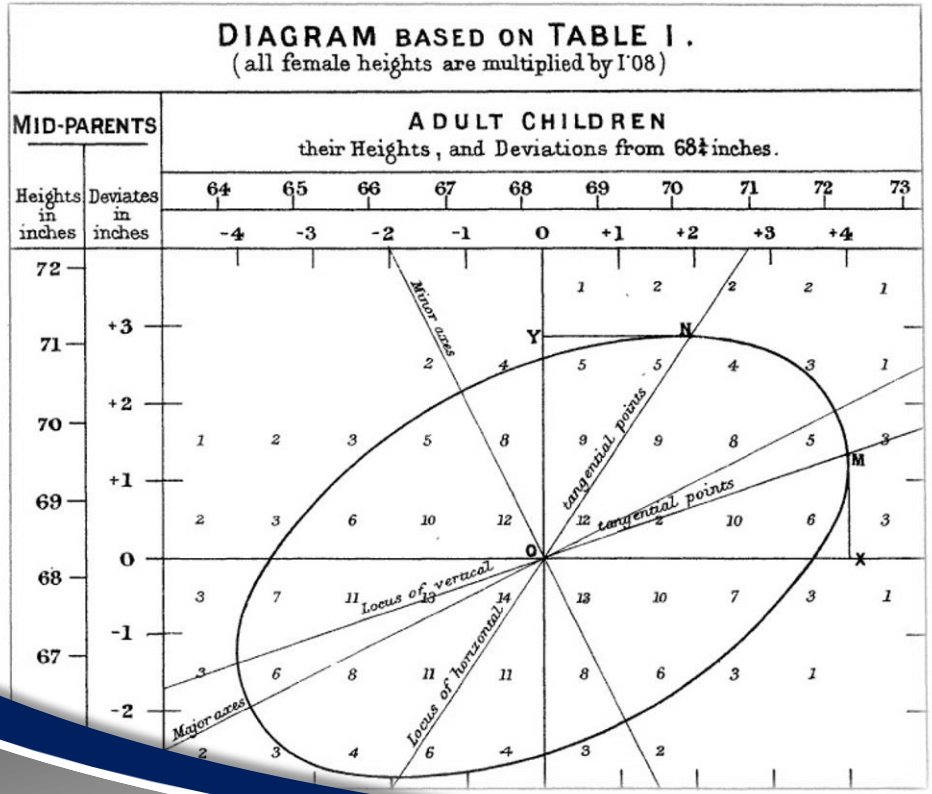
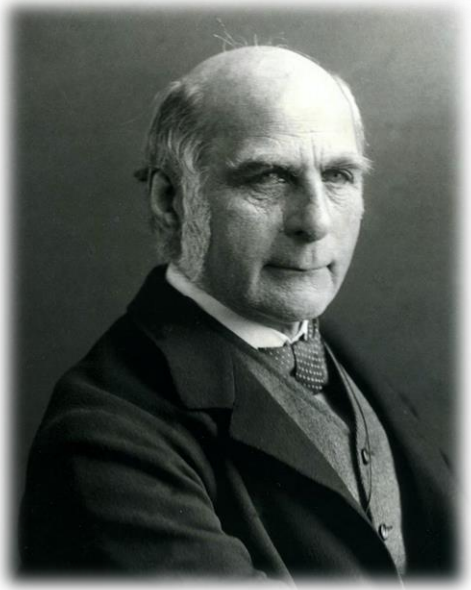




线性回归

城市分析方法系列课程

苏州大学 王灿

大纲

- 线性回归的基本原理
- 虚拟变量
- 多重共线性
- 残差分析与回归诊断

什么是回归

什么是回归

Regression to the mediocrity (1877/2/9)：Galton发现，相比于父母，子女的身高会出现向中心“回归”的趋势，而非两极分化，他试图为此提供**因果**解释。

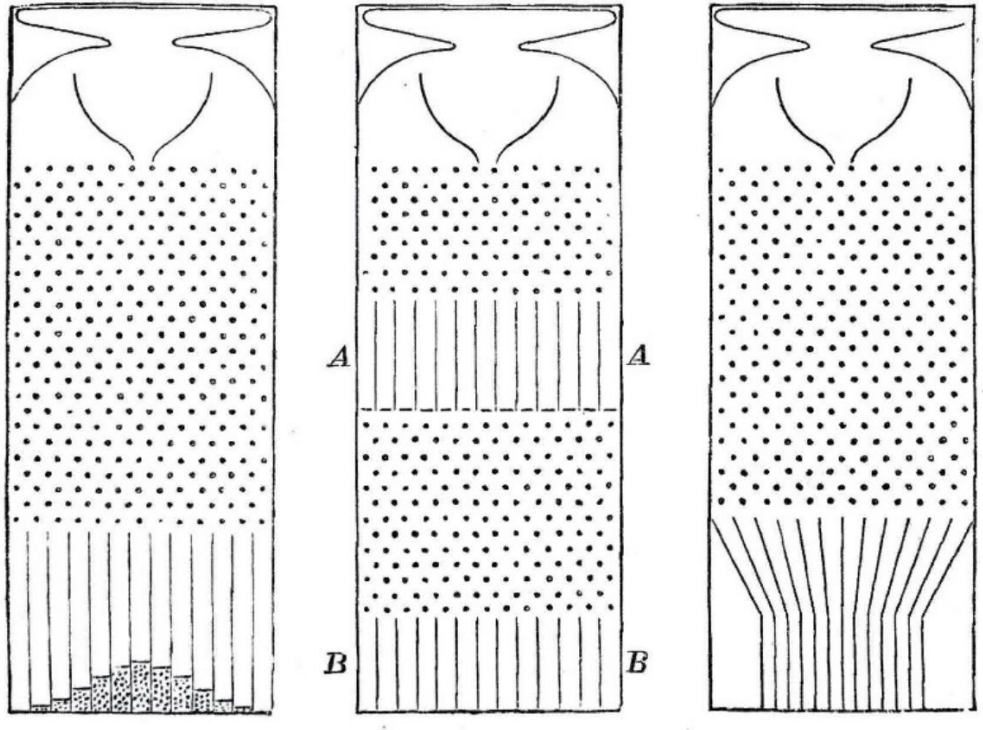


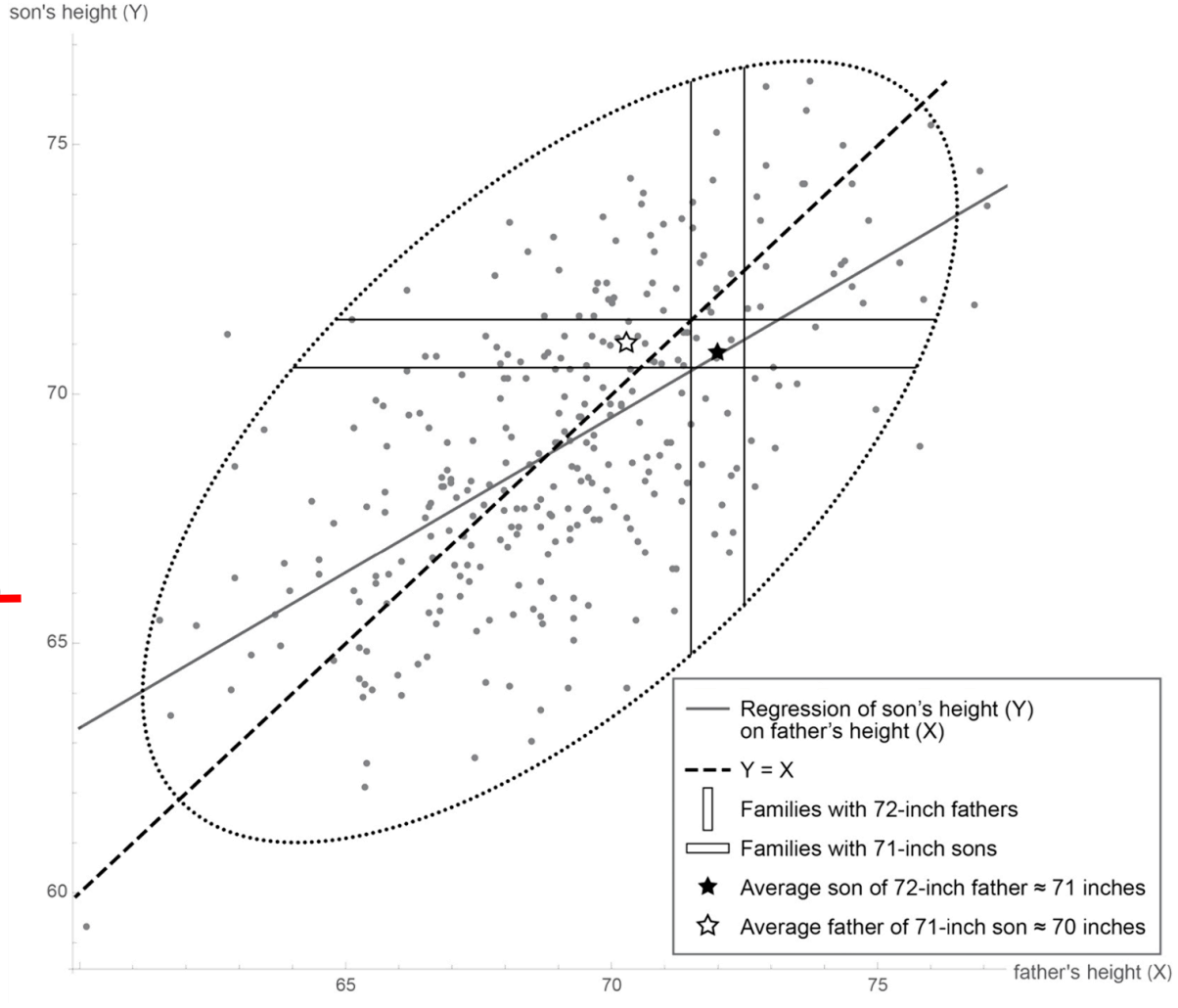
FIGURE 2.1. The Galton board, used by Francis Galton as an analogy for the inheritance of human heights. (a) When many balls are dropped through the pinball-like apparatus, their random bounces cause them to pile up in a bell-shaped curve. (b) Galton noted that on two passes, *A* and *B*, through the Galton board (the analogue of two generations) the bell-shaped curve got wider. (c) To counteract this tendency, he installed chutes to move the “second generation” back closer to the center. The chutes are Galton’s causal explanation for regression to the mean.

The Book of Why: The New Science of Cause and Effect

什么是回归

Disappointed but also fascinated divorce from causation (1886)

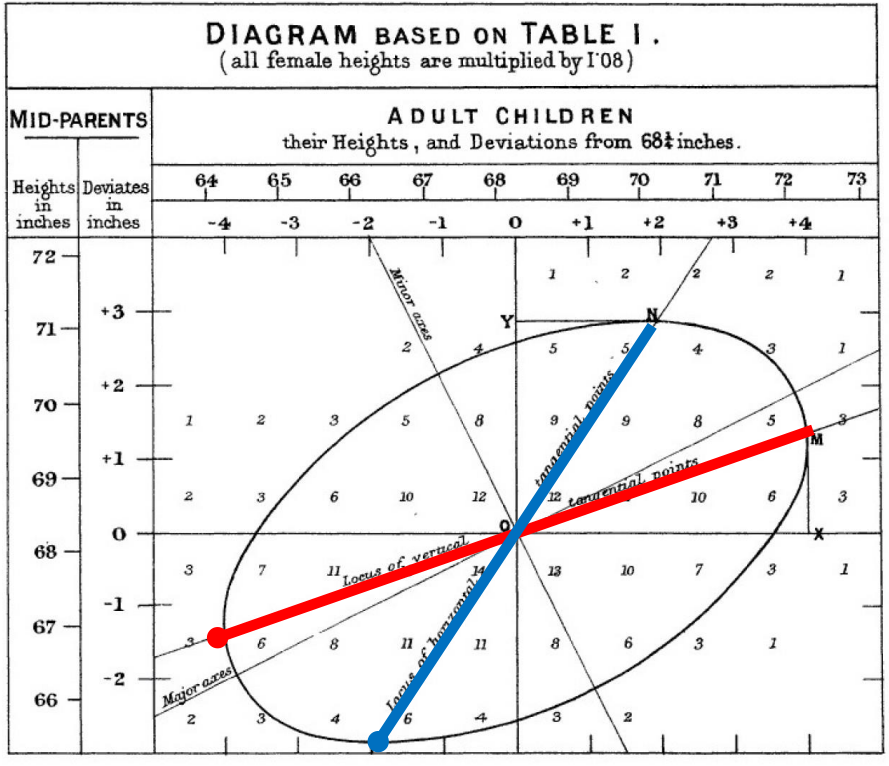
- 不仅子代的身高会向均值回归，父代的身高也会向均值回归，而后者显然不能用因果来解释。
- ~~儿子既比父亲高，又比父亲矮？~~
- 对于身高是 x 的父辈，其子辈身高的最佳预测是 y ？
  对于身高是 x 的子辈，其父辈身高的最佳预测是 y ？



什么是回归

Regression lines：已知一个变量，回归斜率可以帮助预测另一个变量的值。

但是，回归斜率不涉及因果信息！



$$H_{son} = \alpha_1 + 0.5H_{father}$$

$$H_{father} = \alpha_2 + 0.5H_{son}$$

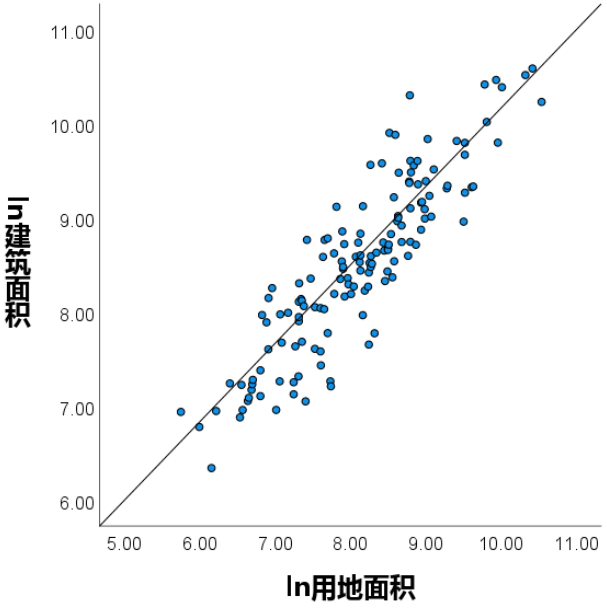
父亲的身高每增加1inch，儿子的身高增加0.5inch。

儿子的身高每增加1inch，父亲的身高增加0.5inch。

FIGURE 2.3. Galton's regression lines. Line OM gives the best prediction of a son's height if you know the height of the father; line ON gives the best prediction of a father's height if you know the height of the son. Neither is the same as the major axis (axis of symmetry) of the scatter plot. (*Source:* Francis Galton, *Journal of the Anthropological Institute of Great Britain and Ireland* [1886], 246–263, Plate X.)

