




心理统计

第二十四讲：Multiple Linear Regression

严超轶

Chao-Gan Yan, Ph.D.

yancg@psych.ac.cn
http://rfmri.org/yan

Institute of Psychology, Chinese Academy of Sciences

1

主要内容

- 1 多元线性回归
- 2 变量选择
- 3 多元回归中系数的解读
- 4 标准化回归模型
- 5 基本假定和workflow

Courtesy of Dr. Chuan-Peng Hu @ NNU

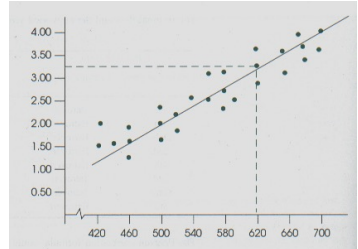
2

一元线性回归回顾

3

研究情境

- GRE与研究生GPA的关系



4

最佳拟合线

- 目标是使误差最小。即，这条线与所有的数据点最近，是最佳拟合线。
- 回归线是给定X，a 和 b，用公式 (线性方程)来预测Y 的值。我们的目标是找出一条线，以对Y作最佳估计。即，这条线使得所有Y 值的估计误差最小。

5

斜率和截距

- 最佳拟合线斜率的公式是：
 $b = SP/SS_X = r S_Y/S_X$
- 最佳拟合线截距的公式是：
 $a = Y - bX$

6

回归表达形式

- 简单线性回归 (Simple Linear Regression)

$$Y = \boxed{a + bX} + \epsilon$$

Data Model Error

Y - 因变量, Dependent variable

X - 自变量, Independent (explanatory) variable

a - 截距, Intercept

b - 斜率, Slope

ϵ - 残差, Residual (error)

7

矩阵表达

矩阵乘法 矩阵加法

$$Y_{n \times 1} = X_{n \times 2} \beta_{2 \times 1} + \epsilon_{n \times 1}$$

$$Y_{n \times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X_{n \times 2} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \quad \beta_{2 \times 1} = \begin{bmatrix} \alpha \\ \beta_1 \end{bmatrix} \quad \epsilon_{n \times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

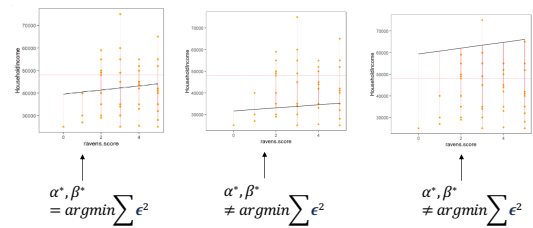
8

最小二乘法

- 要建立一元线性回归方程, 就要先计算方程中的参数 a 和 b 。根据最佳拟合原则, 回归线是指散点图中每一个点沿Y轴方向到该直线的距离平方和最小的那条直线, 即使使误差平方和最小, 这就是常规最小二乘法(ordinary least squares, OLS)的基本思想。

9

最小二乘法



10

基本假定

1. 模型设定假定 (线性预设)
2. 正交预设
3. 残差方差齐性预设
4. 正态分布预设

11

Part 1 多元线性回归

12

问题引入

- 收入与问题解决能力是什么关系？

13

问题引入

- 收入与问题解决能力是什么关系？
- 我们可以使用简单线性回归来解决这个问题。

14

问题引入

- **假如：**同时我们还想知道收入与风险偏好、自我控制是什么关系。

我们还可以使用简单线性回归模型吗？

15

问题引入

- 多个简单回归？
- 如何比较不同自变量相互之间的作用？如何比较不同变量对同一因变量的效应大小？

可能需要**多元线性回归**来回答这个问题。

16

多元线性回归

- 由于一个现象总是同时受到多个因素的影响。如果我们仅考虑个别因素对结果变量的影响，而忽略其他有关变量的影响，则回归模型的参数估计可能是有偏的。
- 而多元线性回归模型适用于分析一个因变量和多个自变量之间的关系。

17

多元线性回归

- 简单线性回归 (Simple Linear Regression)

$$Y = a + bX + \epsilon$$

Y – 因变量, Dependent variable
X – 自变量, Independent (explanatory) variable
a – 截距, Intercept
b – 斜率, Slope
 ϵ – 残差, Residual (error)

18

多元线性回归

- 多元线性回归 (Multiple Linear Regression)

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p + \epsilon$$

Y – 因变量, Dependent variable

X_i – 自变量, Independent (explanatory) variable

b_0 – 截距, Intercept

b_i – 斜率, Slope

ϵ – 残差, Residual (error)

19

多元线性回归

- 线性代数表达

$$y_i = a + b x_i + \epsilon$$

$$\begin{cases} Y_1 = \alpha + \beta_1 X_1 + \epsilon_1 \\ Y_2 = \alpha + \beta_1 X_2 + \epsilon_2 \\ \dots \\ Y_n = \alpha + \beta_1 X_n + \epsilon_n \end{cases}$$

20

多元线性回归

- 线性代数表达

$$y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_p X_{ip} + \epsilon_i$$

$$\begin{cases} Y_1 = b_0 + b_1 X_{11} + b_2 X_{12} + \dots + b_p X_{1p} + \epsilon_1 \\ Y_2 = b_0 + b_1 X_{21} + b_2 X_{22} + \dots + b_p X_{2p} + \epsilon_2 \\ \dots \\ Y_n = b_0 + b_1 X_{n1} + b_2 X_{n2} + \dots + b_p X_{np} + \epsilon_n \end{cases}$$

