第17章 线性回归模型入门

学习目标

- 掌握简单线性回归分析的原理
- 能用SPSS做简单线性回归分析
- 能用SPSS做多重线性回归分析

主要内容

- 17.1 线性回归模型简介
- 17.2 案例
- 17.3 多重线性回归模型入门

17.1 线性回归模型简介

相关分析与回归分析的联系与区别

- 相关分析是研究事物或现象之间有无关系以及关系的方向和密切程度的分析方法。
- 回归分析是研究事物或现象之间数量依存的关系的分析方法。
  - 研究一个连续性变量（因变量）的取值随着其它变量（自变量）的数值变化而变化的趋势。
  - 通过回归方程解释两变量之间的关系显得更为精确，可以计算出自变量改变一个单位时因变量平均改变的单位数量，这是相关分析无法做到的。
  - 除了描述两变量的关系以外，通过回归方程还可以进行预测和控制，这在实际工作中尤为重要。

简单回归分析的原理和要求

线性回归方程

回归直线 $\hat{y}=\hat{a}+\hat{b}x$

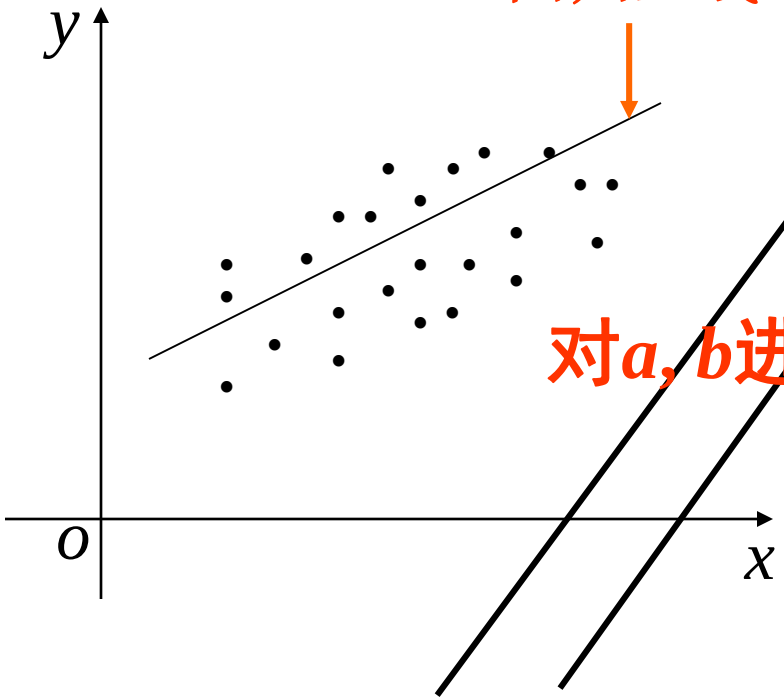
容量为 n 的二维样本：

(x_1, y_1)

(x_2, y_2)

……

(x_n, y_n)



对a, b进行估计

一元线性回归模型 $\begin{cases} y=a+bx+\varepsilon, \\ \varepsilon \sim N(0,\sigma^2), \quad a,b,\sigma^2\text{为常数}. \end{cases}$

简单回归分析的原理和要求

- 如何估计***a***、***b***?
  - 最小二乘法：使各实测点距回归直线的纵向距离的平方和即残差的平方和达到最小的 ***a*** 和 ***b*** 是最优的。此平方和是关于***a***、***b*** 的二元函数 ***Q*(*a*, *b*)**。

$$l_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$l_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{b} = l_{xy} / l_{xx}$$

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

简单回归分析的原理和要求

- **相关系数与回归的显著性检验**

  即使平面上 n 个杂乱无章的样本点也可以得到回归方程，但实际上此时的回归方程毫无意义！

  究竟在什么情况下所配的回归直线才有意义，回归方程真的揭示了 X 和 Y 之间存在线性关系的内在规律？

  问题1： X 和 Y 之间是否有线性回归函数？ 回归显著性检验

  问题2： X 和 Y 之间如果有线性回归函数，是否可以用某个指标来描述 X 和 Y 之间的线性关系的密切程度呢？

  相关系数
  确定系数

简单回归分析的原理和要求

- 平方和分解公式

这里的平方和指的是 y_i 与其均值 $\bar{y}$ 偏差的平方和，反映了y_i 的数据分散程度

$$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\ldots=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$SS_{剩}$ 剩余平方和

$SS_{回}$ 回归平方和

注：$\bar{y}$ 也是 $\hat{y}_i$ 的均值

简单回归分析的原理和要求

- 回归直线拟合程度的度量

  样本相关系数：

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

反映*X*和*Y*线性关系方向和程度的指标！

$$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\ldots=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$SS_{剩}$ 剩余平方和

$SS_{回}$ 回归平方和

$SS_{回}$在S_{yy}中占的比例越大，说明*X*和*Y*线性关系越强

$$\frac{SS_{回}}{S_{yy}}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\ldots=\frac{l_{xy}^{\ 2}}{l_{xx}l_{yy}}=r^2$$

称为确定系数
或决定系数

简单回归分析的原理和要求

- 线性回归的显著性检验

  总体上X与Y之间是否存在显著的线性关系呢？可进行如下检验：

$$H_0: b = 0;\quad H_1: b \neq 0$$

SPSS中提供了检验方法：F检验

F检验

统计量 $F = (n-2)\dfrac{SS_{回}}{SS_{剩}} \sim F(1, n-2)$

简单回归分析的原理和要求

- **个体 *Y* 预测值的区间估计**

  给定自变量 X 的某一个值 X_0，以一定的置信度预测对应的 Y 的观察值 Y_0 的取值范围. 这种预测的取值范围称为预测区间.

