統計分析

第四講

王寧寧， Ph.D

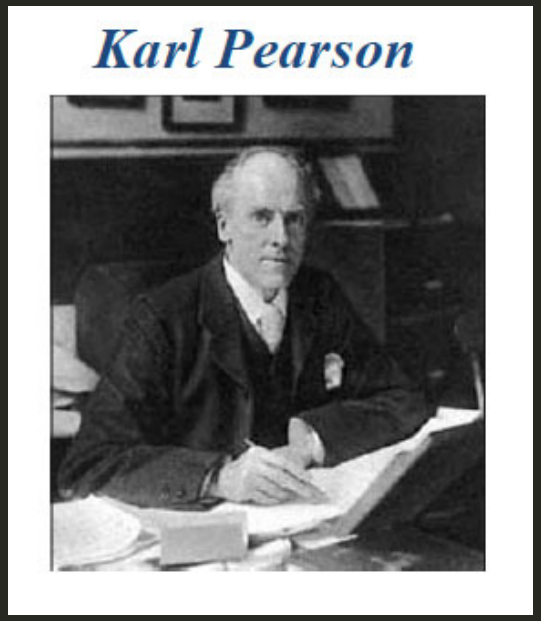
oliningning@qq.com

2022/11/15

主要內容

- 相關分析
  - 簡單相關分析
  - **spearman**秩相關
- 回歸分析
  - 簡單回歸分析
  - 多重綫性回歸

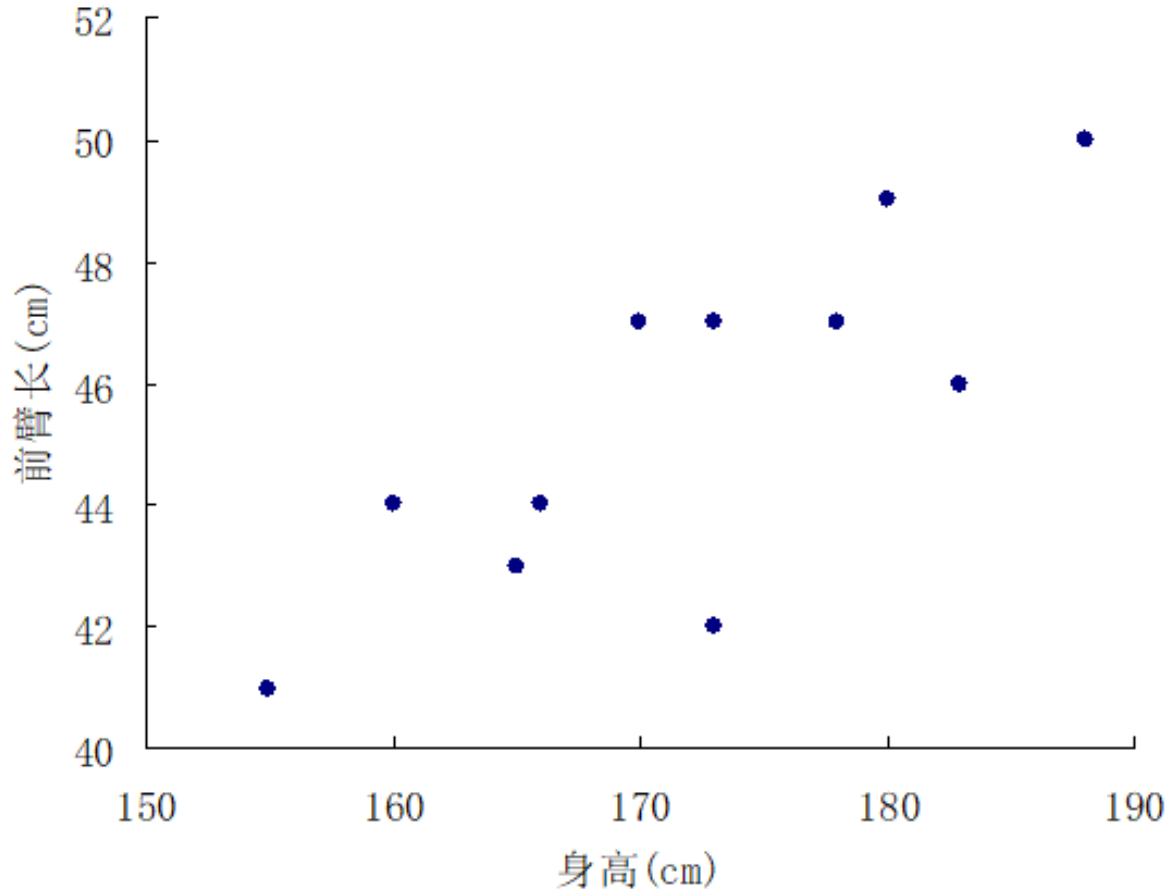
簡單相關分析



> 【例】 根據下表11個成年男青年的身高和前臂長的數據，考慮身高和前臂長的關係。