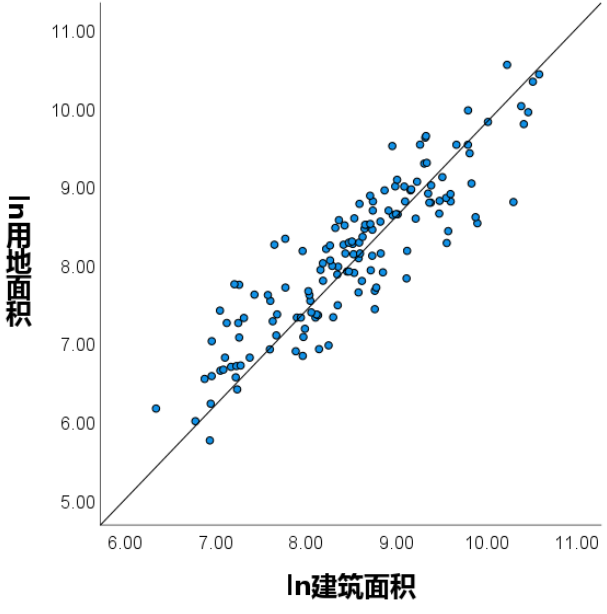
从“相关”到“回归”

用地面积 Vs 建筑面积



$r = 0.890$

$y = 1.667 + 0.844x$



$r = 0.890$

$y = 0.117 + 0.939x$

- 相关分析中的变量地位相同；回归分析则是通过**自变量**对**因变量**进行解释预测；
- 回归分析通常假设了因果关系，但它与相关分析一样，本质上**只能提示相关关系**；
- 相关分析仅关注强度和方向；回归分析则生成**回归方程**。

线性回归

$$y = \hat{y} + \varepsilon$$
$$= a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon$$

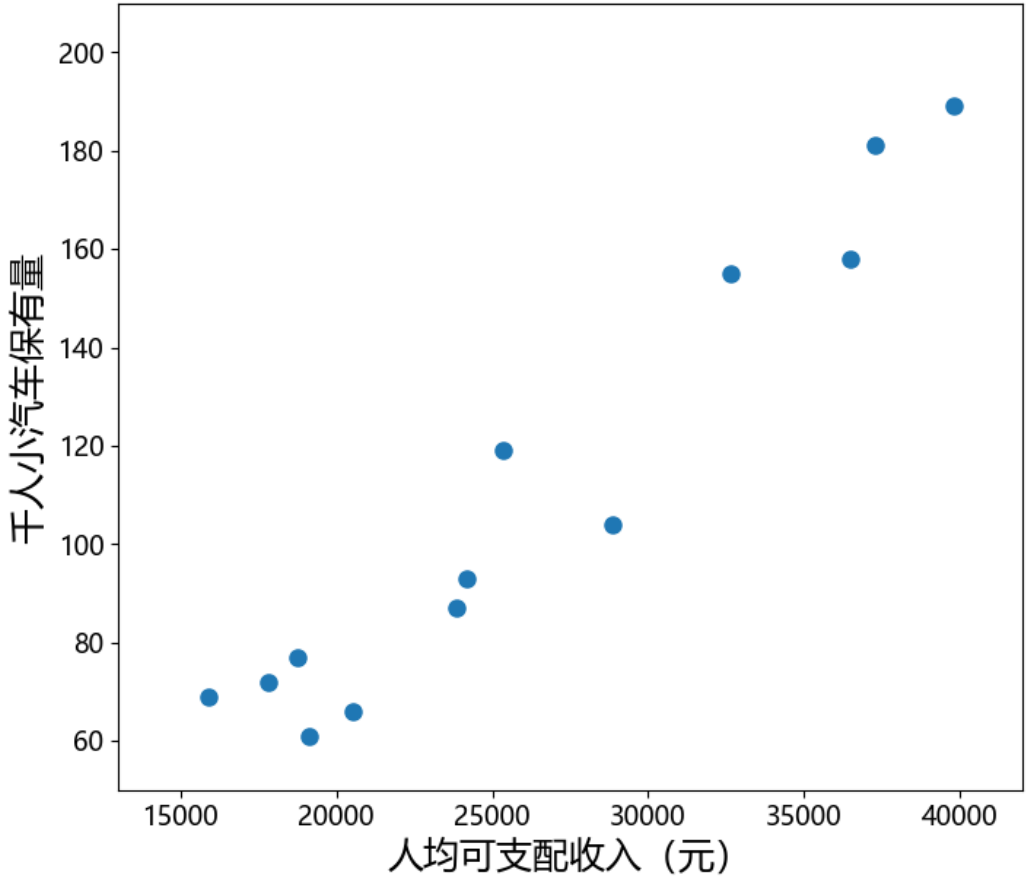
- y：因变量
- $x_1 \sim x_k$：k个自变量，或称解释变量
- a，$b_1 \sim b_k$：回归系数，其中a又称为截距或常数项，也写作b_0
- $\hat{y}$：因变量的估计值
- ε：残差（residual）

从一元线性回归看
线性回归的基本原理

$$y = a + bx + \varepsilon$$

一元线性回归

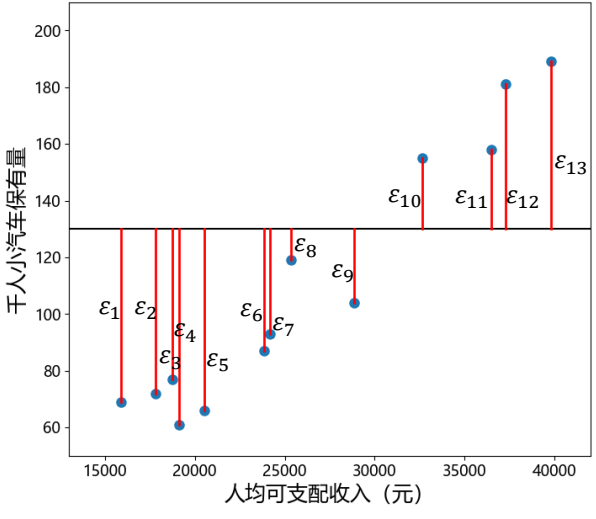
2014年江苏省各地级市的千人小汽车保有量和人均可支配收入



城市	千人小汽车保有量	人均可支配收入（元）
南京	181	37283
无锡	158	36471
徐州	77	18744
常州	155	32662
苏州	189	39780
南通	119	25340
连云港	72	17798
淮安	61	19110
盐城	66	20543
扬州	93	24157
镇江	104	28850
泰州	87	23833
宿迁	69	15888

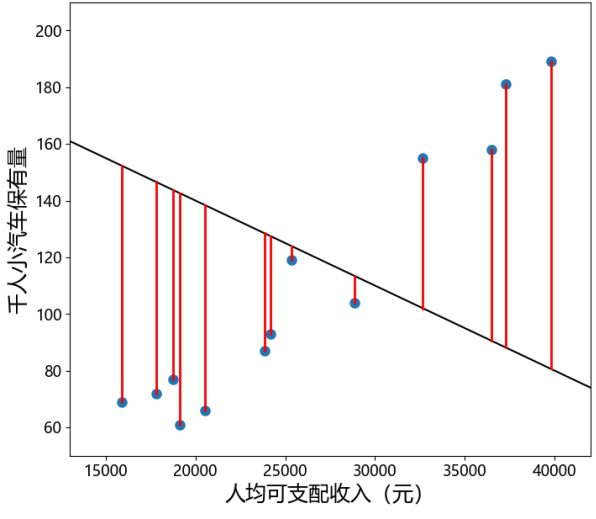
一元线性回归

什么是最优回归线?



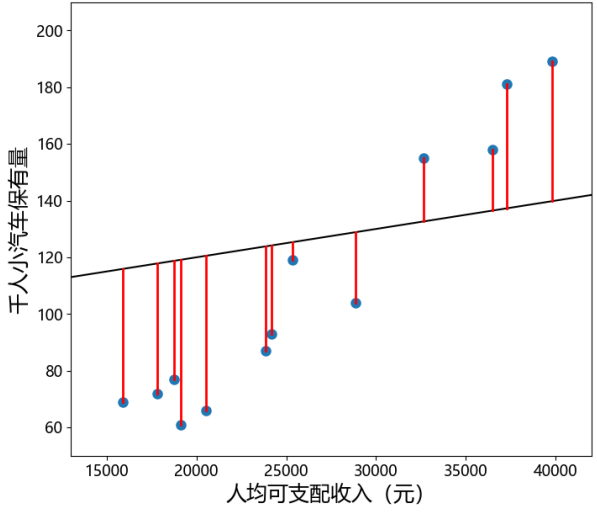
$$y = 130 + 0x$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = 30257$$



$$y = 200 - 0.003x$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = 59616$$



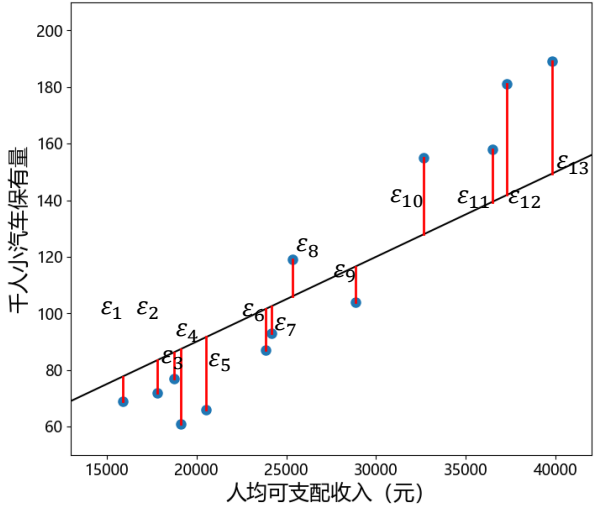
$$y = 100 + 0.001x$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = 20672$$

y：千人小汽车保有量　　x：人均可支配收入（元）

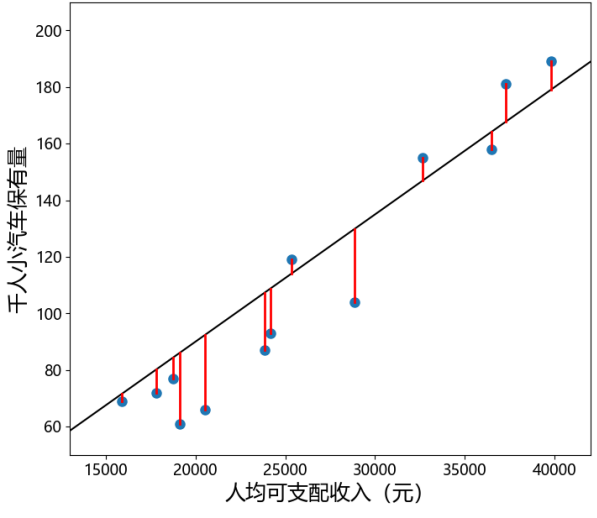
一元线性回归

什么是最优回归线?



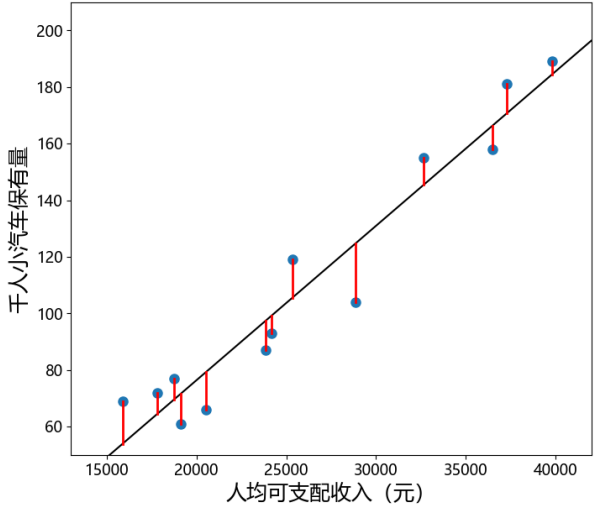
$$y = 30 + 0.003x$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = 6447$$



$$y = 0 + 0.0045x$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = 3175$$



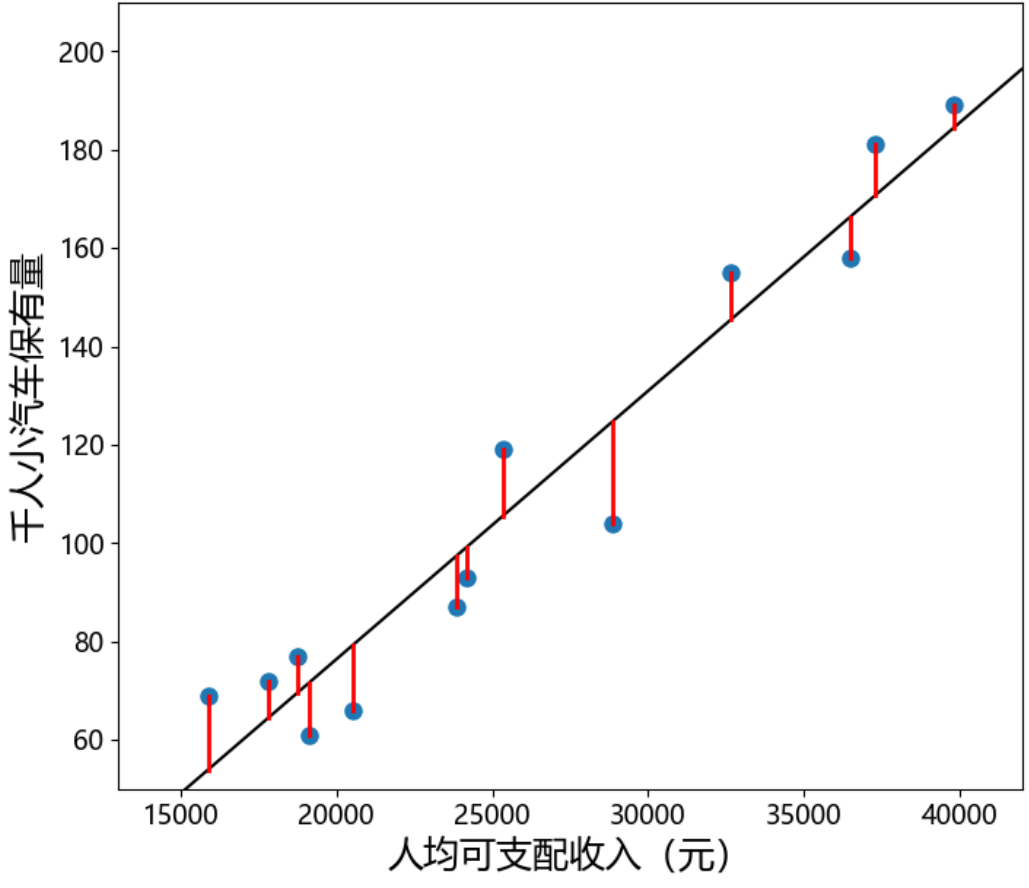
$$y = -32.829 + 0.00546x$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = 1669$$

y：千人小汽车保有量　　x：人均可支配收入（元）

一元线性回归

普通最小二乘法（Ordinary Least Square, OLS）



找到一组最好的a和b，使得

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

最小化!