21

多元线性回归

- 矩阵表达

$$Y_{n \times 1} = X_{n \times 2} \beta_{2 \times 1} + \epsilon_{n \times 1}$$

$$Y = X \beta + \epsilon$$

$$Y_{n \times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X_{n \times 2} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \quad \beta_{2 \times 1} = \begin{bmatrix} \alpha \\ \beta_1 \end{bmatrix} \quad \epsilon_{n \times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

22

多元线性回归

- 矩阵表达

$$Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}$$

$$Y = X \beta + \epsilon$$

$$Y_i = b_0 + \left(\sum_{k=1}^p b_k X_{ik} \right) + \epsilon_i$$

$$Y_{n \times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X_{n \times p} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad \beta_{p \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{bmatrix}$$

$$\epsilon_{n \times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

23

多元线性回归

- 代码表达(r)

$$\text{"Formula"} = Y \sim X$$

Y : 因变量

X : 自变量

24

多元线性回归

- 代码表达(r)

$$\text{"Formula"} = Y \sim X_1 + X_2 + \dots + X_n$$

Y: 因变量

X_i : 自变量

25

多元线性回归

- 收入与问题解决能力是什么关系?

- 数据: “HouseholdIncome”(房产收入)和 “Ravens.score”(瑞文推理测验)这两个变量进行研究(Eisenberg et al., 2019)

(Eisenberg et al. 2019)

26

多元线性回归

- 收入与问题解决能力、风险偏好、自我控制是什么关系?

变量	意义
HouseholdIncome	收入
Ravens.score	瑞文推理测验分数
dospert_cb_survey.financial	财政方面的风险偏好
dospert_cb_survey.health_safety	健康方面的风险偏好
dospert_cb_survey.recreational	娱乐方面的风险偏好
dospert_cb_survey.social	社交方面的风险偏好
brief_self_control_survey.self_control	自我控制
self_regulation_survey.control	自我规范

(Eisenberg et al. 2019)

27

多元线性回归

- 若我们只取前4个样本的数据:

Y	X1	X2	X3	X4	X5	X6	X7
40000	2	2.5	3	1.7	4.2	53	101
19500	4	2	1	1.7	3.3	33	122
60000	3	4.2	3.2	1.8	3.8	40	102
81000	9	2	2.8	1	3	64	133

28

多元线性回归

- 用方程组表示为:

$$\begin{cases} 40000 = \beta_0 + 2 \times \beta_1 + 2.5 \times \beta_2 + 3 \times \beta_3 + 1.7 \times \beta_4 + 4.2 \times \beta_5 + 53 \times \beta_6 + 101 \times \beta_7 + \epsilon_1 \\ 19500 = \beta_0 + 4 \times \beta_1 + 2 \times \beta_2 + 1 \times \beta_3 + 1.7 \times \beta_4 + 3.3 \times \beta_5 + 33 \times \beta_6 + 122 \times \beta_7 + \epsilon_2 \\ 60000 = \beta_0 + 3 \times \beta_1 + 4.2 \times \beta_2 + 3.2 \times \beta_3 + 1.8 \times \beta_4 + 3.8 \times \beta_5 + 40 \times \beta_6 + 102 \times \beta_7 + \epsilon_3 \\ 81000 = \beta_0 + 9 \times \beta_1 + 2 \times \beta_2 + 2.8 \times \beta_3 + 1 \times \beta_4 + 3 \times \beta_5 + 64 \times \beta_6 + 133 \times \beta_7 + \epsilon_4 \end{cases}$$

29

多元线性回归

- 矩阵表示:

$$\begin{bmatrix} 40000 \\ 19500 \\ 60000 \\ 81000 \end{bmatrix} = \beta_0 + \beta_1 \begin{bmatrix} 2 \\ 4 \\ 3 \\ 9 \end{bmatrix} + \beta_2 \begin{bmatrix} 2.5 \\ 2 \\ 4.2 \\ 2 \end{bmatrix} + \beta_3 \begin{bmatrix} 3 \\ 1 \\ 3.2 \\ 2.8 \end{bmatrix} + \beta_4 \begin{bmatrix} 1.7 \\ 1.7 \\ 1.8 \\ 1 \end{bmatrix} + \beta_5 \begin{bmatrix} 4.2 \\ 3.3 \\ 3.8 \\ 3 \end{bmatrix} + \beta_6 \begin{bmatrix} 53 \\ 33 \\ 40 \\ 64 \end{bmatrix} + \beta_7 \begin{bmatrix} 101 \\ 122 \\ 102 \\ 133 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}$$

30

多元线性回归

- 简洁的矩阵形式

$$Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

31

设计矩阵

- 设计矩阵通常是用来描述数据集的一个方法。通常有m行n列，
 - 1、每个行向量 x_i 代表第i个实例
 - 2、每个列向量 x_j 代表第j个特征

32

设计矩阵

$$\begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & x_{16} & x_{17} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & x_{26} & x_{27} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & x_{n4} & x_{n5} & x_{n6} & x_{n7} \end{bmatrix}$$

intercept Raven financial health_safety recreational social self_control self-reg