编号	身高 X	前臂长 Y
1	170	47
2	173	42
3	160	44
4	155	41
5	173	47
6	188	50
7	178	47
8	183	46
9	180	49
10	165	43
11	166	44
合计	1891	500

做散點圖



簡單相關係數

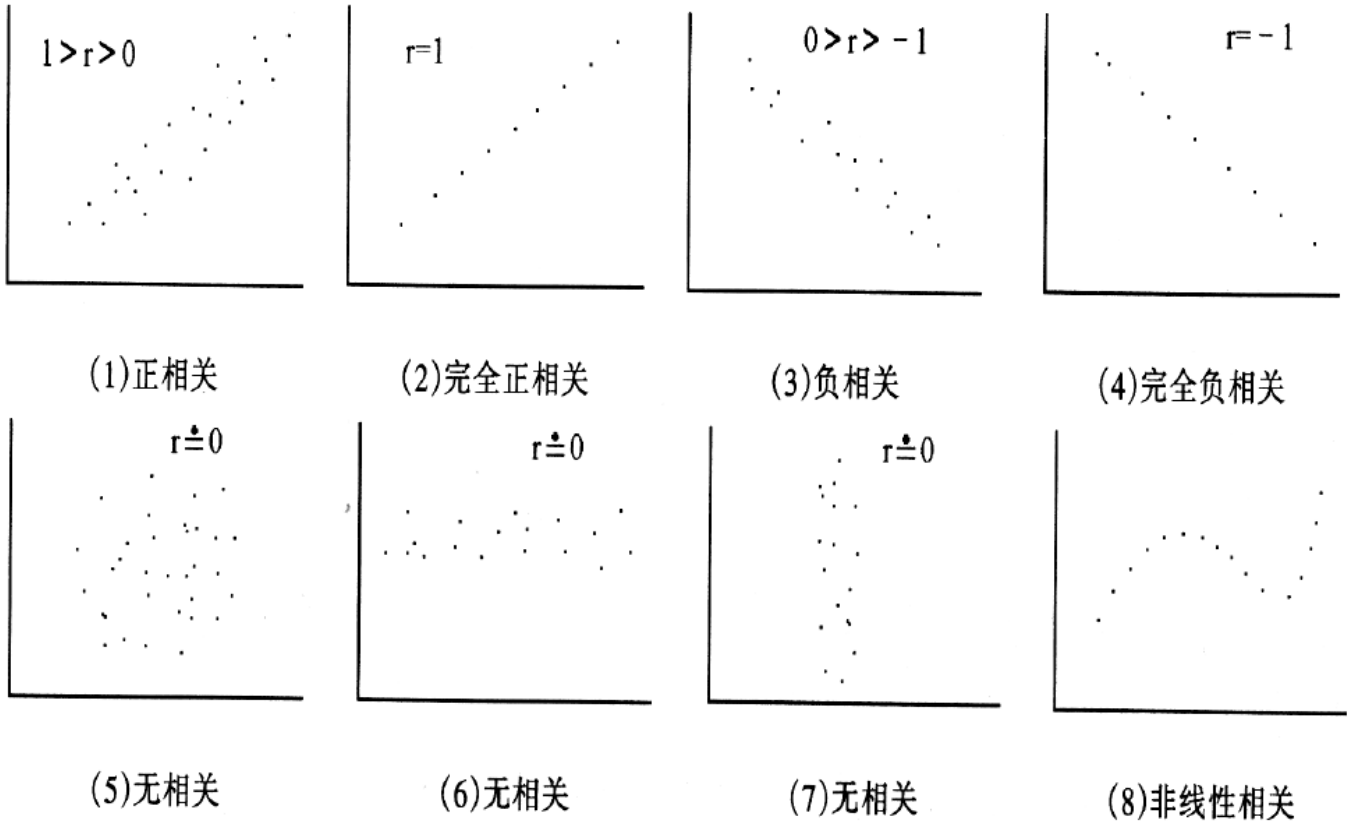
- 又稱**Pearson相關係數**，說明具有綫性關係的兩個數值變量間的密切程度與相關方向的統計量。