其中，$\hat{y}_i = a + b_1x_1 + \cdots + b_kx_k$

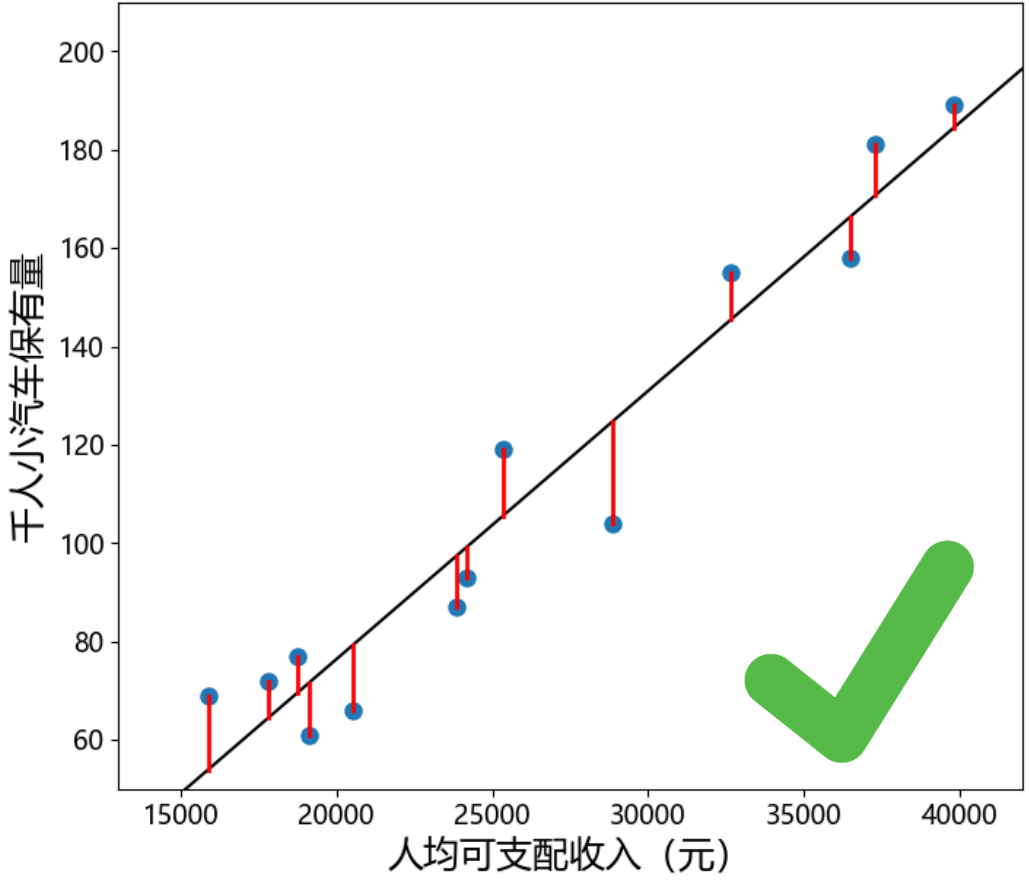
$$b=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sum_i(x_i-\bar{x})^2}=0.00546$$

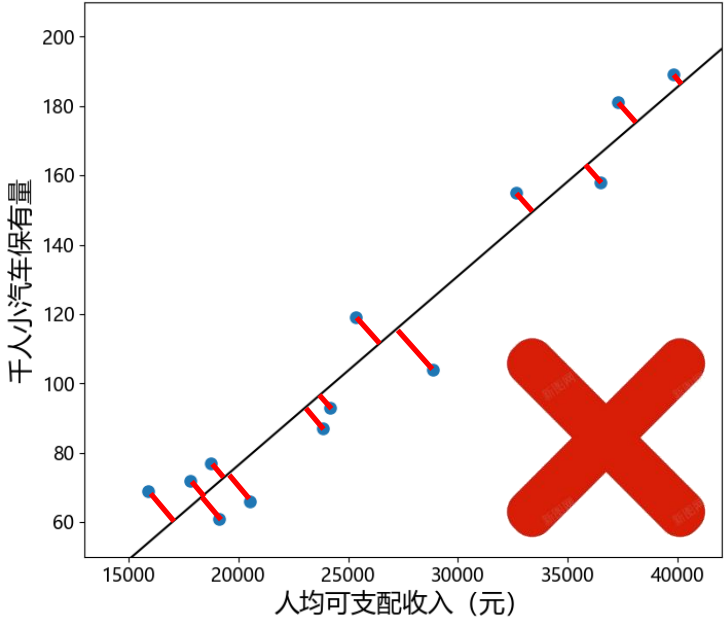
$$a=\bar{y}-b\bar{x}=-32.829$$

$$\hat{y}=-32.829+0.00546x$$

一元线性回归

普通最小二乘法（Ordinary Least Square, OLS）





普通最小二乘法（Ordinary Least Square, OLS）

“Of all the principles that can be proposed for this purpose, I think there is none more general, more exact, or easier to apply, than that which we have used in this work; it consists of **making the sum of the squares of the errors a minimum**. By this method, a kind of equilibrium is established among the errors which, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which most nearly approaches the truth.” ——Legendre (1805)

“在所有为此目的而提出的原则中，我认为没有任何一个比我们在这个工作中所使用的更为普遍、精确或易于应用；这个原则是**使误差的平方和最小化**。通过这种方法，在误差之间建立了一种平衡，这种平衡防止极端值占主导地位，因此适合揭示最接近真实的系统状态。” ——勒让德（1805）

回归方程的解释

$$\hat{y} = -32.865 + 0.00546x$$

- $b = 0.0054$：人均可支配收入每增加1元，千人小汽车保有量增加0.00546辆；或者人均可支配收入每增加1000元，千人小汽车保有量增加5.46辆。

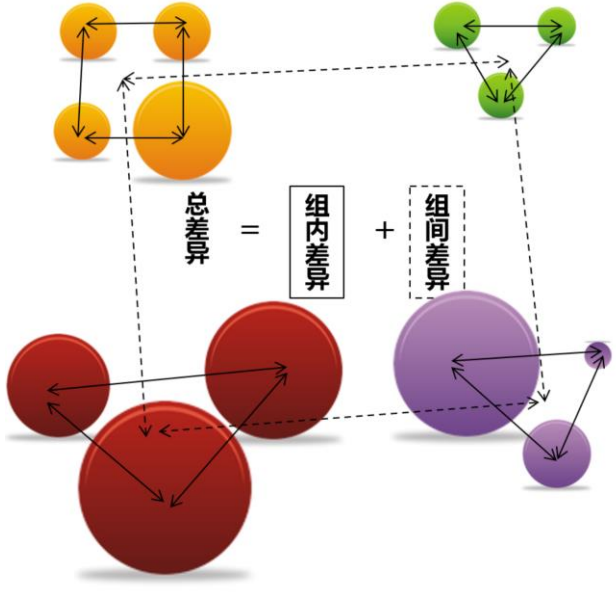
斜率：自变量每增加1个单位，因变量增加多少个单位?

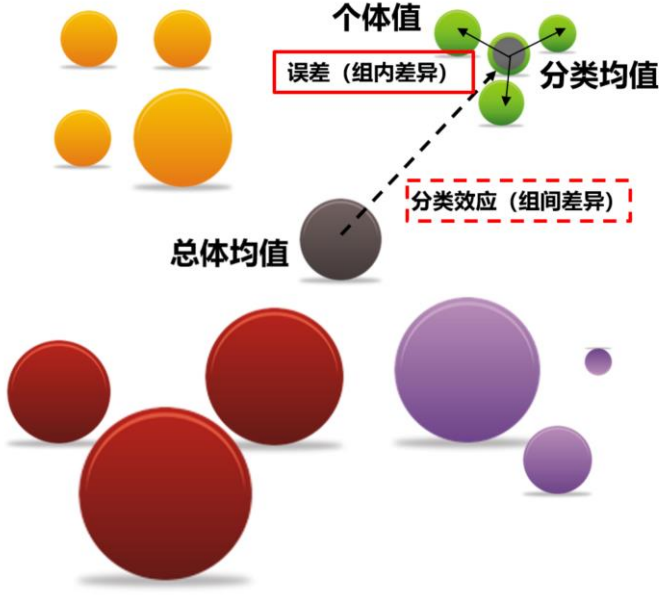
- $a = -32.865$：当人均可支配收入为0时，千人小汽车保有量为-32.826，但这一解释并不具有实际意义。

常数项：自变量为0时，因变量的取值是多少？需要考虑实际意义。

拟合优度

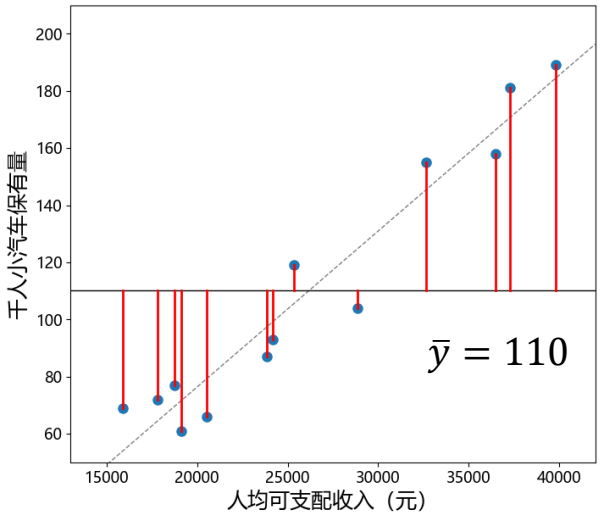
- **拟合优度**（Goodness of Fit）是统计学中的一个指标，用于衡量统计模型拟合观测数据的好坏程度。它可以帮助评估模型对数据的解释能力，以及模型的预测效果是否理想。

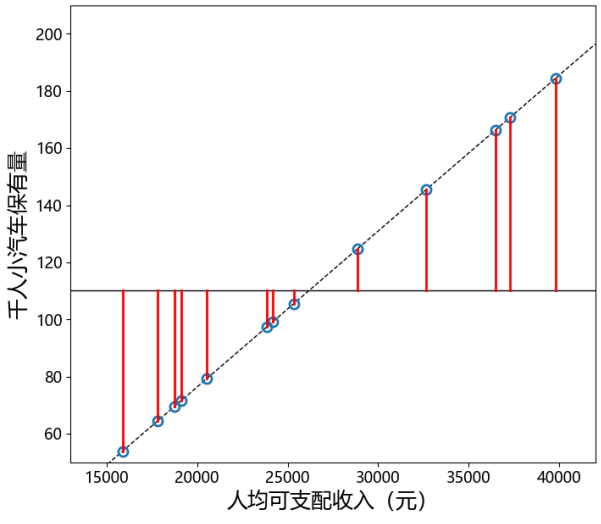


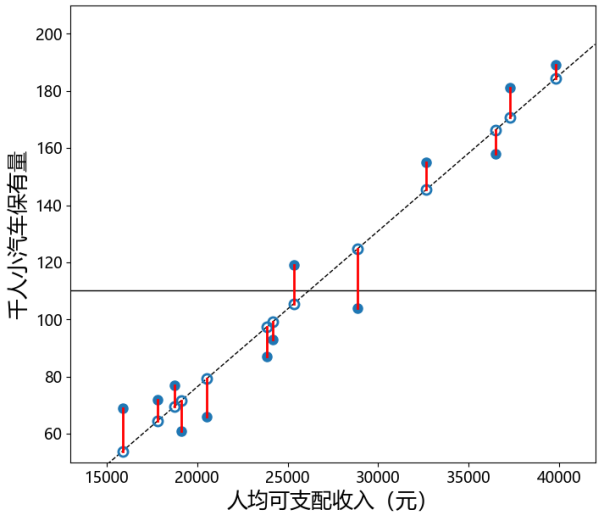


一元线性回归

拟合优度







总平方和（SST） = **回归平方和（SSR）** + **残差平方和（SSE）**

实际值相对于平均值的变异

$$SST = \sum_i (y_i - \bar{y})^2 = 25096.9$$

全部信息量

模型预测值相对于平均值的变异

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2 = 23428.2$$

可解释的信息量

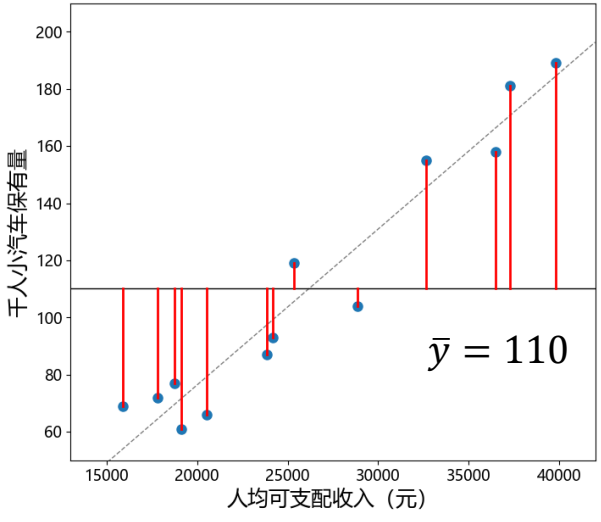
实际值与模型预测值的差异

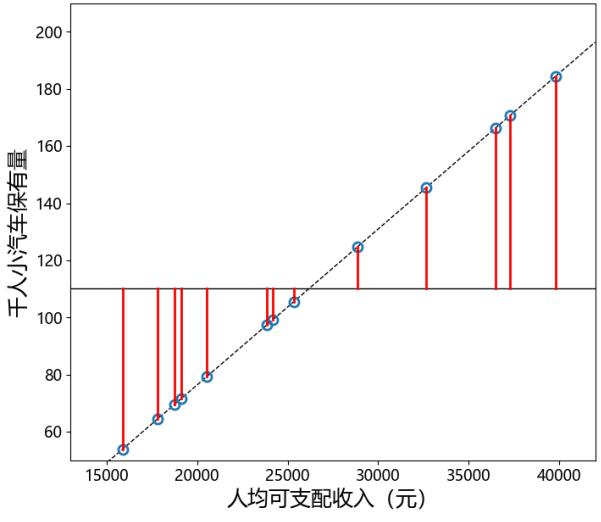
$$SSE = \sum_i (y_i - \hat{y}_i)^2 = 1668.7$$

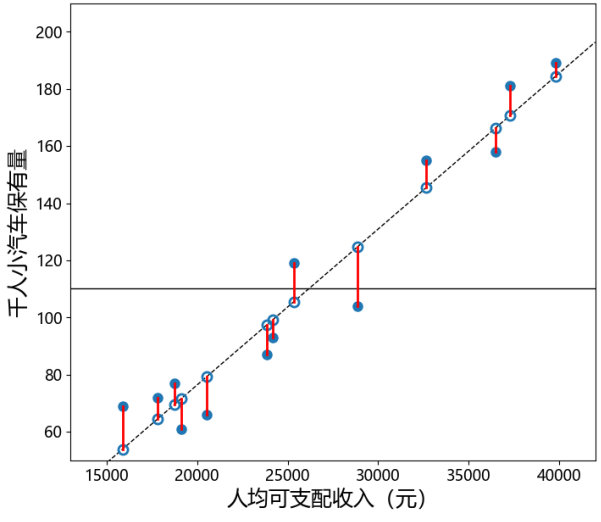
不可解释的信息量

一元线性回归

拟合优度







总平方和（SST）　=　回归平方和（SSR）　+　残差平方和（SSE）

$$r^2 = \frac{SSR}{SST} = \frac{\text{可解释的信息量}}{\text{总信息量}} = \frac{23428.2}{25096.9} = 0.934$$

一元线性回归

拟合优度

- 拟合优度r^2通常介于0~1之间，其值越高表明模型的拟合效果越好，解释力和预测力越强。
- 在线性回归中，r^2又被称为样本决定系数（coefficient of determination），其具体意义是因变量 y 的变异（信息量）可以被自变量 x 所解释的比例。
- 在一元线性回归中，r^2 等于因变量 y 和自变量 x 的Pearson相关系数的平方。
- 虽然我们一般希望r^2越高越好，但它并没有一个绝对的标准。

What's a good value for R-squared?