33

设计矩阵

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{bmatrix}$$

intercept
coefficient of Raven
coefficient of financial
coefficient of health_safety
coefficient of recreational
coefficient of social
coefficient of self_control
coefficient of control

34

设计矩阵

$$\begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14} + \beta_5 x_{15} + \beta_6 x_{16} + \beta_7 x_{17} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \beta_4 x_{24} + \beta_5 x_{25} + \beta_6 x_{26} + \beta_7 x_{27} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \beta_4 x_{n4} + \beta_5 x_{n5} + \beta_6 x_{n6} + \beta_7 x_{n7} \end{bmatrix}$$

income

35

多元回归的参数估计

- 最小二乘法 (OLS)

36

Part 2 变量选择

37

模型设定

- 回归分析依赖于所设定的模型是正确的，模型参数估计和假设检验都建立在这一大前提之下。
- 在实际研究中，研究者通常根据某个理论或某些经验研究结果设定回归模型。
- 在模型设定中存在两类错误：
 - 模型中纳入了某些无关变量；
 - 模型中忽略了某些相关变量。

38

纳入无关自变量

- 回归分析中，在进行模型设定的时候，可能加入了无关的自变量。也就是说，尽管在总体中一个或多个自变量对因变量没有影响，但它们还是被纳入模型中。
- 选择的标准之一：
 - 若新加入的变量的回归系数不显著，可将其看作无关变量。

39

纳入无关自变量

- 无关变量的影响：
 - 若原方程为： $Y = b_0^* + b_1^*X_1 + b_2^*X_2 + \epsilon$
 - X_3 为无关变量： $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \epsilon$
- 纳入无关变量后，方程
 - (1) 回归系数 b_0, b_1, b_2 的方差变大，
 - (2) 模型自由度减小。

40

忽略相关自变量

- 如果在模型设定中忽略了某些本该纳入却未被纳入的有关自变量，可能有两种情况：
 - 1. 所忽略的变量与模型中的其他变量无关；
 - 2. 所忽略的变量与模型中的其他变量相关。
- 在前一种情形下，不会发生忽略变量偏误，后一种情形下，可能会发生忽略变量偏误。

41

忽略相关自变量

- 若真实方程为： $Y = b_0^* + b_1^*X_1 + b_2^*X_2 + b_3^*X_3 + \epsilon$
- 假如 X_3 被忽略： $Y = b_0 + b_1X_1 + b_2X_2 + \epsilon$
 - 在第一种情况下， X_3 变成误差项的一部分。由于 X_3 与 X_2, X_1 不相关，最小二乘估计无偏。
 - 在第二种情况下， X_3 变成误差项的一部分。由于 X_3 与 X_2, X_1 相关，误差项则不能保持独立，违反了最小二乘估计的基本假定。

42

多重共线性(multicollinearity)

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p + \epsilon$$

- 其基本假设之一就是解释变量间相互独立。
- 如果某两个或多个解释变量之间出现了相关性，则称为多重共线性。

43

多重共线性

- 在设计矩阵中，至少有一个变量可以表示为其他变量的线性组合。则变量之间存在多重共线性。即：

$$c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4 + \dots = 0$$

- 若 $\epsilon = 0$ ，则为完全共线性，但在现实情况并不常见，一般 $\epsilon \neq 0$ ，即近似共线性。

44

多重共线性

- 多重共线性的产生是由于变量之间具有相关关系。
- 若解释变量间存在多重共线性，利用最小二乘法推导出的方程没有唯一解。

45

多重共线性的影响

- 完全共线性下参数估计量不存在
- 近似共线性下OLS估计量非有效
- 参数估计量不合理
- 变量的显著性检验失去意义

46

多重共线性的诊断

- 容许度 $TOL_{x_k} = 1 - R_{x_k}^2$
其中， $R_{x_k}^2$ 为复相关系数，以 x_k 为因变量，其他自变量来预测 x_k 的新自变量进行回归，所得到的 R^2 。 $R_{x_k}^2$ 越接近1说明 x_k 能被其他自变量解释；
也就是说， $R_{x_k}^2$ 越接近1，则共线性越严重。

- 方差膨胀因子(variance inflation factor, VIF)
对于自变量 x_k ，

$$VIF_{x_k} = \frac{1}{TOL_{x_k}} = 1/(1 - R_{x_k}^2)$$

当多重共线性严重时， $R_{x_k}^2$ 越接近1， VIF_{x_k} 变大。

47

多重共线性的处理

- 如果完全共线性：直接删除不必要的变量
- 如果是近似共线性，处理依赖于经验
 - 整合解释变量（如降维的方法——主成分分析等）
 - 增加样本量

48

Part 3 多元回归中系数的解读

49

Part 3 多元回归中系数的解读

50

偏效应

$$\hat{y}_i = b_0 + b_1 x_a + b_2 x_b$$

b_0 : $x_a = 0$ 且 $x_b = 0$ 时, 对 y 的预测值

b_1 、 b_2 : 偏回归系数, 反映了各自自变量对 y 的偏效应 (partial effect), 即单独效应

- 偏效应: 其他自变量的值保持不变时 (例如, 等于 0), 变化某个自变量对因变量产生的影响

51

偏效应

$$\hat{y}_i = b_0 + b_1 x_a + b_2 x_b$$

- 偏效应: 其他自变量的值保持不变时 (例如, 等于 0), 变化某个自变量对因变量产生的影响。

$$\Delta \hat{y}_i = b_1 \times \Delta x_{i1} + b_2 \times \Delta x_{i2}$$

$$\Delta x_{i2} = 0 \longrightarrow \Delta \hat{y}_i = b_1 \times \Delta x_{i1} + b_2 \times \Delta x_{i2}$$

52

偏效应反映了什么?