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

簡單相關係數的性質

- 沒有單位，數值的絕對值大小表示相關性強弱
- 越接近1表示相關性越強
- 約接近0表示相關性越弱

簡單相關係數示意圖


1>r>0
r=1
0>r> -1
r= -1
(1)正相关
(2)完全正相关
(3)负相关
(4)完全负相关
r≒0
r≒0
r≒0
(5)无相关
(6)无相关
(7)无相关
(8)非线性相关

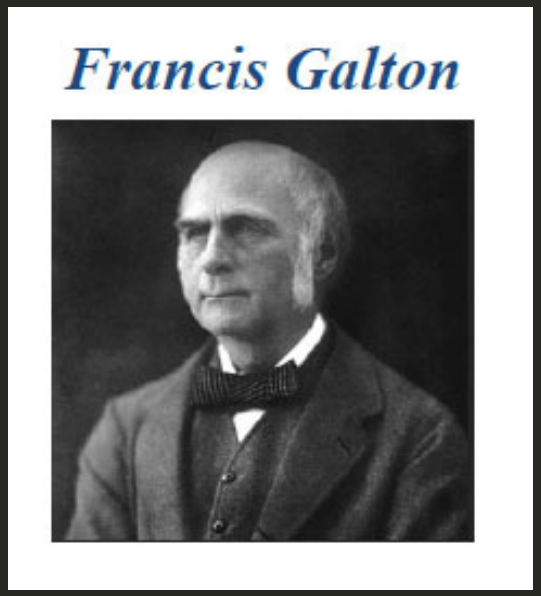
相關係數的假設檢驗

- 計算簡單相關係數： $r = 0.8009$
- 建立檢驗假設並設定檢驗水平
  - H_0: $\rho = 0$
  - H_1: $\rho \neq 0$
  - $\alpha = 0.05$
- 計算P值，得出結論
  - $P < 0.05$，拒絕原假設，在**0.05**的檢驗水平下認爲前臂長與身高存在綫性相關關係。

Spearman 秩相关检验

- 計算秩相關係數：$r_s = 0.8009$
- 建立檢驗假設並設定檢驗水平
  - H_0: $\rho_s = 0$
  - H_1: $\rho_s \neq 0$
  - $\alpha = 0.05$
- 計算P值，得出結論
  - $P < 0.05$，拒絕原假設，在**0.05**的檢驗水平下認爲前臂長與身高存在秩相關關係。