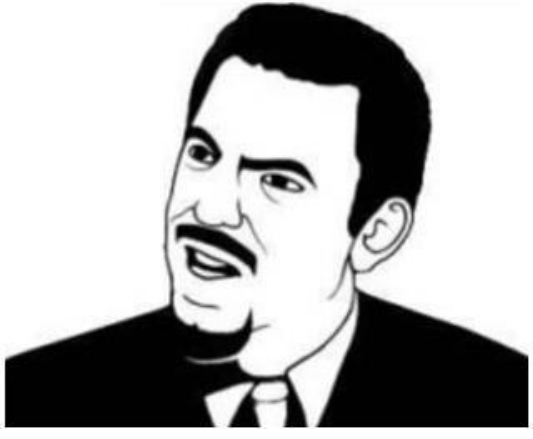
The question is often asked: "what's a good value for R-squared?" or "how big does R-squared need to be for the regression model to be valid?" Sometimes the claim is even made: "a model is not useful unless its R-squared is at least x", where x may be some fraction greater than 50%. The correct response to this question is polite laughter followed by: "That depends!" A former student of mine landed a job at a top consulting firm by being the only candidate who gave that answer during his interview.

https://people.duke.edu/~rnau/rsquared.htm#punchline

一元线性回归

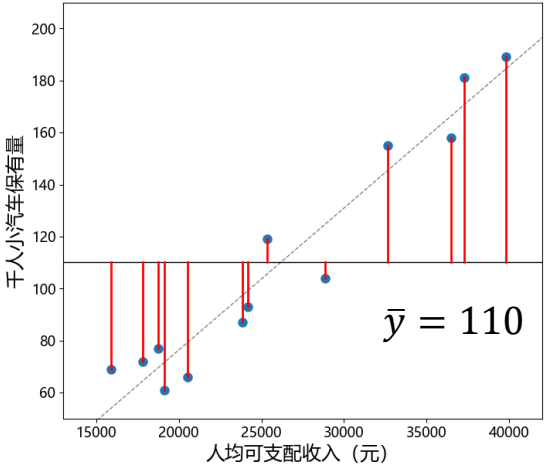
结果是样本的偶然性造成的吗?

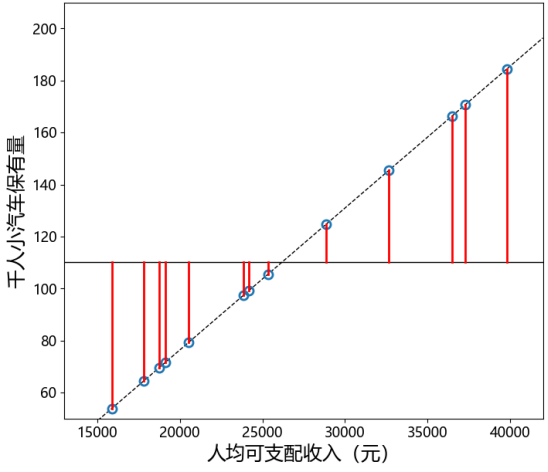
假设检验!

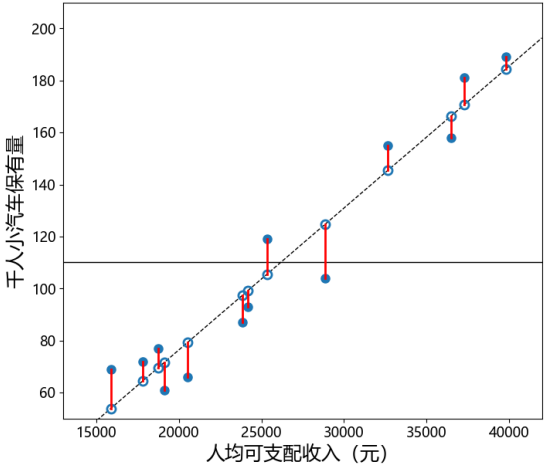


一元线性回归

模型整体的F检验： （除截距以外的）**所有系数均为0，模型无意义**







总平方和（SST） = **回归平方和（SSR）** + **残差平方和（SSE）**

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{23428.2/1}{1668.7/(13-1-1)} = 154.439$$

k：变量个数

n：样本量

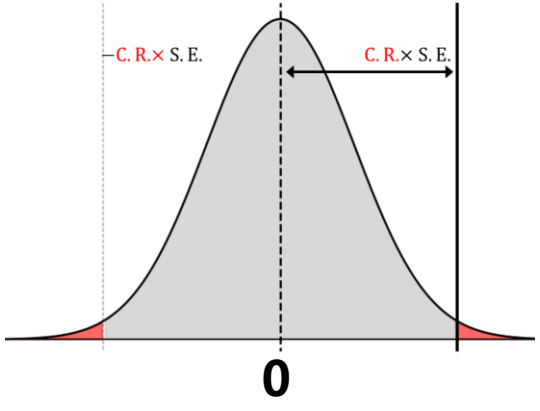
- 如果零假设成立，则$p = P(F \geq 154.439) = 8.6 \times 10^{-8} < 0.05$，所以拒绝零假设，认为得到如此高的F统计量不是抽样的巧合，至少一个模型系数不为0，模型有意义。

一元线性回归

特定回归系数的T检验：系数与0没有显著差异，变量无影响

$$\hat{y} = -32.829 + 0.00546x$$

H0：与0没有显著差异



$$t = \frac{b - 0}{SE_b} = \frac{b}{SE_b} = \frac{0.00546}{0.000439} = 12.427$$

- 如果零假设成立，则$p = P(|t| \geq 12.427) = 8.6 \times 10^{-8} < 0.05$，因此拒绝原假设，认为得到如此高的t统计量不是抽样的巧合，人均可支配收入的影响是显著的。
- 对于一元线性回归而言，T检验与F检验等价。

小结：线性回归的主要结果

- 回归方程：回归系数的大小和正负，如何解释。
- 拟合优度r^2：回归模型对数据的拟合效果。
- 假设检验（F检验、T检验）：回归系数是否具有统计显著性。

拟合优度 vs. 显著性

- r^2反映了样本中的因变量被解释的比例，而显著性反映的是推断总体的信心。
- r^2低但统计显著：影响是可靠的，但还远不足以解释因变量的全部信息。
- r^2高但统计不显著：看似解释了因变量的大部分信息，但是是通过堆砌变量实现的，或者样本量很有限，导致推断总体的信心不足。
- 经验：当样本量增加时，假设检验倾向于显著，而r^2可能会降低。

多元线性回归

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon$$

多元线性回归

京沪高铁站点周边的城镇化

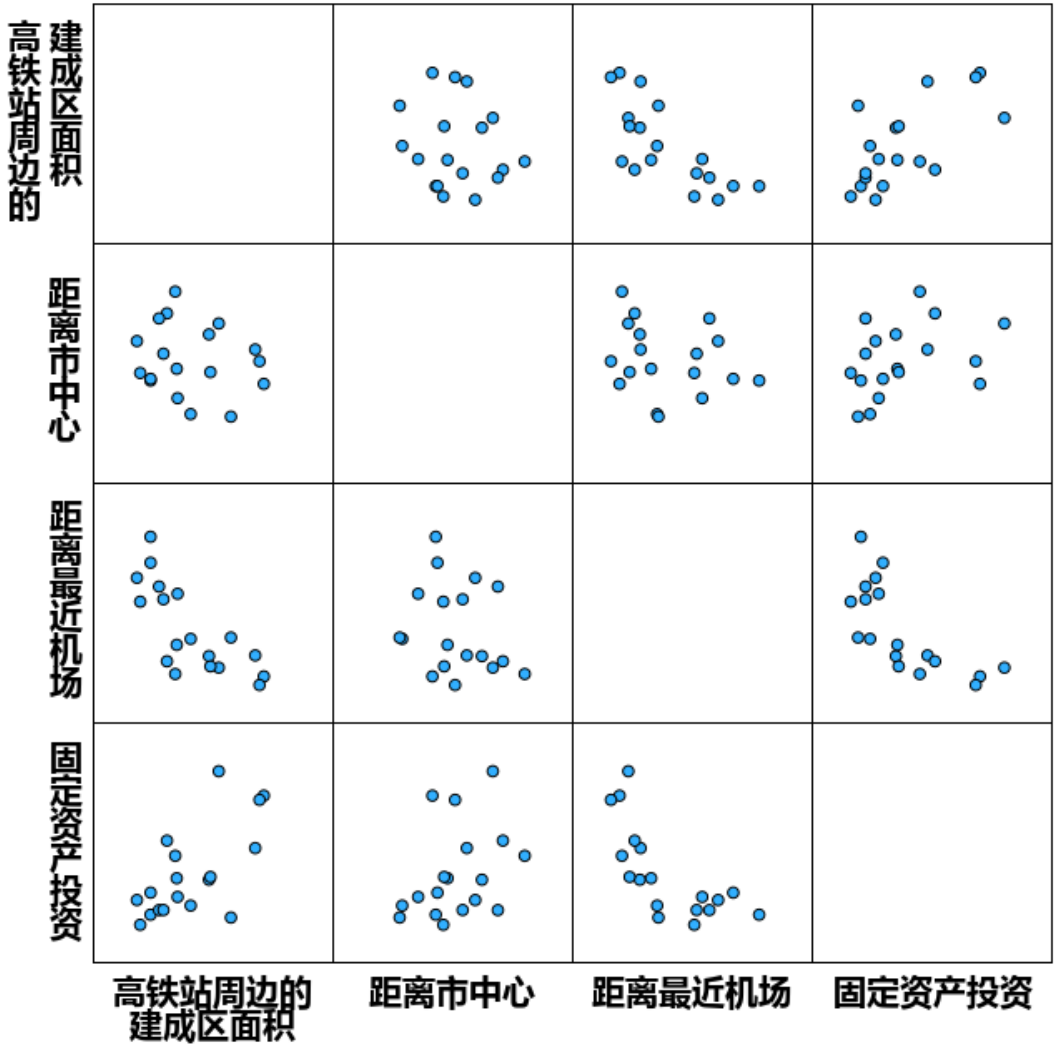
城市	Area (km^2)	D1 (km)	D2 (km)	Invest (亿元)	Type
北京	201.1	7.6	9.3	5493.5	1
上海	194.7	10.3	2.1	5317.7	1
济南	119.4	13.5	26.2	1987.4	2
南京	188.2	11.7	26.8	3306.0	2
泰安	72.6	5.9	78.2	1270.5	3
苏州	56.8	16.0	21.9	3617.8	3
……	……	……	……	……	……

- 因变量：
  - Area：2012年，京沪高铁18个站点周边8km范围内的建成区面积（km^2）
- 解释变量：
  - D1：高铁站到市中心的距离（km）
  - D2：高铁站到最近机场的距离（km）
  - Invest：设站城市的固定资产投资金额（亿元）
  - Type：城市类型，1=直辖市，2=省会，3=其他城市

多元线性回归

散点图

- 观察变量之间的关系，考虑变量转换的需求；
- 检测自变量之间的强相关（共线性）；
- 识别异常值。



多元线性回归

相关分析

Correlations

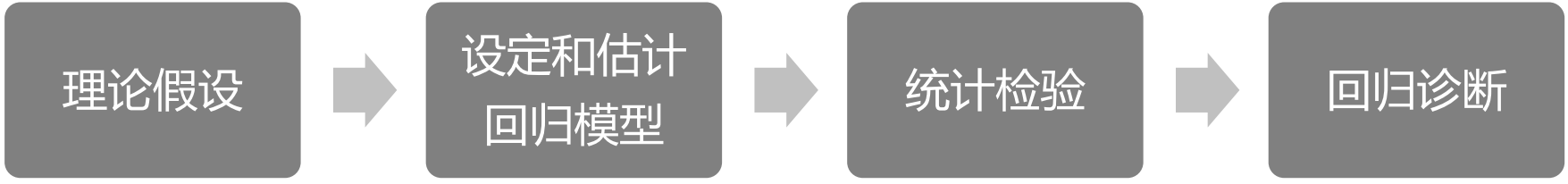
		高铁站周边的建成区面积	距离市中心	距离最近机场	固定资产投资
高铁站周边的建成区面积	Pearson Correlation	1	-.164	-.743**	.661**
	Sig. (2-tailed)		.516	<.001	.003
距离市中心	Pearson Correlation	-.164	1	-.233	.351
	Sig. (2-tailed)	.516		.352	.154
距离最近机场	Pearson Correlation	-.743**	-.233	1	-.711**
	Sig. (2-tailed)	<.001	.352		<.001
固定资产投资	Pearson Correlation	.661**	.351	-.711**	1
	Sig. (2-tailed)	.003	.154	<.001	

**. Correlation is significant at the 0.01 level (2-tailed).

- 理解变量之间的关系；
- 检测共线性；
- 变量选择？ 仅供参考！

多元线性回归

回归分析



- 理论假设：分析研究问题，开展文献回顾，明确因变量和可能的影响因素，提出因变量与自变量关系的理论假设，**尽量避免数据导向的分析！**
- 设定回归模型

$$Area = a + b_{D1}D1 + b_{D2}D2 + b_{Invest}Invest + \varepsilon$$

多元线性回归

估计回归模型：回归系数

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	173.626	32.881		5.280	<.001
	距离市中心	-6.569	2.087	-.441	-3.147	.007
	距离最近机场	-.896	.311	-.539	-2.885	.012
	固定资产投资	.014	.006	.433	2.232	.042

a. Dependent Variable: 高铁站周边的建成区面积

$$\widehat{Area} = 173.626 - 6.569D1 - 0.896D2 + 0.014Invest$$

多元线性回归

估计回归模型：非标准化回归系数

$$\widehat{Area} = 173.626 - 6.569D1 - 0.896D2 + 0.014Invest$$

- **非标准化系数（Unstandardized Coefficients）B**：自变量每增加一个单位（控制其他自变量保持不变），因变量变化多少个单位。
  - $b_{D1} = -6.569$：高铁站与市中心的距离每增加1km，站点周边建成区面积减少6.569km^2。
  - $b_{D2} = -0.896$ ：高铁站与最近机场的距离每增加1km，站点周边建筑成面积减少0.896km^2 。
  - $b_{Invest} = 0.014$：城市固定资产投资每增加1亿元，站点周边建成区面积增加0.014km^2 。
- 特别地，**常数项**表示所有自变量取值均为0时的因变量取值，需要考虑实际意义。
  - $a = 173.626$：3个自变量均为0时的站点周边建成区面积，无实际意义。
- 系数的正负表示效应的方向：两个距离为负效应，固定资产投资为正效应。

多元线性回归

估计回归模型：回归系数

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	173.626	32.881		5.280	<.001
	距离市中心	-6.569	2.087	-.441	-3.147	.007
	距离最近机场	-.896	.311	-.539	-2.885	.012
	固定资产投资	.014	.006	.433	2.232	.042

a. Dependent Variable: 高铁站周边的建成区面积

- 非标准化系数的大小反映了效应的强度，但是在比较时，存在单位和量纲问题。
- 在比较不同自变量的影响强度时，通常使用标准化系数。

$$\widehat{Z\_Area} = -0.441Z\_D1 - 0.539Z\_D2 + 0.433Z\_Invest$$

多元线性回归

估计回归模型：标准化回归系数

$$\widehat{Z\_Area} = -0.441Z\_D1 - 0.539Z\_D2 + 0.433Z\_Invest$$

- **标准化系数（Standardized Coefficients）Beta**：将自变量和因变量分别标准化以后拟合的回归模型系数，可以解释为：自变量每增加一个**标准差**（控制其他自变量保持不变），因变量变化多少个**标准差**。
  - $\beta_{D1} = -6.569$：高铁站与市中心的距离每增加1个标准差（4.129km，3.7~18.6km），站点周边建成区面积减少0.441个标准差（0.441 × 61.434=27.092km^2）。
  - $\beta_{D2} = -0.896$：高铁站与最近机场的距离每增加1个标准差（36.949km，2.1~125.7km），站点周边建筑成面积减少0.539个标准差（0.896 × 61.434=55.045km^2）。
  - $\beta_{Invest} = 0.433$：城市固定资产投资每增加1个标准差（1909.547亿元，107~6651亿元），站点周边建成区面积增加0.433个标准差（0.433 × 61.434=26.600km^2）。