- 如果我们觉得一个变量和另一个变量是有因果关系的, 两个变量之间的关系必须满足三个标准:
 - (1) 变量之间存在关联
 - (2) 变量的适当时间顺序
 - (3) 排除产生这种关联的其他解释
- 在实际研究中, 我们要正确看待不同变量间的关系, 避免错误地将相关关系解释为因果关系。

53

相关关系和因果关系

- 在观察研究中, 由于所有变量都是事后调查得到的, 我们在看待变量之间的影响时, 可能会犯以下错误。

54

混淆 (confounding)

- 各种解释变量之间存在相关关系，无法确定哪个变量对结果产生了影响。
- 例如，研究发现，温度越高，泳池溺死人数越多。但我们不能确定泳池溺死人数是否是受温度的影响。与温度相关的其他变量，比如进入泳池的人数等可能也会影响到泳池溺死人数。

55

虚假关联(spurious association)

- 虚假关联是混淆的特例，两个变量之间的关联是由于第三个变量对两个各自产生了影响。
- 例如，研究发现冷饮销量越多，去泳池的人数越多。这两者之间没有直接联系，这两个变量都是由于温度升高导致，由于温度升高导致买冷饮的人变多，由于温度升高导致去泳池的人变多。

56

条件独立(conditionally independent)

- 给定第三个事件，如果 $p(X,Y|Z) = p(X|Z)p(Y|Z)$ ，则称X和Y是条件独立事件，也就是说，当Z发生时，X发生与否与Y发生与否是无关的。
- 例如，研究发现冷饮销量越多，去泳池的人数越多。冷饮的销量和去泳池的人数这两个变量就是条件独立的，两者之间没有直接关联，都是由温度这一变量导致的。

57

小结

- 变量之间存在强烈的相关关系，我们错误地将不合适的变量纳入了回归方程中，有可能得出荒谬的结论。
- 如果这些变量被同时纳入回归方程，就表现为多重共线性。如果只有某个变量纳入方程，在对变量进行解释时，一定要关注它的实际意义。
- 在对变量进行选择时除了关注模型本身的指标，还要关注变量的实际意义。

58

Part 4 标准化回归模型

59

标准化回归模型

- 回归模型中涉及的自变量具有不同的测量单位，因此回归系数之间不具有直接的可比性。
- 采用标准化回归系数，将自变量转变为无量纲的变量。

60

标准化回归模型

- 设真实的回归模型为：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i(p-1)} + \varepsilon_i$$

- 所有变量减去其均值（改变原始变量位置，转换后变量的均值为0）
- 再除以其标准差（测度转换，转换后变量的方差为1）
- 得到所有变量的标准Z值

61

标准化回归模型

- 设真实的回归模型为：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i(p-1)} + \varepsilon_i$$

- 得到所有变量的标准Z值

$$y_i^* = \frac{y_i - \bar{y}}{S_y}$$

$$x_{ik}^* = \frac{x_{ik} - \bar{x}_k}{S_{xk}}$$

62

标准化回归模型

- 设真实的回归模型为：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i(p-1)} + \varepsilon_i$$

- 标准化的回归模型

$$\rightarrow y_i^* = \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \dots + \beta_{p-1}^* x_{i(p-1)}^* + \varepsilon_i^*$$

（常数项为0，没有截距项）

63

标准化回归模型

- 非标准化系数与标准化系数的关系

$$y_i - \bar{y} = \beta_1 (x_{i1} - \bar{x}_1) + \beta_2 (x_{i2} - \bar{x}_2) + \dots + \beta_{p-1} (x_{i(p-1)} - \bar{x}_{p-1}) + (\varepsilon_i - \bar{\varepsilon})$$

除以样本量的标准差 S_y

$$\frac{y_i - \bar{y}}{S_y} = \frac{\beta_1}{S_y} (x_{i1} - \bar{x}_1) + \frac{\beta_2}{S_y} (x_{i2} - \bar{x}_2) + \dots + \frac{\beta_{p-1}}{S_y} (x_{i(p-1)} - \bar{x}_{p-1}) + \frac{\varepsilon_i - \bar{\varepsilon}}{S_y}$$

$$= \beta_1^* \left(\frac{x_{i1} - \bar{x}_1}{S_{x1}} \right) + \beta_2^* \left(\frac{x_{i2} - \bar{x}_2}{S_{x2}} \right) + \dots + \beta_{p-1}^* \left(\frac{x_{i(p-1)} - \bar{x}_{p-1}}{S_{x(p-1)}} \right) + \frac{\varepsilon_i - \bar{\varepsilon}}{S_y}$$

$$= \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \dots + \beta_{p-1}^* x_{i(p-1)}^* + \varepsilon_i^*$$

$$\rightarrow \beta_k^* = \beta_k \frac{S_{xk}}{S_y}$$

64

标准化回归模型

- 非标准化系数与标准化系数的关系

- 标准化系数：

- 回归参数估计值处于 (-1,1)
- 可在标准化尺度上进行比较

- 非标准化系数：

- 提供更多关于数据的信息
- 提供基于实际单位的自变量对因变量的效应

65

Part 5 基本假定和workflow

66

多元线性回归的模型预设(OLS)

