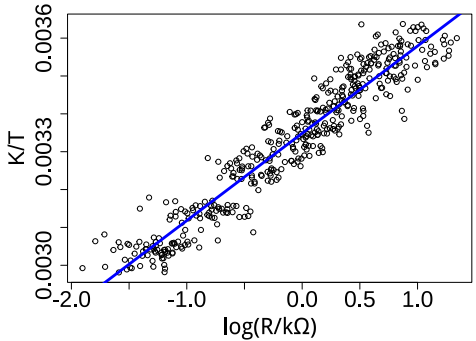
- 选取拒绝域 $|t| \geq t_{\alpha/2}(n-2)$ 进行 t 检验。

区间估计

类似 $\hat{b}$， $\hat{a}$ 以及 $\hat{Y} = \hat{a} + \hat{b}x$ 都可经线性变换构造 t 分布统计量，进行区间估计。

多元线性回归

热敏电阻的更精确模型



$$\frac{K}{T} = a + b \log \frac{R}{\mathrm{k}\Omega}$$

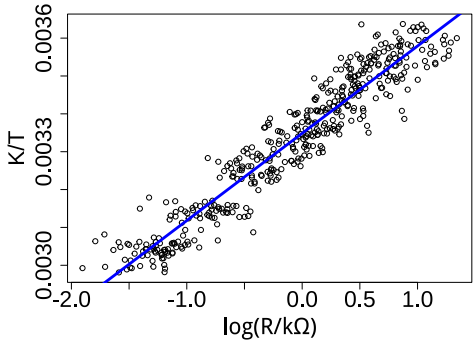
$$y = a + bx$$

- 加入更高阶项，可以描述更复杂的关系。

$$\frac{K}{T} = b_0 + b_1 \log \frac{R}{\mathrm{k}\Omega} + b_2 \left(\log \frac{R}{\mathrm{k}\Omega} \right)^2 + b_3 \left(\log \frac{R}{\mathrm{k}\Omega} \right)^3 \cdots$$

- 如何确定高阶项的系数？

热敏电阻的更精确模型



$$\frac{K}{T} = a + b \log \frac{R}{\text{k}\Omega}$$

$$y = a + bx$$

- 加入更高阶项，可以描述更复杂的关系。

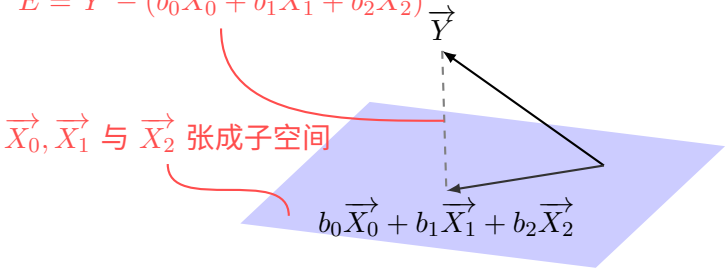
$$\frac{K}{T} = b_0 + b_1 \log \frac{R}{\text{k}\Omega} + b_2 \left(\log \frac{R}{\text{k}\Omega} \right)^2 + b_3 \left(\log \frac{R}{\text{k}\Omega} \right)^3 \cdots$$

- 如何确定高阶项的系数？

线性回归的推广

- 令 $\vec{X}_0 = \vec{1}, \vec{X}_1 = \vec{x}, \vec{X}_2 = \vec{x^2}$:

$$\vec{E} = \overrightarrow{Y} - (b_0\overrightarrow{X_0} + b_1\overrightarrow{X_1} + b_2\overrightarrow{X_2})$$



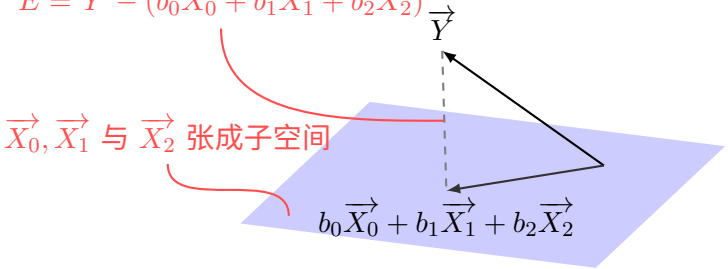
推广

- 替换投影子空间的基
- 填加新的基，构成更高维的子空间，实现多元线性回归。

线性回归的推广

- 令 $\vec{X}_0 = \vec{1}, \vec{X}_1 = \vec{x}, \vec{X}_2 = \vec{x^2}$:

$$\vec{E} = \overrightarrow{Y} - (b_0\overrightarrow{X_0} + b_1\overrightarrow{X_1} + b_2\overrightarrow{X_2})$$



推广

- 替换投影子空间的基
- 填加新的基，构成更高维的子空间，实现多元线性回归。

多元线性回归概念

- 一组观测值 $\vec{Y}$ 若与多组自变量 $\vec{X}_0, \vec{X}_1, \vec{X}_2, \cdots, \vec{X}_p$ 有关，

$$\vec{Y} = (\vec{X}_0, \vec{X}_1, \vec{X}_2, \cdots, \vec{X}_p) \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} + \vec{E}$$

- 可以使用最小二乘法得到 $\hat{b}_0, \hat{b}_1, \hat{b}_2, \cdots, \hat{b}_p$.