- 标准化系数没有常数项。

估计回归模型：拟合优度

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.871[a]	.759	.707	33.26394

a. Predictors: (Constant), 固定资产投资, 距离市中心, 距离最近机场

- r^2又被称为样本决定系数（coefficient of determination），其具体意义是因变量 y 的变异（信息量）可以被自变量 x 所解释的比例。
- 然而，多元线性回归中r^2 存在一个严重问题：当引入新的自变量时，即使它对因变量的解释没有意义，也会使r^2增大。

估计回归模型：拟合优度

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.871[a]	.759	.707	33.26394

a. Predictors: (Constant), 固定资产投资, 距离市中心, 距离最近机场

- 多元线性回归一般使用调整r^2，即r^2_{adj}。该指标会随着解释变量数量的增加而降低，从而综合考虑了模型的拟合效果和简约性。

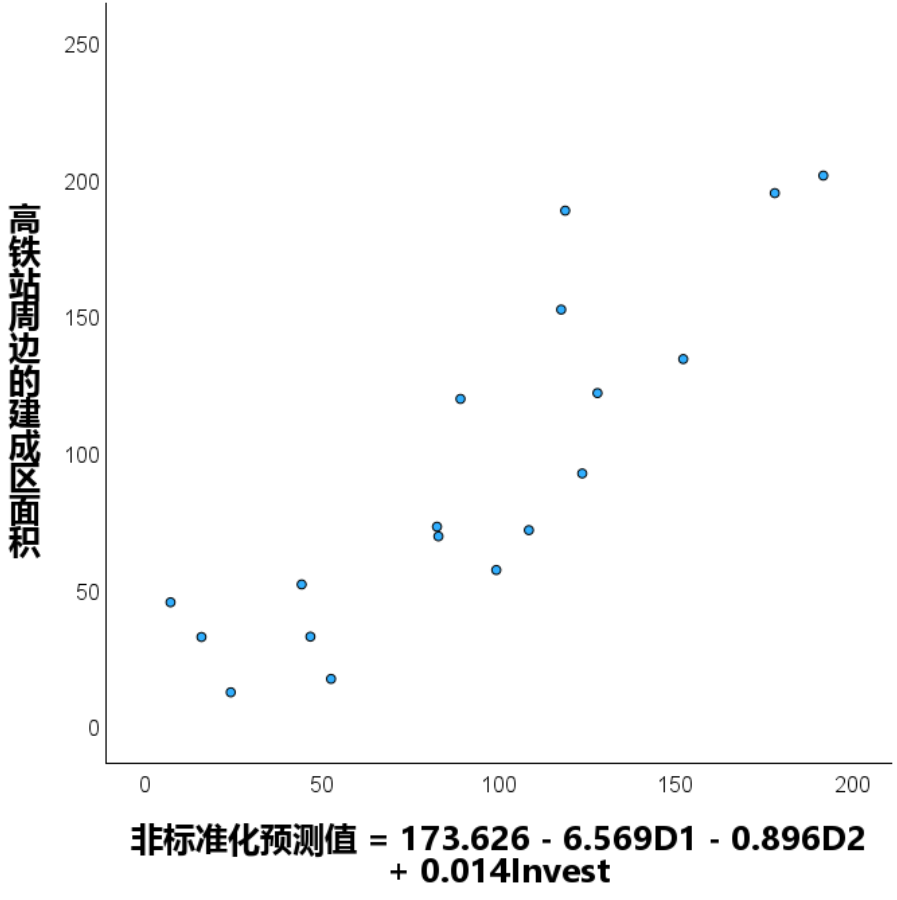
$$r^2_{adj} = 1 - \frac{n-1}{n-k-1}(1-r^2)$$

$n = 18$：样本量

$k = 3$ ：解释变量的数量

多元线性回归

估计回归模型：拟合优度



Model Summary

Model	R	R Square	Adjusted R Square
1	.871[a]	.759	.707

- 复相关系数：测量一个变量与其他多个变量之间线性相关程度的指标，相当于因变量的实际值与预测值之间的相关系数。

多元线性回归

统计检验：模型整体的F检验

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	48735.282	3	16245.094	14.682	<.001[b]
	Residual	15490.855	14	1106.490		
	Total	64226.137	17			

a. Dependent Variable: 高铁站周边的建成区面积

b. Predictors: (Constant), 固定资产投资, 距离市中心, 距离最近机场

- 零假设：除常数项以外的所有系数均为0，即$b_{D1} = b_{D2} = b_{Invest} = 0$。
- 原理：将总平方和分解为回归平方和与残差平方和两部分，比较前者与后者，构建F统计量。
- 结果：$p = P(F \geq 14.682) = 0.00013 < 0.05$，拒绝零假设，模型有意义。

多元线性回归

统计检验：回归系数的T检验

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	173.626	32.881		5.280	<.001
	距离市中心	-6.569	2.087	-.441	-3.147	.007
	距离最近机场	-.896	.311	-.539	-2.885	.012
	固定资产投资	.014	.006	.433	2.232	.042

a. Dependent Variable: 高铁站周边的建成区面积

- 零假设：特定的回归系数与0没有显著差异。
- 原理：t=(B-0)/S.E.=B/S.E.。
- 结果：所有回归系数的p值均小于0.05，均拒绝零假设，具有统计显著性。
- 注意：“距离市中心”在单独的两两相关分析中不显著，但在多元线性回归中是显著的。

虚拟变量

(Dummy Variables)

虚拟变量

分类变量作为自变量

城市	Area (km^2)	D1 (km)	D2 (km)	Invest (亿元)	Type
北京	201.1	7.6	9.3	5493.5	1
上海	194.7	10.3	2.1	5317.7	1
济南	119.4	13.5	26.2	1987.4	2
南京	188.2	11.7	26.8	3306.0	2
泰安	72.6	5.9	78.2	1270.5	3
苏州	56.8	16.0	21.9	3617.8	3
……	……	……	……	……	……

因变量：数值变量

自变量：分类变量

方差分析（ANOVA）

F=12.082, p<0.001, eta=0.785

虚拟变量

分类变量作为自变量

城市	Area (km^2)	D1 (km)	D2 (km)	Invest (亿元)	Type
北京	201.1	7.6	9.3	5493.5	1
上海	194.7	10.3	2.1	5317.7	1
济南	119.4	13.5	26.2	1987.4	2
南京	188.2	11.7	26.8	3306.0	2
泰安	72.6	5.9	78.2	1270.5	3
苏州	56.8	16.0	21.9	3617.8	3
……	……	……	……	……	……

因变量：数值变量　　其他数值自变量　　自变量：分类变量

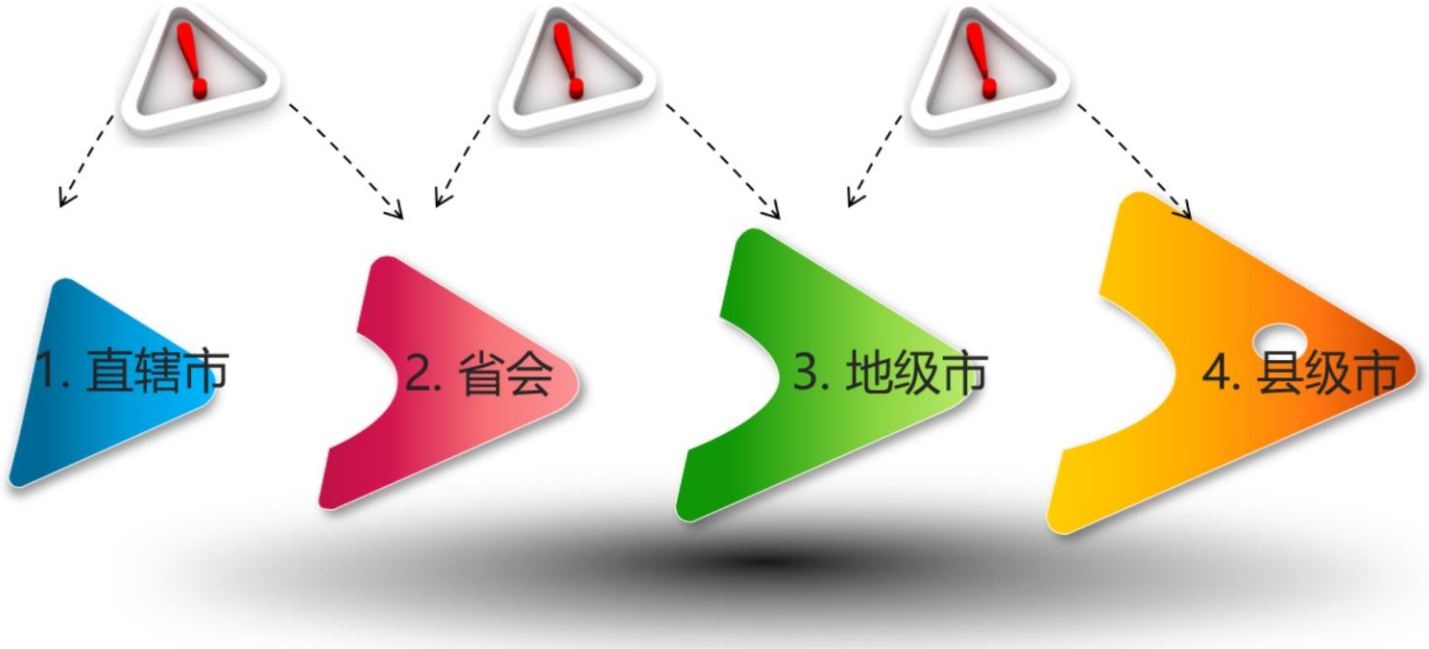
线性回归

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k + \varepsilon$$

$$Area = a + b_{D1}D1 + b_{D2}D2 + \boldsymbol{b_{Type}Type} + \varepsilon$$

分类变量作为自变量

- 数字编码没有数值意义：1→2，2→3，3→4的效应可以完全不同。



虚拟变量

分类变量作为自变量

1=直辖市
2=省会
3=一般城市

是直辖市吗?

是省会吗?

城市	Area (km²)	D1 (km)	D2 (km)	Type	Type1	Type2
北京	201.1	7.6	9.3	1	1	0
上海	194.7	10.3	2.1	1	1	0
济南	119.4	13.5	26.2	2	0	1
南京	188.2	11.7	26.8	2	0	1
泰安	72.6	5.9	78.2	3	0	0
苏州	56.8	16.0	21.9	3	0	0
……	……	……	……	……	……	……

因变量：数值变量　　其他数值自变量　　自变量：分类变量 ⇒ 虚拟变量

1=是
0=否

线性回归

~~$Area = a + b_{D1}D1 + b_{D2}D2 + b_{Type}Type + \varepsilon$~~

$$Area = a + b_{D1}D1 + b_{D2}D2 + b_{Type1}Type1 + b_{Type2}Type2 + \varepsilon$$

虚拟变量

将分类变量转换为虚拟变量

- **虚拟变量**/哑变量/哑元（dummy variable）：在回归分析等模型中对分类变量的编码方式，通常取值为0或1，用于表示是否属于某一类别。

类别	dummy1：是否属于第1类	dummy2：是否属于第2类	dummy3：是否属于第3类	dummy4：是否属于第4类	dummy5：是否属于第5类
第1类	1	0	0	0	0
第2类	0	1	0	0	0
第3类	0	0	1	0	0
第4类	0	0	0	1	0
第5类	0	0	0	0	1

- 有5个类别的分类变量，需要使用5-1=4个虚拟变量。
- 剩下一个是参照水平（reference level）
  - 第5类是参照水平，不属于任何虚拟变量，在所有4个虚拟变量上取值均为0。
  - 另外4类分别对应1个虚拟变量，在该虚拟变量取值为1，在其他虚拟变量上取值为0。

虚拟变量

将分类变量转换为虚拟变量

- **虚拟变量**/哑变量/哑元（dummy variable）：在回归分析等模型中对分类变量的编码方式，通常取值为0或1，用于表示是否属于某一类别。
- 有N个类别的分类变量，需要使用N-1个虚拟变量。
  - 如果使用全部N个虚拟变量，将导致完全共线性，无法估计。
- 参照水平（reference level）：必然有一个类别不属于任何虚拟变量。
  - 参照水平在N-1个虚拟变量上的取值均为0。
  - 非参照水平在N-1个虚拟变量上的取值有且仅有一个为1，其余均为0。
  - 参照水平的选取是任意的，但通常选取最有意义/最常见/第1个/最后1个的类别。

虚拟变量

在回归中纳入虚拟变量

- 以“其他城市”为参照水平，设置2个虚拟变量：Type1=直辖市，Type2=省会。

$$\widehat{Area} = 173.873 - 5.640D1 - 0.872D2 + \mathbf{72.299Type1} + \mathbf{74.112Type2}$$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	173.873	22.315		7.792	<.001
	距离市中心	-5.640	1.500	-.379	-3.759	.002
	距离最近机场	-.872	.201	-.524	-4.338	<.001
	直辖市	72.299	18.805	.451	3.845	.002
	省会城市	74.112	20.039	.390	3.698	.003

a. Dependent Variable: 高铁站周边的建成区面积

虚拟变量

虚拟变量回归系数的解释

- 以“一般城市”为参照水平，设置2个虚拟变量：Type1=直辖市，Type2=省会。

$$\widehat{Area} = 173.873 - 5.640D1 - 0.872D2 + \mathbf{72.299}\boldsymbol{Type1} + \mathbf{74.112}\boldsymbol{Type2}$$

- **非标准化系数（Unstandardized Coefficients）B**：

  虚拟变量由0变为1，即分类自变量由“参照水平”变化为“虚拟变量对应的类别”（控制其他自变量保持不变），因变量变化多少个单位。

  - $b_{Type1} = 72.299$：直辖市与一般城市相比，站点周边建成区面积高72.299km^2。
  - $b_{Type2} = 74.112$：省会城市与一般城市相比，站点周边建成区面积高74.112km^2 。

- **为什么是N-1个虚拟变量**?