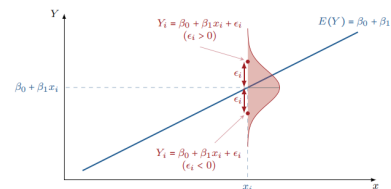
1. 模型设定假定(线性预设)
2. 正交预设
3. 残差方差齐性预设
4. 残差正态分布预设
5. 变量间不存在多重共线性

67

线性预设

- Y的条件均值是所有自变量X的线性函数

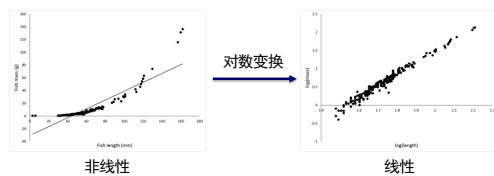
$$E(y|X) = X\beta$$



68

线性预设

- Y的条件均值是所有自变量X的线性函数



69

正交预设

- 误差项 ϵ 和 x 不相关
 - $COV(X, \epsilon) = 0$
 - 等价于 $E(\epsilon) = 0$ 和 $E(X'\epsilon) = 0$
 - 确保对模型参数的最小二乘法估计是无偏的。

70

残差方差齐性

$$Var(\epsilon_i) = \sigma_i^2 = \sigma^2$$

- 残差的方差不受X变量取值的影响。
- 若残差方差不齐性，误差较大的观测值将对拟合模型产生更大的影响。

71

正态分布预设

- 该预设规定残差项 ϵ 独立且同分布。

$$COV(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$$

- 实际中无法确定 ϵ 的分布。
 - 对于大样本数据，可根据中心极限定理对参数进行统计推断。然而在小样本情况下，我们只有假定 ϵ 服从正态分布时才能使用t检验。

72

变量间不存在多重共线性

- 多重共线性：线性回归模型中的解释变量之间存在相关关系。

73

WorkFlow

- 例子
 - 因变量：
 - HouseholdIncome (家庭收入)
 - “自变量”：
 - 智力：瑞文推理测验；
 - 风险偏好：道德、金融风险、健康风险、娱乐风险、社交风险
 - 自我控制
 - 自我调节

74

WorkFlow

- 先前的workflow:
 1. 提出与研究假设对应的统计假设
 2. 选择统计模型
 3. 确定显著性水平
 4. 基于 H_0 的统计模型，计算统计量
 5. 做出统计推断

75

WorkFlow



哪个变量应该进入模型？
自变量如何影响到我们假设？

76

1.提出与研究假设对应的统计假设

- 回归分析中的两个假设检验
 - 模型整体检验
 - $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$
 - $H_1: \beta_i$ 不全为0(至少有一个自变量与因变量之间存在显著性的线性关系)
 - 偏回归系数的显著性检验
 - H_0 : 单个回归系数 $\beta_k = 0$
 - H_1 : 单个回归系数 $\beta_k \neq 0$

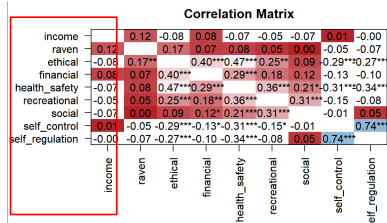
77

了解数据

	income	raven	ethical	financial	health_safety	recreational	social	self_control	self_regulation
Valid	521	521	521	521	521	521	521	521	521
Mean	48065.161	9.438	2.212	2.993	1.610	2.893	3.884	45.626	119.030
Std. Deviation	31279.537	4.299	0.862	1.075	0.588	1.150	0.881	10.266	19.850
Skewness	0.902	0.979	0.954	0.963	0.875	0.973	0.992	0.986	0.962
Shapiro-Wilk	< .001	< .001	< .001	< .001	< .001	< .001	0.005	< .001	< .001
P-value of Shapiro-Wilk	0.000	0.000	1.000	1.000	1.000	1.000	1.000	18.000	51.000
Minimum	240000.000	16.000	5.333	6.167	4.667	6.500	7.000	65.000	150.000
Maximum									

78

了解数据



- 我们希望找到几个自变量能够用于预测收入。
- 由图可见，瑞文推理测验分数与收入相关程度较其他变量更高。自我控制与自我调节相关系数高达0.74，后续同时选择这两个变量可能存在多重共线性。

79

检验线性回归的预设

- 检验如下预设：
 - 模型设定假定 (线性预设)
 - 正交预设
 - 残差方差齐性
 - 残差正态分布
 - 不存在多重共线性

80

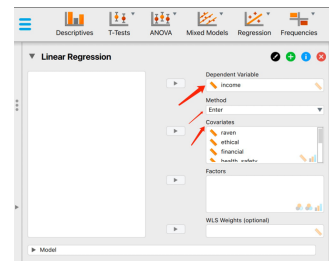
检验线性回归的预设

- 检验如下预设：
 - 模型设定假定 (线性预设)
 - 正交预设
 - 残差方差齐性
 - 残差正态分布
 - 不存在多重共线性
- 线性模型中，预设的检验与模型需要首先确定模型 (依赖于残差)