簡單回歸分析



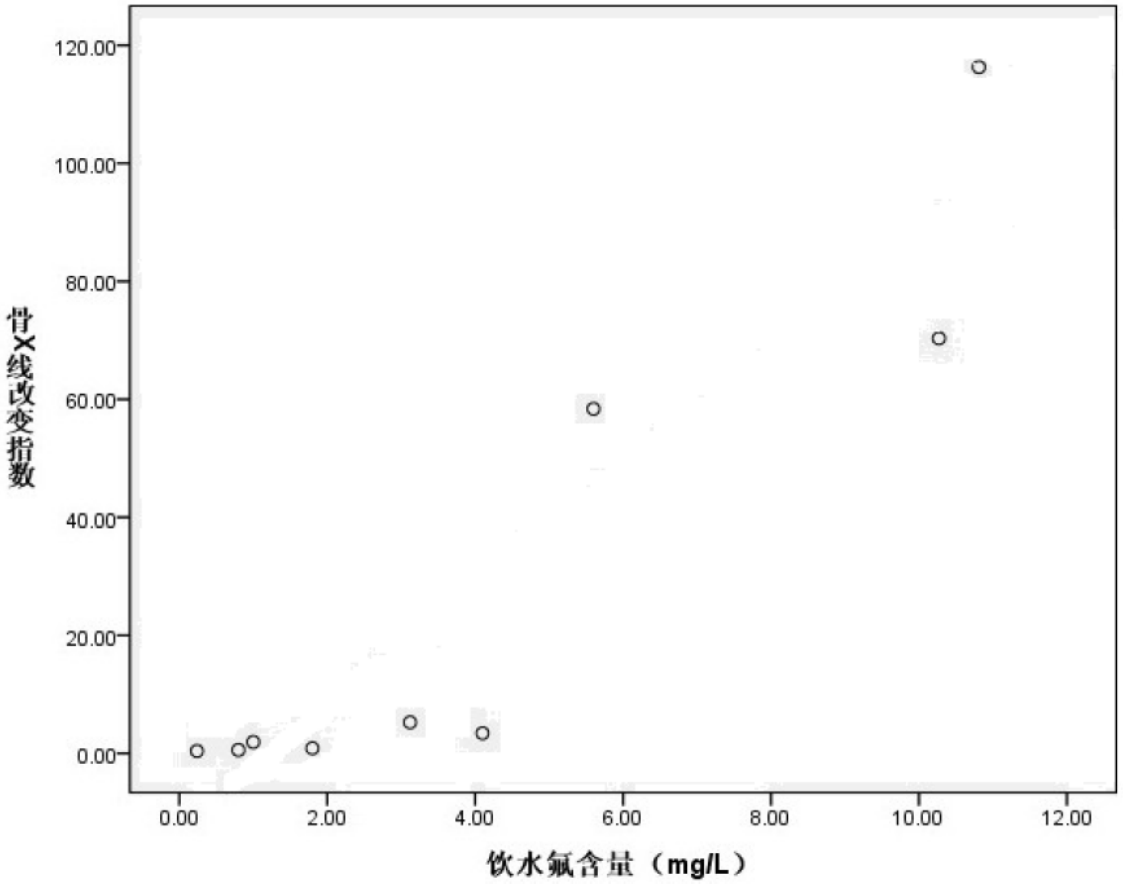
$$\hat{Y} = a + bX$$

- X: 自變量、解釋變量
- Y: 因變量、應變量、被解釋變量
- $\hat{Y}$: 因變量的預測值
- $Y - \hat{Y}$: 殘差
- a: 截距項、常數項
- b: 回歸係數

【例】 根據下表研究飲水氟含量與骨X綫改變指數的關係。

调查对象	饮水氟含量（*X*）	骨X线改变指数（*Y*）
1	0.24	0.40
2	0.80	0.56
3	1.00	1.91
4	1.80	0.86
5	3.12	5.25
6	4.10	3.40
7	5.60	58.38
8	10.27	70.33
9	10.81	116.30
合计	37.74	257.39

畫散點圖


120.00
100.00
80.00
60.00
40.00
20.00
0.00
骨×线改变指数
0.00
2.00
4.00
6.00
8.00
10.00
12.00
饮水氟含量（mg/L）

估計回歸方程

- 利用最小二乘原理: 殘差平方和最小
- $$\hat{Y} = -13.409 + 9.94X$$
- 得到樣本方程，總體是否有回歸關係?
- 假設檢驗

回歸方程的假設檢驗

- 建立檢驗假設並設定檢驗水平
  - $H_0: \beta = 0$
  - $H_1: \beta \neq 0$
  - $\alpha = 0.05$
- 計算卡方檢驗統計量: 方差分析