  因为只能估计（N-1）个相对差异。

虚拟变量

在回归中纳入虚拟变量

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	173.873	22.315		7.792	<.001
	距离市中心	-5.640	1.500	-.379	-3.759	.002
	距离最近机场	-.872	.201	-.524	-4.338	<.001
	直辖市	72.299	18.805	.451	3.845	.002
	省会城市	74.112	20.039	.390	3.698	.003

a. Dependent Variable: 高铁站周边的建成区面积

- 虚拟变量的标准化回归系数：解释方式不变。
- 虚拟变量的标准误、t值、p值：解释方式不变，均显著。

虚拟变量

在回归中纳入虚拟变量

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.937[a]	.878	.840	24.56719

a. Predictors: (Constant), 省会城市, 直辖市, 距离市中心, 距离最近机场

$$\widehat{Area} = 173.873 - 5.640D1 - 0.872D2 + 72.299Type1 + 74.112Type2$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.871[a]	.759	.707	33.26394

a. Predictors: (Constant), 固定资产投资, 距离市中心, 距离最近机场

$$\widehat{Area} = 173.626 - 6.569D1 - 0.896D2 + 0.014Invest$$

- 拟合优度的解释方式不变。

小结

多元线性回归的主要结果

- 回归系数：
  - 显著性：一切的前提，不显著则无意义。
  - 方向：正效应还是负效应。
  - 大小：效应的强度，如果多变量比较，要用标准化系数。
    - 非标准化系数：自变量每增加1个单位，因变量变化多少个单位。
    - 非标准化系数（虚拟变量）：与参照水平相比，因变量变化多少个单位。
    - 标准化系数：自变量每增加1个标准差，因变量变化多少个标准差。
- 拟合优度——调整r^2：回归模型对数据的拟合效果，同时考虑了模型的简约性。

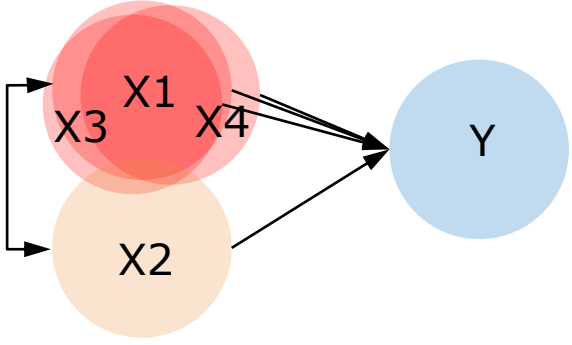
多重共线性与逐步回归

（Multicollinearity & Stepwise Regression）

多重共线性

许多自变量

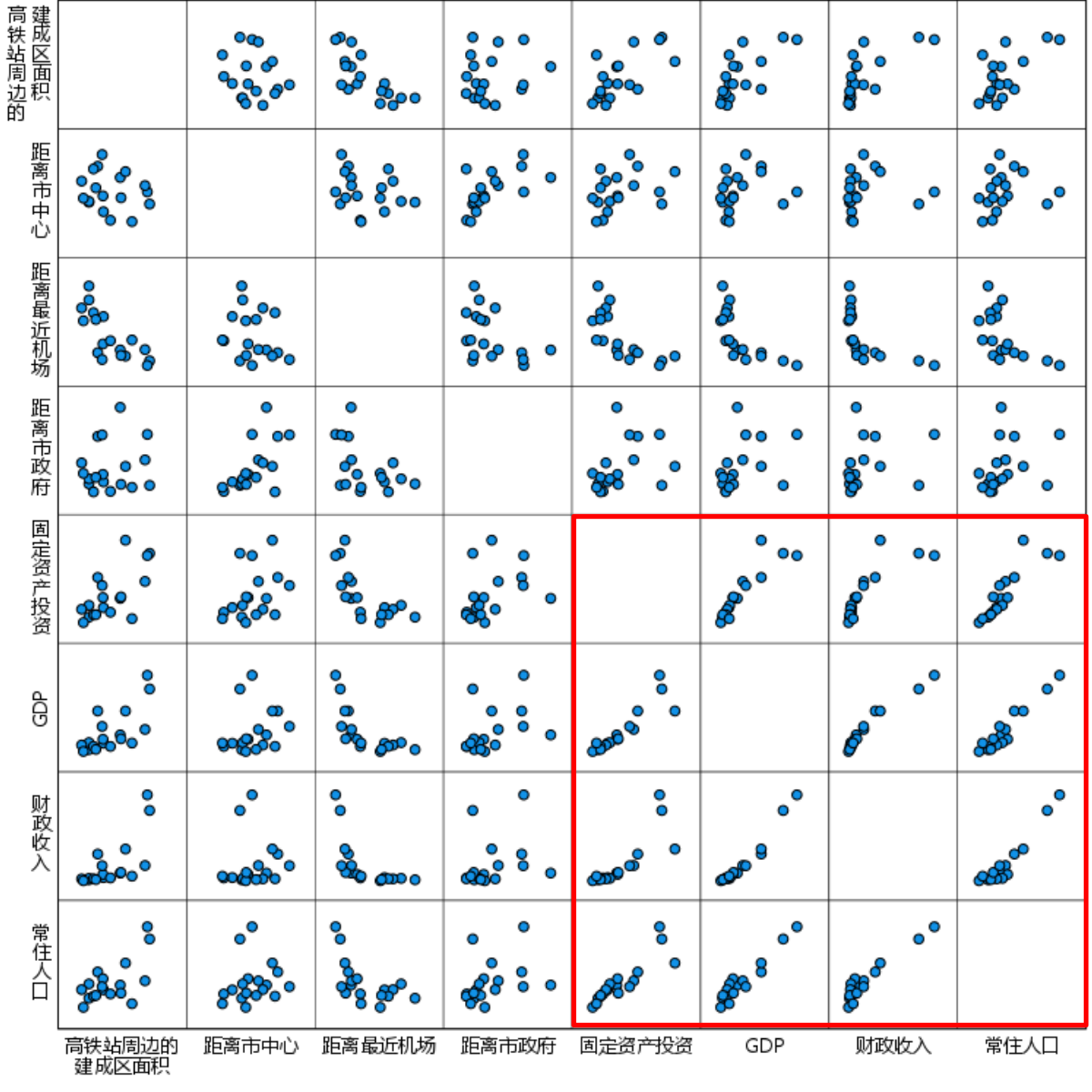
- 研究者在准备数据时，常常会收集尽可能多的解释变量。
  - 高铁站的可达性：到市中心的距离、到最近机场的距离、到市政府的距离（D3）。
  - 城市的整体经济发展水平：固定资产投资、城市类型、GDP、常住人口（Population）、财政收入（Revenue）。



- 这些自变量之间有可能彼此高度相关，存在大量的信息重叠。
- 如果直接将所有变量纳入模型，则有可能会出现多重共线性问题。

多重共线性

- 多重共线性（multi-collinearity）：自变量之间高度相关，某些自变量可以通过其他自变量的线性组合来近似解释。



多重共线性

- 多重共线性（multicollinearity）：自变量之间高度相关，某些自变量可以通过其他自变量的线性组合来近似解释。

Correlations

		高铁站周边的建成区面积	距离市中心	距离最近机场	距离市政府	固定资产投资	GDP	财政收入	常住人口
高铁站周边的建成区面积	Pearson Correlation	--							
距离市中心	Pearson Correlation	-.164	--						
	Sig. (2-tailed)	.516							
距离最近机场	Pearson Correlation	-.743**	-.233	--					
	Sig. (2-tailed)	<.001	.352						
距离市政府	Pearson Correlation	.161	.609**	-.456	--				
	Sig. (2-tailed)	.523	.007	.057					
固定资产投资	Pearson Correlation	.661**	.351	-.711**	.374	--			
	Sig. (2-tailed)	.003	.154	<.001	.127				
GDP	Pearson Correlation	.694**	.221	-.696**	.417	.901**	--		
	Sig. (2-tailed)	.001	.379	.001	.086	<.001			
财政收入	Pearson Correlation	.702**	.087	-.620**	.301	.829**	.976**	--	
	Sig. (2-tailed)	.001	.730	.006	.226	<.001	<.001		
常住人口	Pearson Correlation	.645**	.148	-.582*	.343	.879**	.959**	.955**	--
	Sig. (2-tailed)	.004	.557	.011	.164	<.001	<.001	<.001	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

多重共线性的问题

- 自变量之间的高度相关性使得它们对模型的贡献难以区分，使估计的标准误变大（即所谓的“方差膨胀”），具体有以下几点表现：
  - 回归系数解释困难：出现与预期相反、难以解释的结果。
  - 回归系数不稳定：对数据的变化十分敏感，增加或删除一条记录容易使系数发生较大变化；有些系数的估计值会异常的大。
  - 显著性检验失效：假设检验中的t值偏小，预期显著的变量变得不显著。
  - 误导性的拟合优度：R^2可能看上去很高，但并不意味着模型是理想的或稳健的。

多重共线性

$$Area \sim D1 + D2 + D3 + Invest + Type1 + Type2 + GDP + Population + Revenue$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.970[a]	.940	.873	21.91626

a. Predictors: (Constant), 常住人口, Type=省会城市, 距离市中心, 距离最近机场, 距离市政府, Type=直辖市, 固定资产投资, 财政收入, GDP

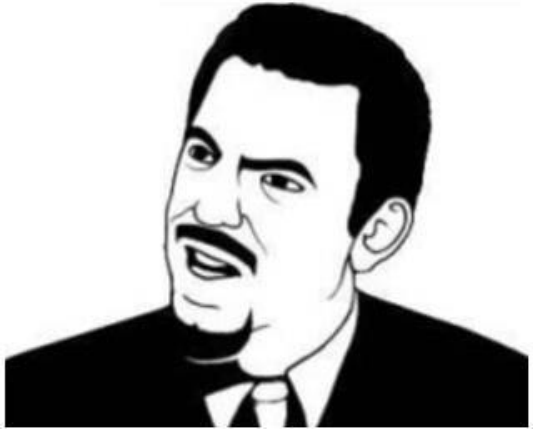
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.937[a]	.878	.840	24.56719

a. Predictors: (Constant), Type=省会城市, Type=直辖市, 距离市中心, 距离最近机场

$$Area \sim D1 + D2 + Type1 + Type2$$

R^2提高了，棒!?



多重共线性

$$Area \sim D1 + D2 + D3 + Invest + Type1 + Type2 + GDP + Population + Revenue$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	176.900	24.382		7.255	<.001
	距离市中心	-3.283	2.184	-.221	-1.503	.171
	距离最近机场	-.729	.280	-.438	-2.608	.031
	距离市政府	-5.383	2.516	-.420	-2.139	.065
	固定资产投资	-.006	.020	-.197	-.320	.757
	Type=直辖市	42.610	61.415	.266	.694	.507
	Type=省会城市	109.450	25.148	.576	4.352	.002
	GDP	.016	.016	1.000	.819	.438
	财政收入	.008	.085	-.102	-.091	.930
	常住人口	-.036	.044	-.346	-.827	.432

a. Dependent Variable: 高铁站周边的建成区面积

符号与预期相反

T检验不显著

多重共线性

完全共线性（perfect multicollinearity）

- 某些自变量可以通过其他自变量的线性组合来完美解释。

Type	Type1： 是否是直辖市	Type2： 是否是省会城市	Type3： 是否是其他城市
1. 直辖市	1	0	0
2. 省会城市	0	1	0
3. 其他城市	0	0	1

$$Type3 = 1 - Type1 - Type2$$

- 当发生完全共线性时，某个变量完全是冗余的，模型将不可估计。虚拟变量陷阱（dummy variable trap）就是典型的完全共线性。

多重共线性的诊断

- 容忍度（tolerance）和方差膨胀因子（variance inflation factor, VIF）
  - 对于模型中的一个解释变量 x_i，可以将其作为因变量，将其他解释变量作为自变量，建立一个新的线性回归模型，得到该模型的样本决定系数为r_i^2；
  - 容忍度：$Tol_i = 1 - r_i^2$
  - 方差膨胀因子：容忍度的倒数，$VIF = \frac{1}{Tol_i} = \frac{1}{1 - r_i^2}$
- 诊断标准（经验法则）：
  - VIF<5：不存在多重共线性问题；
  - 5<VIF<10：可能存在一定的多重共线性问题；
  - VIF>10：存在严重的多重共线性问题。

多重共线性

多重共线性的诊断

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	176.900	24.382		7.255	<.001		
	距离市中心	-3.283	2.184	-.221	-1.503	.171	.347	2.878
	距离最近机场	-.729	.280	-.438	-2.608	.031	.265	3.778
	距离市政府	-5.383	2.516	-.420	-2.139	.065	.194	5.154
					320	.757	.020	50.680
					694	.507	.051	19.632
					352	.002	.427	2.341
					819	.43	.005	199.404
					091	.930	.006	169.679
					827	.432	.043	23.437

a.

$$GDP \sim D1 + D2 + D3 + Invest + Type1 + Type2 + Population + Revenue$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.997[a]	.995	.991	470.19066

a. Dependent Variable: GDP

b. Predictors: (Constant), 常住人口, Type=省会城市, 距离市中心, 距离最近机场, 距离市政府, Type=直辖市, 固定资产投资, 财政收入

GDP系数的方差与没有多重共线性的情况相比，增大了近200倍!

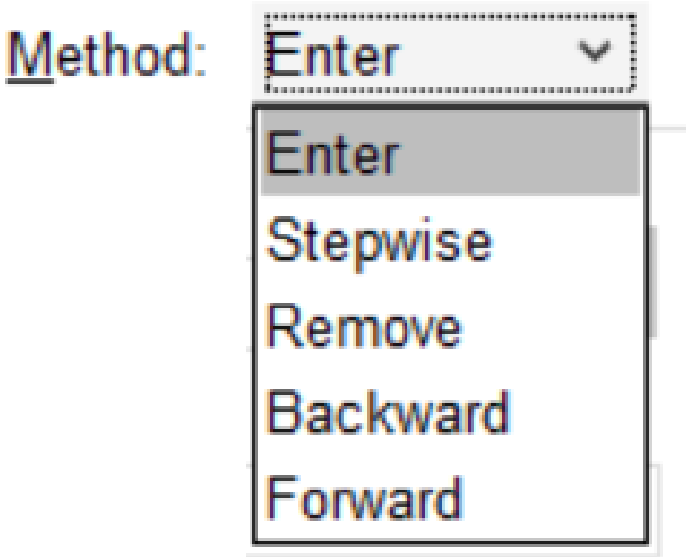
多重共线性

如何应对多重共线性?