81

线性假定&正交

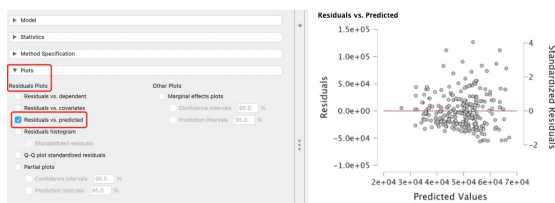
- 全模型: $\text{income} \sim \text{raven} + \text{financial} + \text{health_safety} + \text{recreational} + \text{social} + \text{self_control} + \text{self_regulation}$
- data: example24.csv



82

线性假定&正交

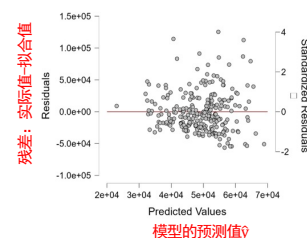
- 全模型: $\text{income} \sim \text{raven} + \text{financial} + \text{health_safety} + \text{recreational} + \text{social} + \text{self_control} + \text{self_regulation}$



83

线性假定&正交

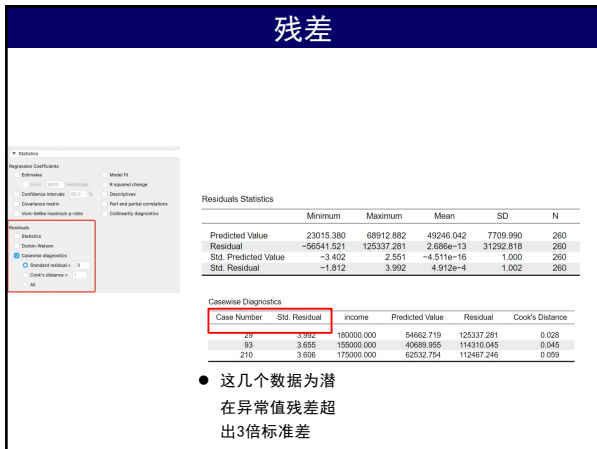
- 全模型: $\text{income} \sim \text{raven} + \text{financial} + \text{health_safety} + \text{recreational} + \text{social} + \text{self_control} + \text{self_regulation}$