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12537.411	1	12537.411	43.324	.000[b]
	Residual	2025.728	7	289.390		
	Total	14563.139	8			

a. Dependent Variable: 骨X线改变指数

b. Predictors: (Constant), 饮水氟含量

- 計算P值，得出結論
  - $P < 0.05$，拒絕原假設，原回歸方程有統計學意義。

簡單回歸的其他問題

- 方程的整體擬合情況： 決定係數 R^2
- 決定係數跟簡單相關係數的關係
- 单个系数的假设检验
- 回歸分析的假定條件條件

多重線性回歸

【例】為研究大氣污染物一氧化氮（NO）的濃度是否受到汽車流量、氣候狀況等因素的影響，選擇24個工業水平相近的城市的一個交通點，統計：

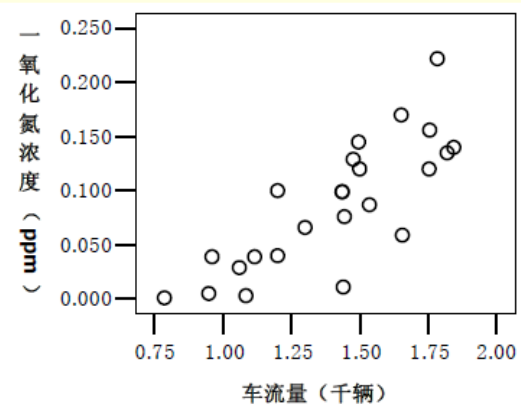
- 單位時間過往的汽車數（千輛）:X1
- 同時在低空的相同高度測定了該時間段平均氣溫（℃）:X2
- 空氣濕度（%）:X3
- 風速（m/s）:X4
- 空氣中一氧化氮（NO）的濃度（ppm）:Y

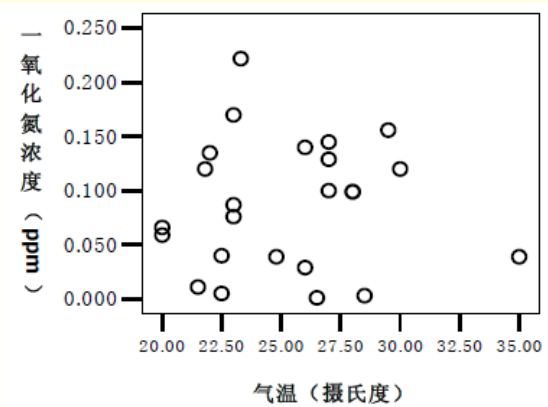
一氧化氮（Y）	车流量（X_1）	气温（X_2）	气湿（X_3）	风速（X_4）	一氧化氮（Y）	车流量（X_1）	气温（X_2）	气湿（X_3）	风速（X_4）
0.066	1.300	20.0	80	0.45	0.005	0.948	22.5	69	2.00
0.076	1.444	23.0	57	0.50	0.011	1.440	21.5	79	2.40
0.001	0.786	26.5	64	1.50	0.003	1.084	28.5	59	3.00
0.170	1.652	23.0	84	0.40	0.140	1.844	26.0	73	1.00
0.156	1.756	29.5	72	0.90	0.039	1.116	35.0	92	2.80
0.120	1.754	30.0	76	0.80	0.059	1.656	20.0	83	1.45
0.040	1.200	22.5	69	1.80	0.087	1.536	23.0	57	1.50
0.120	1.500	21.8	77	0.60	0.039	0.960	24.8	67	1.50
0.100	1.200	27.0	58	1.70	0.222	1.784	23.3	83	0.90
0.129	1.476	27.0	65	0.65	0.145	1.496	27.0	65	0.65
0.135	1.820	22.0	83	0.40	0.029	1.060	26.0	58	1.83
0.099	1.436	28.0	68	2.00	0.099	1.436	28.0	68	2.00