- **无视**：具体问题具体分析
  - 如果模型的主要目的是预测，而非解释，则多重共线性的问题不大。
  - 如果VIF>10，但p值都显著：”Multicollinearity is only a problem, if it is a problem”
- **自变量筛选**：从多个变量中派代表
  - 根据理论知识，对于高度相关的解释变量高度，只保留一个变量
  - 逐步回归（stepwise regression）
- **自变量聚合**：把多个变量变成1个
  - 主成份分析（principal component analysis, PCA）
- **更复杂的回归模型**：对不稳定性进行惩罚
  - 岭回归（ridge regression），套索回归（lasso regression）

逐步回归

自变量如何进入回归方程



- **Enter**：不经筛选，强制进入。
- **Remove**：不经筛选，强制移除。
- **Forward**（前向选择）：从空模型开始，一个一个做准入筛选，一旦进入，就不会被移除。
- **Backward**（后向移除）：一开始全部进入，然后一个一个做淘汰筛选。
- **Stepwise**（逐步回归）：从空模型开始，一个一个做准入筛选；同时，每当有新变量进入时，对当前保留在回归方程中的变量做淘汰筛选，是Forward与Backward的结合。

逐步回归

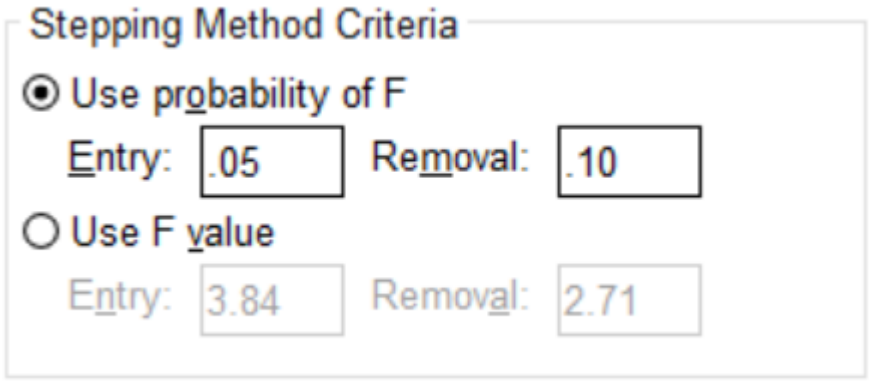
自变量选择过程

- 逐步回归是一种典型的**数据导向**的自变量选择方法，选择过程如下：
  - 对于 k 个解释变量，逐步回归首先将分别拟合 k 个一元线性回归模型，并选择p值最小且达到进入阈值（如$\alpha = 0.05$）的自变量 x_i，第一个引入模型；
  - 在已包含 x_i 的基础上，再将剩余 $k-1$ 个解释变量分别纳入，得到 $k-1$ 个二元线性回归模型，并选择p值最小且达到进入阈值的自变量 x_j ，第二个引入模型，依次类推。
  - 每当引入一个新变量以后，重新考察当前留在模型中的变量是否仍然显著，找到p值最大且超过移除阈值（如$\alpha = 0.1$）的变量将被剔除。
  - 如此往复，直至模型外的自变量均无统计意义（未达到进入阈值），模型内的自变量均有统计意义（未超过移除阈值）。

逐步回归

自变量选择标准

- 逐步回归是一种典型的**数据导向**的自变量选择方法。



进入阈值

更高的进入阈值会放宽进入标准，纳入更多自变量

移除阈值

更低的移除阈值会使筛选更严格，移除更多自变量

- 零假设：在纳入或移除某个自变量前后，模型的拟合效果没有显著差异 → F检验。

逐步回归

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	距离最近机场	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	距离市中心	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	固定资产投资	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
4	Type=省会城市	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: 高铁站周边的建成区面积

- 模型估计共计4步，每一步纳入1个自变量，依次是：距离最近机场、距离市中心、固定资产投资、是否是省会城市（虚拟变量），没有变量被移除。
- 模型的拟合优度逐步提高，最终$r^2_{adj} = 0.81$。

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.743[a]	.553	.525	42.37084
2	.820[b]	.673	.629	37.41996
3	.871[c]	.759	.707	33.26394
4	.924[d]	.855	.810	26.79329

a. Predictors: (Constant), 距离最近机场

b. Predictors: (Constant), 距离最近机场, 距离市中心

c. Predictors: (Constant), 距离最近机场, 距离市中心, 固定资产投资

d. Predictors: (Constant), 距离最近机场, 距离市中心, 固定资产投资, Type=省会城市

逐步回归

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
4	(Constant)	161.662	26.798		6.033	<.001		
	距离最近机场	-.710	.258	-.427	-2.751	.017	.464	2.154
	距离市中心	-7.480	1.710	-.502	-4.375	<.001	.847	1.180
	固定资产投资	.016	.005	.510	3.218	.007	.446	2.244
	Type=省会城市	61.799	21.100	.325	2.929	.012	.907	1.102

a. Dependent Variable: 高铁站周边的建成区面积

$$\widehat{Area} = 161.662 - 7.480D1 - 0.710D2 + 0.016Invest + 61.799Type2$$

- 模型中纳入的所有变量均显著。
- 回归系数的符号均可解释。
- 多重共线性不复存在（VIF<5）。

虚拟变量的注意事项

- 成组出现的虚拟变量**同进同出**。
  - 逐步回归仅把Type2（省会城市）纳入模型，而剔除了Type1（直辖市）。当我们认为城市类型对因变量确有影响时，一般把将Type1、Type2同时纳入模型。
  - 如果未能同进同出，有啥问题?
    - Type1 & Type2：Type2是省会城市相对于一般城市的效应
    - 仅有Type2：Type2是省会城市相对于所有其他城市（包括直辖市）的效应。
    - 如果有4个成组的虚拟变量，只保留2个，解释将更加麻烦。
  - 如何同进同出?
    - 强制进入（Enter）：Type1、Type2
    - 逐步筛选（Stepwise）：Invest、GDP、Revenue、Population

逐步回归

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Type=省会城市, Type=直辖市[b]	.	Enter
2	距离最近机场	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	距离市中心	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: 高铁站周边的建成区面积

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients B	Sig.	Collinearity Statistics Tolerance	VIF
1	(Constant)	63.528	<.001		
	Type=直辖市	113.030	<.001	.975	1.026
	Type=省会城市	90.285	.010	.975	1.026
2	(Constant)	108.237	<.001		
	Type=直辖市	75.121	.012	.684	1.463
	Type=省会城市	64.884	.034	.858	1.165
	距离最近机场	-.729	.019	.667	1.498
3	(Constant)	173.873	<.001		
	Type=直辖市	72.299	.002	.683	1.465
	Type=省会城市	74.112	.003	.845	1.183
	距离最近机场	-.872	<.001	.643	1.554
	距离市中心	-5.640	.002	.925	1.081

a. Dependent Variable: 高铁站周边的建成区面积

$$\widehat{Area} = 173.873 - 5.640D1 - 0.872D2 + 72.299Type1 + 74.112Type2$$

逐步回归

逐步回归的批判

Stopping stepwise: Why stepwise selection is bad and what you ld use instead

13 min read · Sep 23, 2018

Why no stepwise regression?

What is stepwise regression?

Many multiple regression programs can choose variab
stepwise? Because the automatic procedure fi

The problem with stepwise?

The problem is multiple comparis
so the program will compare 2k m

Step away from stepwise

Gary Smith*

*Correspondence:
gsmith@pomona.edu
Department of Economics,
Pomona College, 425 N.
College Avenue, Claremont,
CA 91711, USA

Abstract

Background: Stepwise regression is a popular data-mining tool that uses statistical significance to select the explanatory variables to be used in a multiple-regression model.

Findings: A fundamental problem with stepwise regression is that some real explanatory variables that have causal effects on the dependent variable may happen to not be statistically significant, while nuisance variables may be coincidentally significant. As a result, the model may fit the data well in-sample, but do poorly out-of-sample.

Conclusion: Many Big-Data researchers believe that, the larger the number of possible explanatory variables, the more useful is stepwise regression for selecting explanatory variables. The reality is that stepwise regression is less effective the larger the number of potential explanatory variables. Stepwise regression does not solve the Big-Data problem of too many explanatory variables. Big Data exacerbates the failings of stepwise regression.

Keywords: Stepwise regression, Data mining, Big Data

Using stepwise regression to address multicollinearity is not appropriate

Wen-Feng Xi, MD, Qing-Wei Jiang, MD, and Ai-Ming Yang, MM✉

逐步回归的批判

- 逐步回归是以数据（而非理论）为导向的，而样本数据的偶然性将为选择结果带来许多风险。
  - 两个高度相关的自变量，一进一出：纳入模型的可能实际无影响，而真正有影响可能由于数据偶然性被踢出模型。
  - 很可能高估回归系数的显著性。
  - 有可能高估拟合优度。

逐步回归

理论导向 vs. 数据导向

理论假设 → 设定和估计回归模型 → 统计检验 → 回归诊断

- 假设先行
- 变量选择基于理论
- 解释框架明确
- 偏向于验证性验究

理论导向

数据导向

- 没有明确的假设
- 通过数据发现模式
- 模型设定高度依赖数据
- 偏向于探索性研究

逐步回归

如果不做逐步回归，那怎么办？

- 要做理论导向的研究，从Field Knowledge中明确自变量。
- 如果没有多重共线性的问题，则纳入全部自变量，即使它不显著。
    - 一般而言，遗漏变量比纳入无关变量的问题更严重。
    - 当你在验证一个关系时，不显著本身也是有意义的结论。
- 如果有多重共线性的问题，可以使用其他针对多重共线性的方法。如果不得不基于数据进行自变量筛选，两害取其轻。
- 如果在前期开展探索性研究，且有大量备选自变量，逐步回归可提供一定程度的参考。

残差分析与回归诊断

(Residual Analysis)

线性回归的假设

- 线性假设：自变量与因变量之间具有线性关系 → **残差**没有特定曲线模式。
- 正态性假设：**残差**应当服从正态分布。
- 样本独立性假设：各个样本相互独立 → **残差**之间相互独立。
- 同方差假设：**残差**的方差保持不变。
- 无异常样本假设：**残差**中没有虚假的离群值。
- 无多重共线性假设：自变量之间没有强相关性。

残差分析

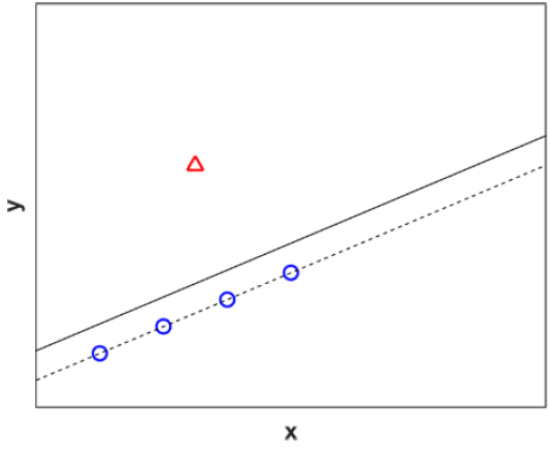
$$y = \underbrace{a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k}_{\hat{y}} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} = y - \hat{y}$$

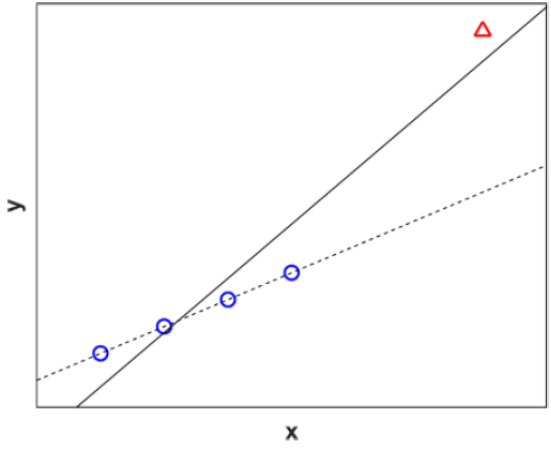
- 残差（residual）：回归分析中，因变量的实际值与预测值的差。
- 残差分析（residual analysis）：通过分析残差的模式、分布和其他统计特性，判断模型是否符合假设条件，是否存在问题。
- 主要手段：**残差图** + 统计量和假设检验。

是否有异常样本

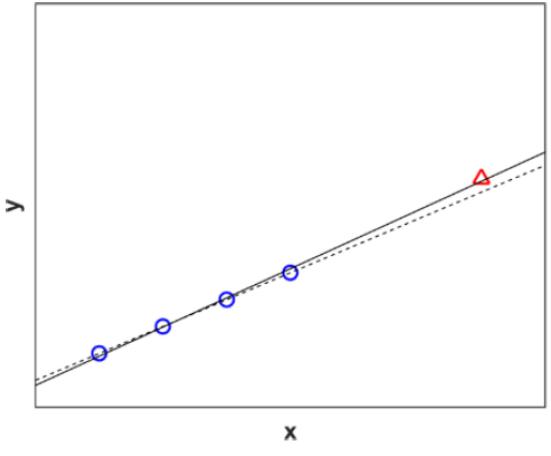
- 不同类型的异常样本：
  - 离群样本：残差较大、拟合效果较差的样本；
  - 高杠杆值样本：在自变量的取值上与大部分样本显著不同的样本；
  - **强影响样本**：对回归结果有强烈影响的样本，存在扭曲结果的风险。



差异程度大，低杠杆



强影响：差异程度大，高杠杆



差异程度小，高杠杆

是否有强影响样本

- 库克距离（Cook's distance）D_i：同时考虑了差异程度和杠杆效应。
  - 解释：当删除某个样本后，回归系数的变化影响，综合测度对系数的影响。
  - 判断标准：常用阈值为1.0，$D_i > 1.0$的样本即为强影响样本。

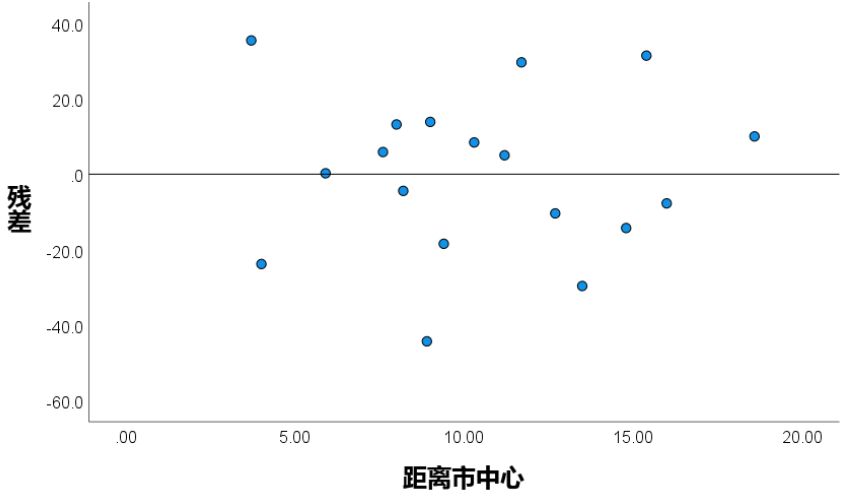
	蚌埠	北京	沧州	常州	滁州	德州	济南	昆山	廊坊
Cook's Distance	.04613	.01087	.00198	.02260	.00095	.00945	.59185	.22199	.09351
	南京	曲阜	上海	苏州	泰安	天津	无锡	徐州	枣庄
Cook's Distance	.59185	.06762	.01816	.00990	.00000	.07450	.04507	.01865	.13629