- 参考线接近水平，符合线性预设

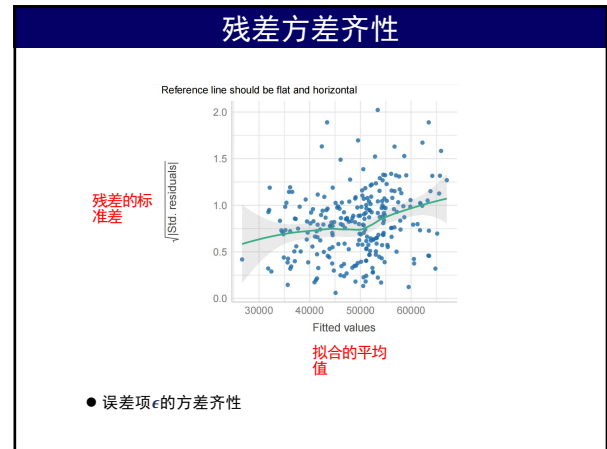
84

残差



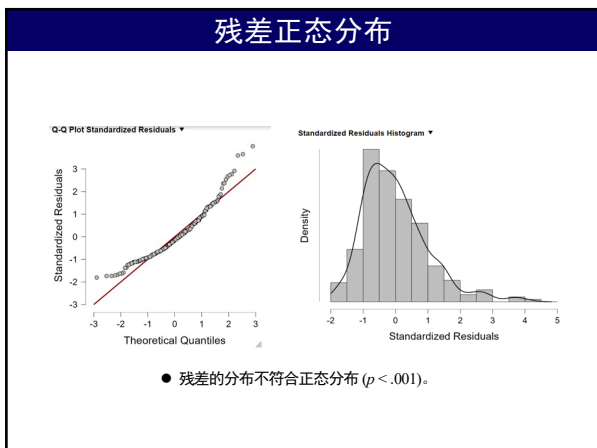
85

残差方差齐性



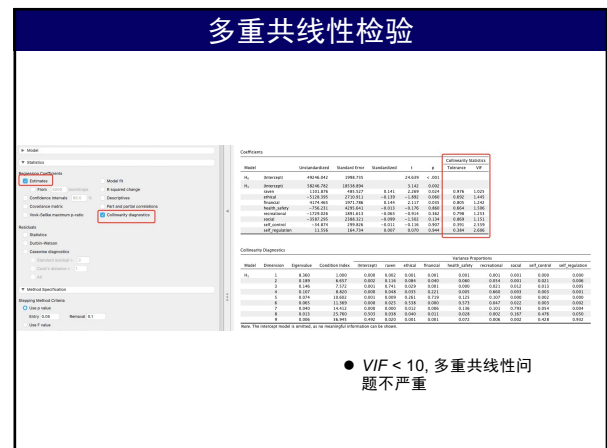
86

残差正态分布



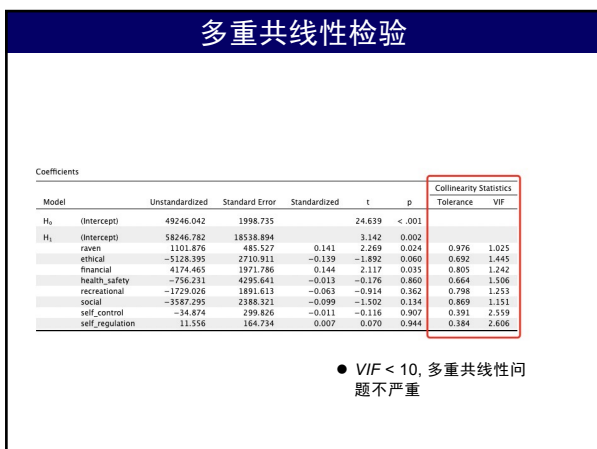
87

多重共线性检验



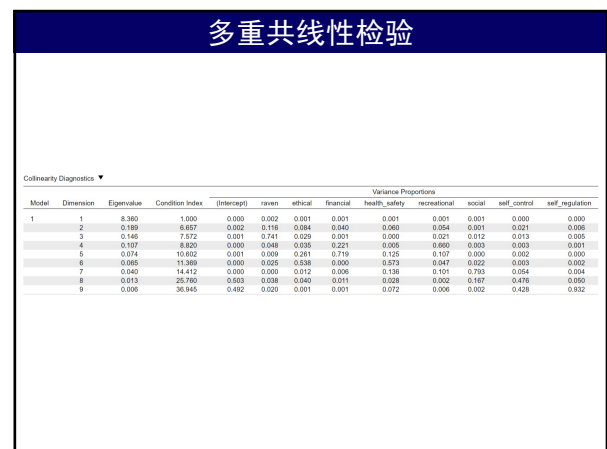
88

多重共线性检验



89

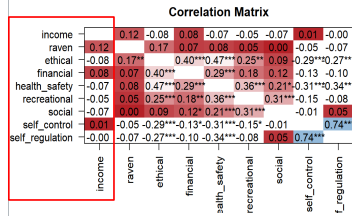
多重共线性检验



90

2.选择统计模型

- 向前选择法
 - 由图可见，瑞典推理测验分数与收入相关程度较其他变量更高。



91

3.确定显著性水平

- 确定显著性水平
 - 使用传统的 $\alpha = .05$

92

4. 计算统计量

- 模型1: $\text{income} \sim \text{raven}$
 - 模型整体检验:

Model Summary - income

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p
1	0.000	0.000	0.000	32228.627	0.000		0	259	
2	0.132	0.017	0.013	32010.507	0.017	4.542	1	258	0.034

93

4. 计算统计量

- 模型1: $\text{income} \sim \text{raven}$

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	95% CI	
							Lower	Upper
1	(Intercept)	49246.042	1998.735		24.639	< .001	45310.203	53181.882
2	(Intercept)	39471.725	4997.687		7.898	< .001	29630.312	49313.138
	raven	1031.361	483.961	0.132	2.131	0.034	78.364	1984.398

该模型在统计上解释了一小部分但统计上显著的变异:

$R^2=0.132$, $t=2.13$, $p=0.034$, $\text{adj. } R^2=0.013$

94

4. 计算统计量

- 增加ethical (道德风险偏好水平), financial (金融风险偏好水平)
进入回归模型

95

4. 计算统计量

- 回归系数检验

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	95% CI		Collinearity Statistics	
							Lower	Upper	Tolerance	VIF
1	(Intercept)	41584.919	7833.830		5.449	< .001	26561.621	56628.218		
	raven	1132.009	483.394	0.144	2.342	0.020	180.075	2083.944	0.983	1.017
	ethical	-6079.186	2458.000	-0.165	-2.473	0.014	-10918.674	-1238.722	0.940	1.190
	financial	3556.425	1916.926	0.123	1.855	0.065	-218.527	7331.377	0.851	1.178

Note. The following covariates were considered but not included: health_safety, recreational, social, self_control, self_regulation.

保留ethical

可能考虑剔除
financial

96

4. 计算统计量

● 模型整体检验:

Model Summary - income ▼

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p	Durbin-Watson		
										Autocorrelation	Statistic	p
1	0.145	0.021	0.013	32013.304	0.021	2.748	2	257	0.066	0.067	1.867	0.277
2	0.209	0.044	0.033	31699.293	0.023	6.117	1	256	0.014	0.055	1.889	0.364

Note. Null model includes raven, financial

增加ethical后Adj. R²变大, 保留

97

4. 计算统计量

● 模型整体检验:

ANOVA ▼

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	5.632e+9	2	2.816e+9	2.748	0.066
	Residual	2.634e+11	257	1.025e+9		
	Total	2.690e+11	259			
2	Regression	1.178e+10	3	3.926e+9	3.907	0.009
	Residual	2.572e+11	256	1.005e+9		
	Total	2.690e+11	259			