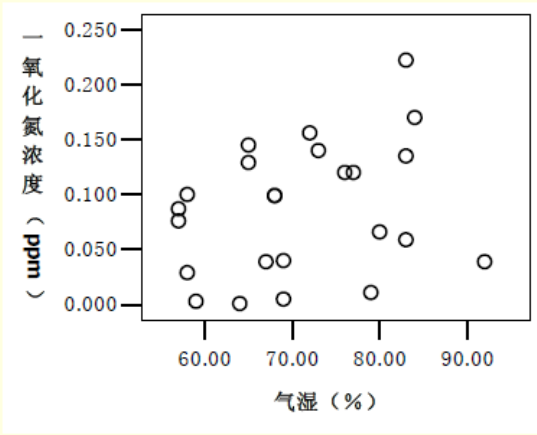
资料来源：数据选自《卫生统计学》第5版（方积乾主编）人民卫生出版社

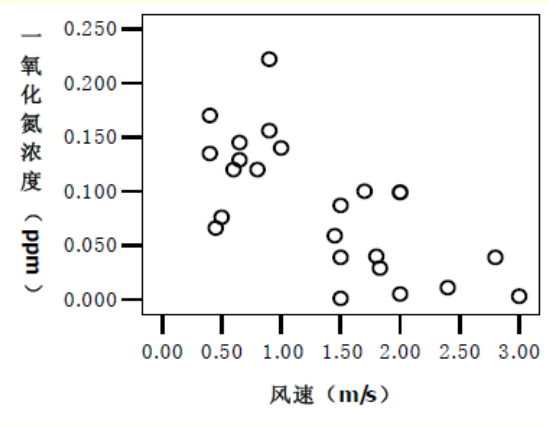
如何做多重線性回歸?

確認自變量與因變量的綫性關係



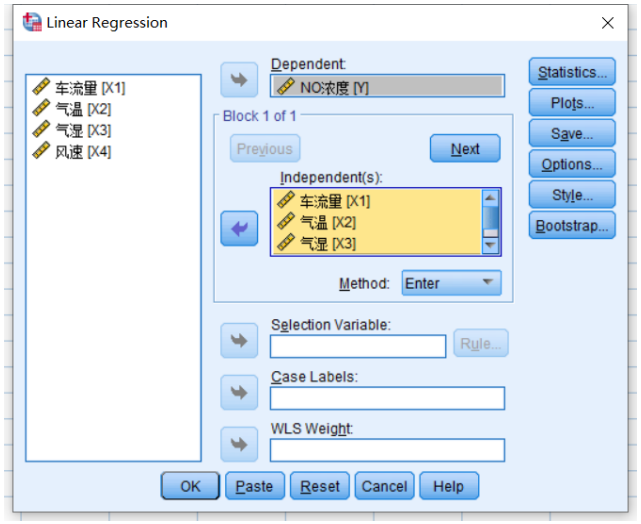






第一步：估計

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$



结果：

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.064	4	.016	17.590	.000[b]
	Residual	.017	19	.001		
	Total	.081	23			

a. Dependent Variable: NO浓度

b. Predictors: (Constant), 风速, 气湿, 气温, 车流量

Coefficients[a]

Double-click to activate

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-.142	.069		-2.048	.055	-.286	.003
	车流量	.116	.027	.592	4.227	.000	.059	.174
	气温	.004	.002	.273	2.364	.029	.001	.008
	气湿	-6.552E-6	.001	-.001	-.009	.993	-.001	.001
	风速	-.035	.011	-.448	-3.208	.005	-.057	-.012

a. Dependent Variable: NO浓度

估計結果

$$\hat{Y} = -0.142 + 0.116 \times X_1 + 0.004 \times X_2 + 0.000006552 \times X_3 - 0.035 \times X_4$$

估計方法：最小二乘法

最小二乘法：殘差平方和最小

第二步：檢驗

1. 檢驗整個方程是否有意義

2. 檢驗每一個偏回歸係數是否有意義

检验方程是否有统计学意义？

- 第一步：建立 H_0 和 H_1.
  - H_0:
    $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
  - H_1： 至少有一個 $\beta_i \neq 0$.
  - $\alpha = 0.05$
- 第二步：F統計量见右边方差分析表，在 H_0成立的條件下，F統計量服從自由度。。。
- 第三步：結論，p<0.05，在0.05的檢驗水平下，拒絕 H_0。。。