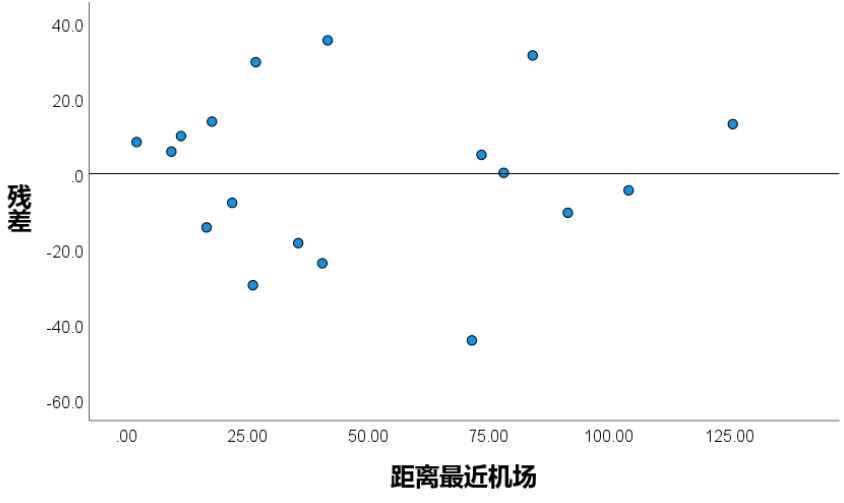
如果有强影响点：

- 检查是否存在测量或录入错误，如果有误，纠正错误或剔除出数据集；
- 如果无误，思考强影响点存在的原因：遗漏重要变量？分属不同群体？……

是否符合线性假设

- 自变量与因变量之间具有线性关系 → 残差没有特定曲线模式。
- 残差图：“自变量—残差”散点图。



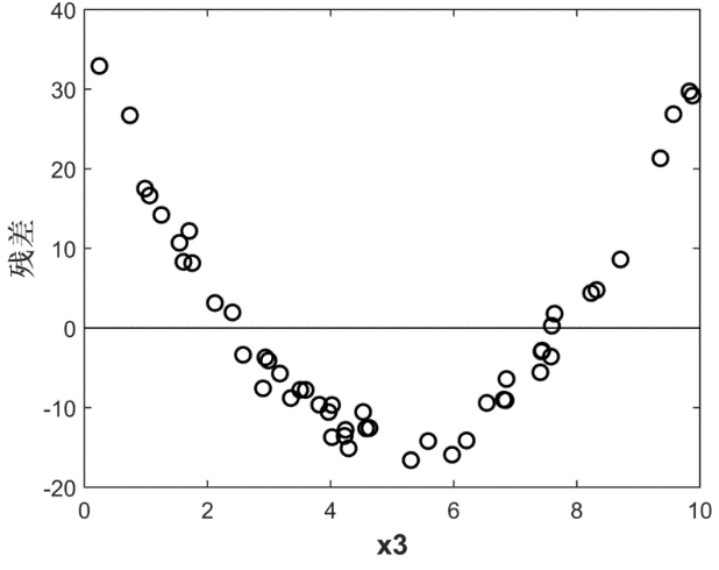


- 各点基本上平均分布在0水平线的两侧，没有表现出特定的曲线模式。

残差分析与回归诊断

是否符合线性假设

- 自变量与因变量之间具有线性关系 → 残差没有特定曲线模式。



真实模型：

$y = 10 + 3x_1 - 4x_2 + 2x_3^2 + \varepsilon$

（x_3与y具有二次关系）

线性回归模型：

$\hat{y} = -27.8 + 3.5x_1 - 4.4x_2 + 20.3x_3$

（假设x_3与y之间是线性的）

如果不满足：对自变量进行变换，使新变量满足线性假设。

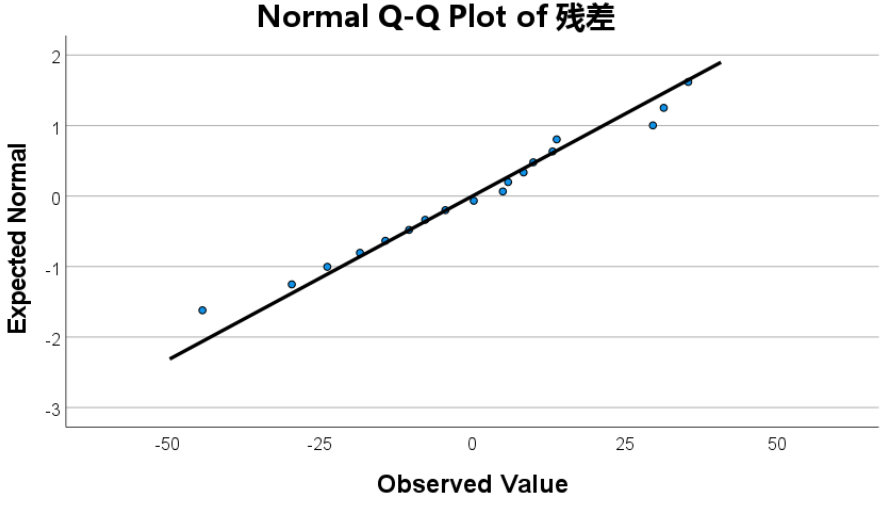
残差分析与回归诊断

是否符合正态性假设

- 残差（不是自/因变量）应当服从正态分布。
- 残差图：残差的Q-Q图，散点靠近对角线。
- 假设检验：正态性检验。

如果不满足：

- 大样本下可放松假设；
- 进一步检验异常值；
- 检查遗漏变量；
- 对自/因变量进行变换；
- 调整模型结构。



Tests of Normality

	Shapiro-Wilk		
	Statistic	df	Sig.
残差	.980	18	.949

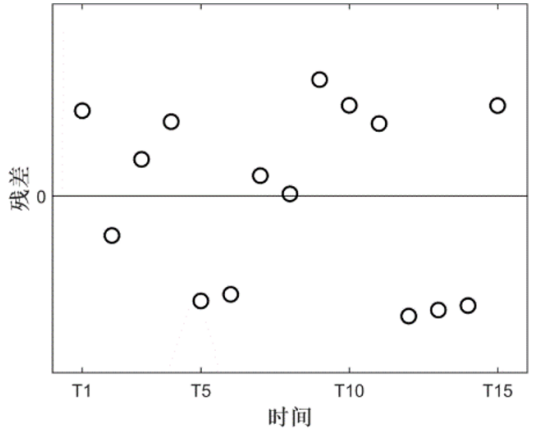
是否符合独立性假设

- 各个样本相互独立 → 残差之间相互独立。
- 对该假设最常见的违反是**序列相关** (serial correlation)，又称**自相关** (autocorrelation)，包括时间序列相关、空间序列相关。
- 非常危险的时间序列相关——伪回归！
  - 例子：人的身高和树的高度、离婚率与人均鸡肉消费量、冰淇淋销量和性犯罪数量、年龄与收入……
  - 原因：自变量和因变量在时间序列中表现出相似的趋势。

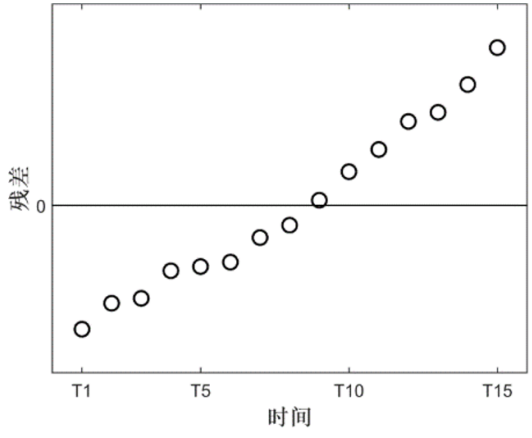
残差分析与回归诊断

是否符合独立性假设：时间序列相关

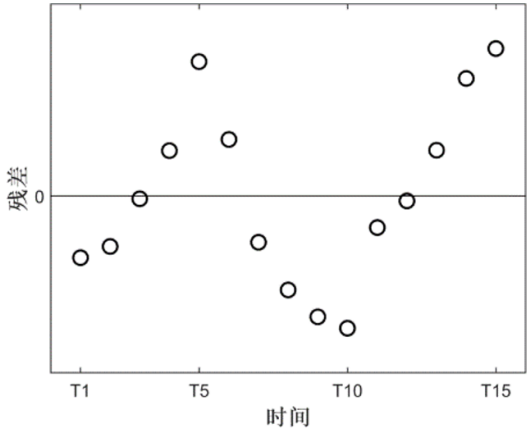
- 残差图：“时间—残差”散点图，理想状态：残差在0的上下摆动，无明显趋势。
- Durbin-Watson统计量：0~4之间，越接近2，越不存在序列相关。



理想状态



残差随时间递增



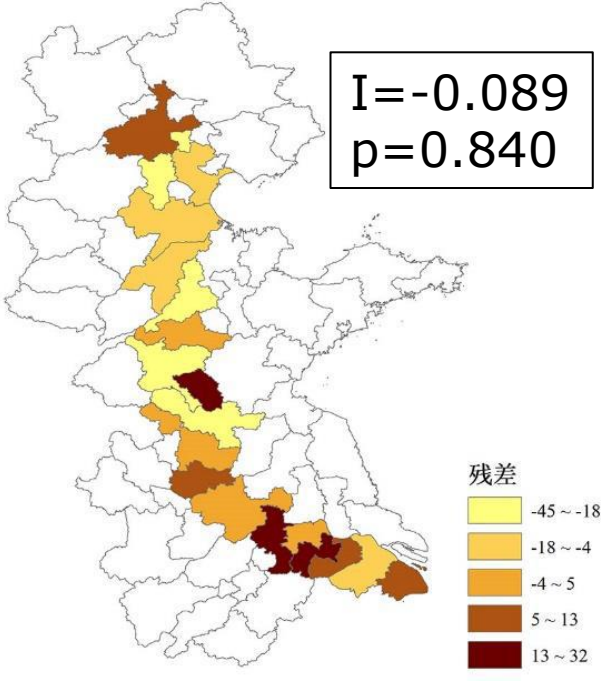
残差周期性变化

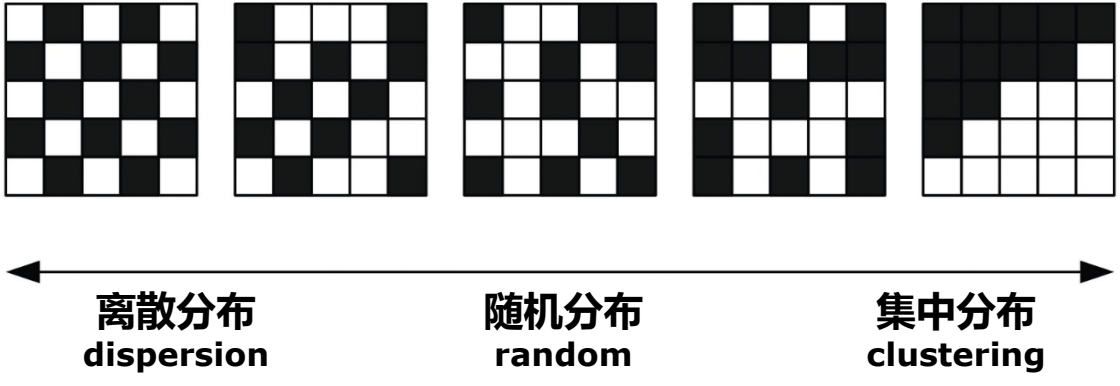
如果不满足：尝试差分变换；使用专门的时间序列分析模型。

残差分析与回归诊断

是否符合独立性假设：空间序列相关

- 残差地图：残差的空间分布，理想状态：随机分布，没有明显的空间模式。
- 残差的空间自相关指标 Moran's I：接近于0，p值不显著，则无空间序列相关。

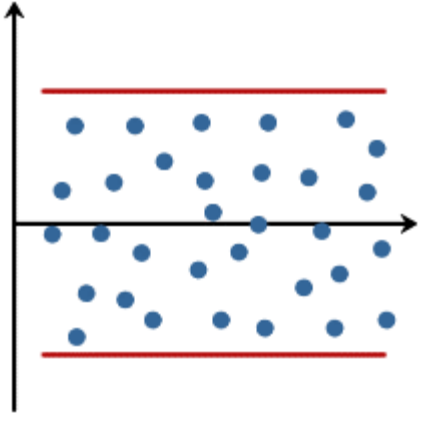




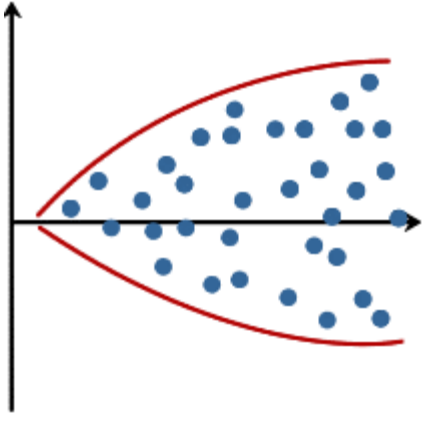
如果不满足：使用空间回归模型。

是否符合同方差假设

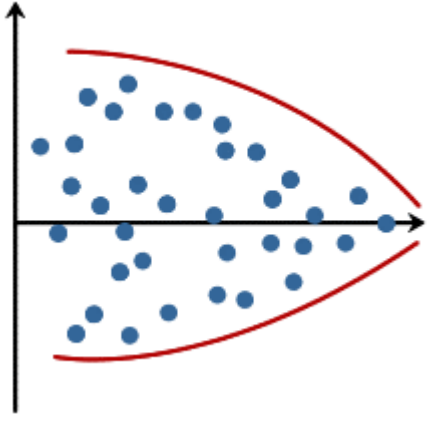
- 同方差（homoskedasticity）：残差的方差保持不变，波动范围稳定。
- 残差图：”因变量预测值—残差”散点图，检查异方差（heteroscedasticity）的典型表现：残差随着预测值的增大而呈现发散或收紧趋势。



同方差：理想状态



异方差：残差发散

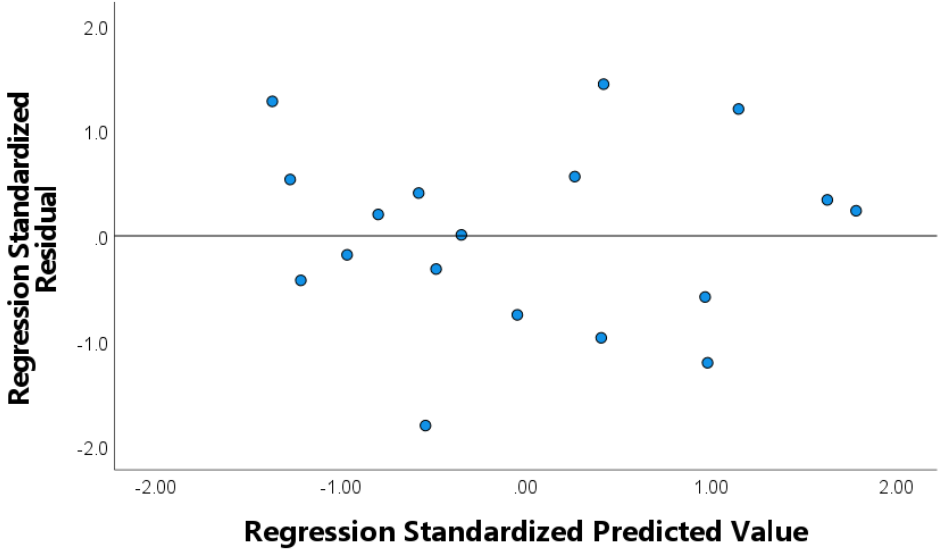


异方差：残差收紧

残差分析与回归诊断

是否符合同方差假设

- 同方差（homoskedasticity）：残差的方差保持不变，波动范围稳定。
- 残差图：”因变量预测值—残差”散点图，检查异方差（heteroscedasticity）的典型表现：残差随着预测值的增大而呈现发散或收紧趋势。



如果不满足：

- 无视：主要影响显著性；
- 可尝试对因变量进行对数或平方根变换；
- 加权最小二乘法 (WLS)：残差的方差越大，权重越小；
- 稳健估计：修正标准误。

小结

多元线性回归的问题和对策

- 多重共线性：自变量之间高度相关，对模型的贡献难以区分
  - 可能造成回归系数不显著、难以解释、不稳定等问题。
  - 可以通过VIF发现多重共线性。
  - 对策：具体分析；理论导向的变量筛选；变量聚合；岭回归和拉索回归。
- 逐步回归：数据导向的自变量筛选
  - 曾经很流行，国内依然在流行，谨慎使用。
  - 遗漏变量比无关变量更严重，不妨纳入所有从理论出发的自变量，即使不显著。
- 残差分析：模型是否满足一系列假设
  - 强影响样本：回归方程可能被扭曲，务必通过库克距离检查。
  - 序列相关：可能是伪回归，有时空背景时务必检查。
  - 线性关系、正态性、同方差：体现专业性和严谨性。