Note. Null model includes raven, financial

98

4. 计算统计量

● 剔除financial后

Model Summary - income ▼

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p	Durbin-Watson		
										Autocorrelation	Statistic	p
1	0.176	0.031	0.023	31849.543	0.031	4.101	2	257	0.018	0.061	1.878	0.322

Note. Null model includes raven, ethical

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	8.320e+9	2	4.160e+9	4.101	0.018
	Residual	2.607e+11	257	1.014e+9		
	Total	2.690e+11	259			

Note. Null model includes raven, ethical

不含financial时Adj. R²小, 故保留financial

99

4. 计算统计量

- 增加recreational (娱乐风险偏好水平), social (社交风险偏好水平) 进入回归模型

100

4. 计算统计量

● 模型整体检验:

Model Summary - income

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p	Durbin-Watson		
										Autocorrelation	Statistic	p
1	0.248	0.081	0.042	31542.284	0.061	3.279	5	254	0.007	0.062	1.875	0.308

Note. Null model includes raven, ethical, financial, recreational, social

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.631e+10	5	3.262e+9	3.279	0.007
	Residual	2.527e+11	254	9.949e+8		
	Total	2.690e+11	259			

Note. Null model includes raven, ethical, financial, recreational, social

101

4. 计算统计量

● 回归系数检验

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	95% CI		Collinearity Statistics	
							Lower	Upper	Tolerance	VIF
1	(Intercept)	57511.584	10987.513		5.234	< .001	35873.352	79149.817		
	raven	1105.443	481.395	0.141	2.296	0.022	157.408	2053.478	0.981	1.019
	ethical	-5245.148	2516.033	-0.142	-2.085	0.038	-10200.092	-290.205	0.794	1.259
	financial	4714.657	1925.343	0.142	2.137	0.034	322.798	7998.127	0.835	1.188
	recreational	-1768.085	1843.964	-0.064	-0.959	0.339	-5399.472	1863.341	0.830	1.204
	social	-3629.331	2343.767	-0.100	-1.549	0.123	-8246.023	986.362	0.892	1.121

Note. The following covariates were considered but not included: health_safety, self_control, self_regulation

剔除recreational

102

4. 计算统计量

● 回归系数检验

Model Summary - income

Model	R	R ²	Adjusted R ²	RMSE
H ₀	0.239	0.057	0.042	31537.297
H ₁	0.246	0.061	0.042	31542.284

Note. Null model includes raven, ethical, financial, social

剔除recreational

103

4. 计算统计量

Model Summary - income

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p	Durbin-Watson	Statistic	p
1	0.239	0.057	0.039	31599.315	0.057	3.084	5	254	0.010	0.063	1.874	0.303

Note. Null model includes raven, ethical, financial, social, self_control

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.540e+10	5	3.079e+9	3.084	0.010
	Residual	2.530e+11	254	9.985e+8		
	Total	2.690e+11	259			

Note. Null model includes raven, ethical, financial, social, self_control

增加self control, adj. R²减小, 故剔除该变量

104

4. 计算统计量

Model Summary - income

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p	Durbin-Watson	Statistic	p
1	0.239	0.057	0.039	31599.315	0.057	3.084	5	254	0.010	0.063	1.874	0.301

Note. Null model includes raven, ethical, financial, social, self_regulation

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.540e+10	5	3.080e+9	3.084	0.010
	Residual	2.536e+11	254	9.985e+8		
	Total	2.690e+11	259			

Note. Null model includes raven, ethical, financial, social, self_regulation

增加self regulation, adj. R²减小, 故剔除该变量

105

4. 计算统计量

● 模型整体检验:

Model Summary - income

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p	Durbin-Watson	Statistic	p
1	0.239	0.057	0.042	31537.297	0.057	3.870	4	255	0.005	0.063	1.874	0.301

Note. Null model includes raven, ethical, financial, social

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.540e+10	4	3.849e+9	3.870	0.005
	Residual	2.536e+11	255	9.946e+8		
	Total	2.690e+11	259			

Note. Null model includes raven, ethical, financial, social

106

4. 计算统计量

● 回归系数检验

Coefficients

					95% CI		Collinearity Statistics		
Model	Unstandardized	Standard Error	Standardized	t	p	Lower	Upper	Tolerance	VIF
1	(Intercept)	56630.147	10947.259	5.173	< .001	35071.595	78188.700		
	raven	1097.920	481.255	0.140	2.281	150.179	2045.661	0.982	1.019
	ethical	-5792.386	2450.059	-0.157	-2.364	-10617.314	-967.458	0.837	1.194
	financial	3987.299	1920.467	0.138	2.076	205.304	7769.294	0.839	1.192
	social	-4276.434	2243.517	-0.118	-1.907	-8696.615	139.747	0.973	1.028

Note. The following covariates were considered but not included: health_safety, recreational, self_control, self_regulation.

107

5. 统计推断（做出决策）

- 使用简单线性模型检验, 发现选择“Ravens.score”(瑞文推理测验), ethical (道德风险偏好水平), financial (金融风险偏好水平), social (社交风险偏好水平) 预测家庭收入效果较好。结果表明该模型显著地解释了一部分变异, $R^2=0.057$, $F(4, 259) = 3.87$, $p = 0.005$, adj. $R^2 = 0.042$ 。

108