热敏电阻的多元线性回归

$$\frac{K}{T} = 3.326 \times 10^{-3} + 3.641 \times 10^{-3} \log \frac{R}{\mathrm{k}\Omega} - 3.624 \times 10^{-4} \left(\log \frac{R}{\mathrm{k}\Omega} \right)^3$$

系数	值	显著?
b_0	$3.326(2) \times 10^{-3}$	是
b_1	$3.641(51) \times 10^{-3}$	是
b_2	$-1.112(50730) \times 10^{-6}$	否
b_3	$-3.624(507) \times 10^{-4}$	是

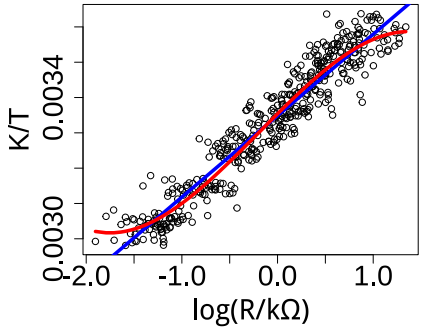
- 红色线对应 Starnhart-Hart 公式。
- 参数越多，拟合得越好，但模型越复杂。

Steinhart and Hart *Calibration curves for thermistors*, 1968

热敏电阻的多元线性回归

$$\frac{K}{T} = 3.326 \times 10^{-3} + 3.641 \times 10^{-3} \log \frac{R}{\mathrm{k}\Omega} - 3.624 \times 10^{-4} \left(\log \frac{R}{\mathrm{k}\Omega} \right)^3$$

系数	值	显著?
b_0	$3.326(2) \times 10^{-3}$	是
b_1	$3.641(51) \times 10^{-3}$	是
b_2	$-1.112(50730) \times 10^{-6}$	否
b_3	$-3.624(507) \times 10^{-4}$	是



- 红色线对应 Starnhart-Hart 公式。
- 参数越多，拟合得越好，但模型越复杂。

Steinhart and Hart *Calibration curves for thermistors*, 1968

例子：辐射热计量

- 每次有放射性衰变，二氧化碲 $\mathrm{TeO_2}$ 晶体都会“发烧”升温，

$$^{130}\mathrm{Te} \to {}^{130}\mathrm{Xe} + 2e^- + \underbrace{2\bar{\nu}_e}_{?}$$

- 通过给晶体“量体温”，确定中微子是否它自身的反粒子，以及中微子质量。

中子嬗变掺杂锗热敏电阻的线性回归

“体温计”为“中子嬗变掺杂（NTD）”的锗热敏电阻。由线性回归标定：

$$\sqrt{\frac{T_0}{T}} = a + b\log\left(\frac{R}{R_0}\right)$$

精度达到 μK，是一般热敏电阻的 1000 倍。

例子：辐射热计量

- 每次有放射性衰变，二氧化碲 $\mathrm{TeO_2}$ 晶体都会“发烧”升温，

$$^{130}\mathrm{Te} \to {}^{130}\mathrm{Xe} + 2e^- + \underbrace{2\bar{\nu}_e}_{?}$$

- 通过给晶体“量体温”，确定中微子是否它自身的反粒子，以及中微子质量。

中子嬗变掺杂锗热敏电阻的线性回归

“体温计”为“中子嬗变掺杂（NTD）”的锗热敏电阻。由线性回归标定：

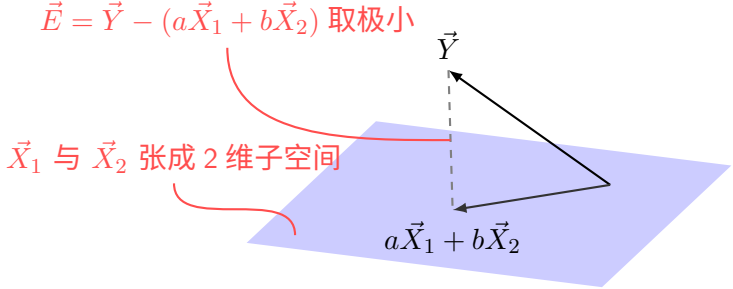
$$\sqrt{\frac{T_0}{T}} = a + b\log\left(\frac{R}{R_0}\right)$$

精度达到 μK，是一般热敏电阻的 1000 倍。

总结

线性回归

1. 线性回归是检验相关性的重要数理统计方法。
2. 最小二乘法是线性回归的解法，可看作线性空间的投影。



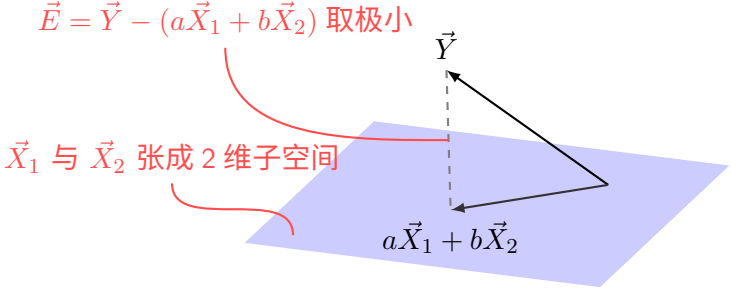
3. 假设残差来自正态总体，

$$\epsilon_i = y_i - (a + bx_i) \sim N(0, \sigma^2)$$

线性回归的结果需要经过 F 或 t 假设检验，确认回归的显著性，才能完成科学解读。

线性回归

1. 线性回归是检验相关性的重要数理统计方法。
2. 最小二乘法是线性回归的解法，可看作线性空间的投影。



3. 假设残差来自正态总体，

$$\epsilon_i = y_i - (a + bx_i) \sim N(0, \sigma^2)$$

线性回归的结果需要经过 F 或 t 假设检验，确认回归的显著性，才能完成科学解读。

推广：打破残差的正态分布假设

残差不服从正态分布时，以线性模型解决问题 $\rightarrow$ 广义线性回归。

- 泊松回归：观测量是计数
- 伽马回归：观测量是正实数
- 逻辑回归：观测量是 $[0,1]$ 区间内的实数
- 二项回归：观测是是 $0,1\cdots,N$ 的整数