方差分析表：

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.064	4	.016	17.590	.000[b]
	Residual	.017	19	.001		
	Total	.081	23			

a. Dependent Variable: NO浓度

b. Predictors: (Constant), 风速, 气湿, 气温, 车流量

$$\hat{Y} = -0.142 + 0.116 \times X_1 + 0.004 \times X_2 + 0.000006552 \times X_3 - 0.035 \times X_4$$

每个偏回归系数是否有统计学意义

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-.142	.069		-2.048	.055	-.286	.003
	车流量	.116	.027	.592	4.227	.000	.059	.174
	气温	.004	.002	.273	2.364	.029	.001	.008
	气湿	-6.552E-6	.001	-.001	-.009	.993	-.001	.001
	风速	-.035	.011	-.448	-3.208	.005	-.057	-.012

a. Dependent Variable: NO浓度

- 注意與一元綫性回歸的區別
- 標準化偏回歸係數的解釋

模型評估

- 決定係數： R^2
- 調整的決定係數

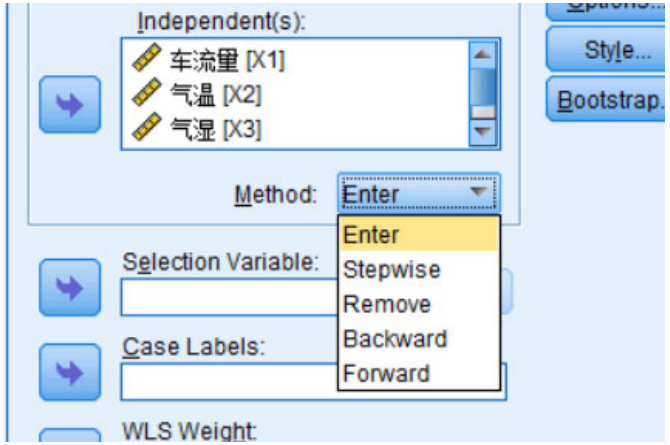
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.887[a]	.787	.743	.0301501

a. Predictors: (Constant), 风速, 气湿, 气温, 车流量

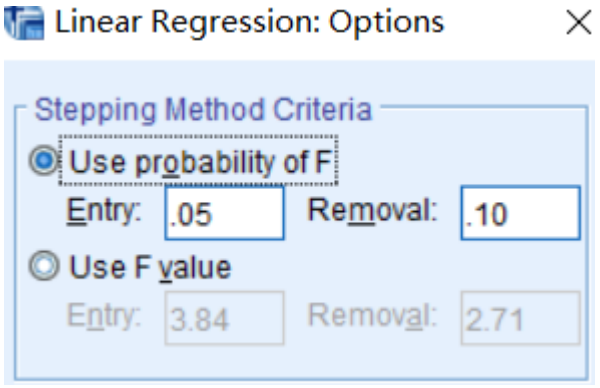
模型選擇

spss中：



对应方法：

- 逐步法
- 剔除法
- 后向法
- 前向法



前向法的結果

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-.135	.035		-3.829	.001	-.209	-.062
	车流量	.158	.025	.808	6.432	.000	.107	.210
2	(Constant)	-.050	.049		-1.027	.316	-.151	.051
	车流量	.122	.027	.623	4.476	.000	.065	.179
	风速	-.025	.011	-.325	-2.338	.029	-.048	-.003
3	(Constant)	-.142	.058		-2.452	.024	-.263	-.021
	车流量	.116	.025	.592	4.699	.000	.065	.168
	风速	-.035	.010	-.448	-3.316	.003	-.057	-.013
	气温	.004	.002	.273	2.430	.025	.001	.008

a. Dependent Variable: NO浓度

$$\hat{Y} = -0.142 + 0.116 \times X_1 + 0.004 \times X_2 - 0.035 \times X_4$$

线性模型的假定

- 綫性（L）
- 獨立性（I）
- 正態性（N）
- 方差齊性（E）

謝謝大家！