  Y_0的置信度为1-a的置信区间是：

$$(\hat{Y}_0-\delta(X_0),\hat{Y}_0+\delta(X_0)) \quad \text{其中}$$

$$\hat{Y}_0=\hat{a}+\hat{b}X_0$$

$$\delta(X_0)=t_{\frac{\alpha}{2}}(n-2)\bullet\sqrt{\frac{SS_{剩}}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(X_0-\overline{X})^2}{l_{xx}}}$$

回归模型的适用条件

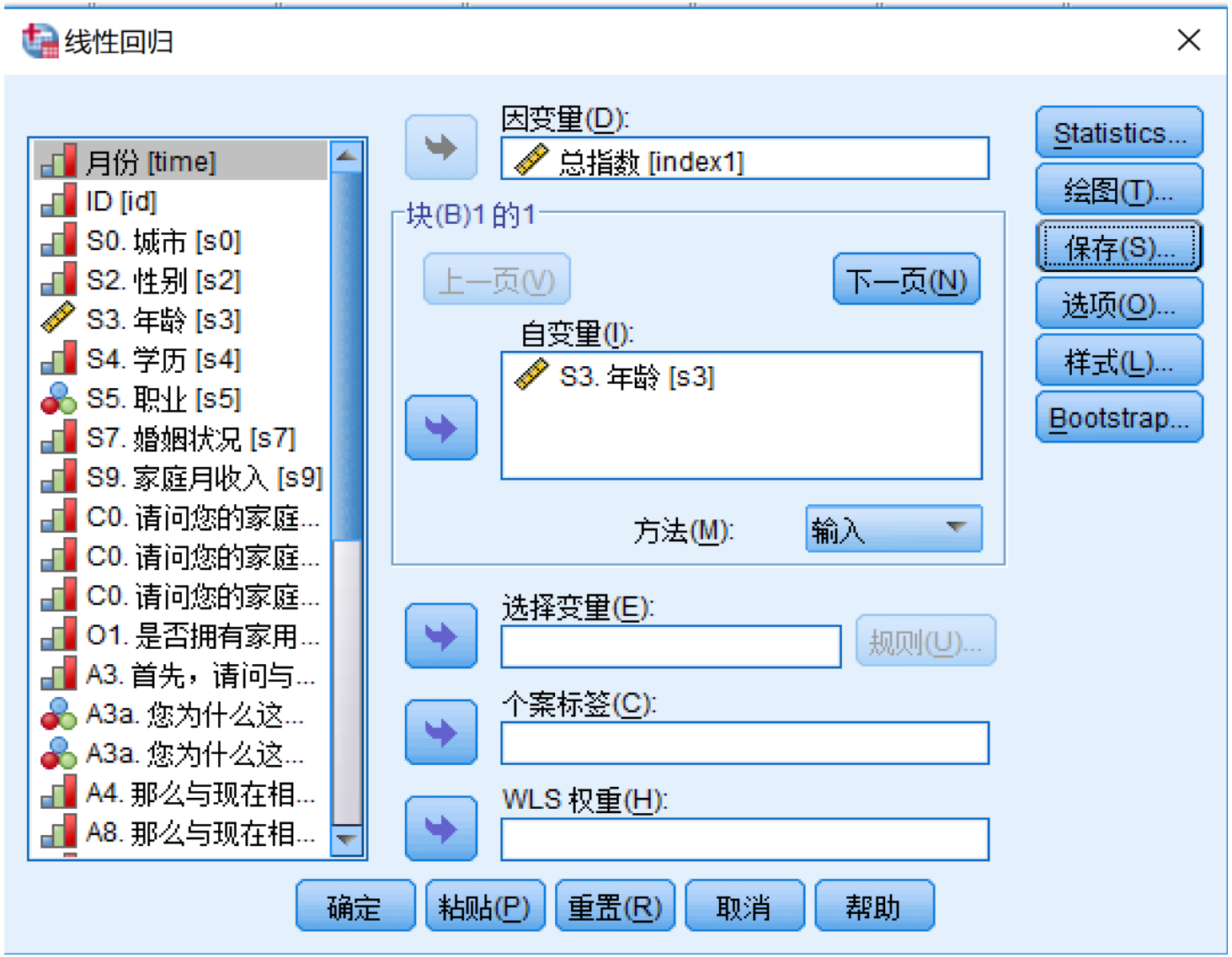
- 线性趋势：自变量与应变量总体上呈现线性关系
- 独立性：因变量*Y*的取值要相互独立，反映到模型中实际上要求残差间相互独立
- 正态性：就自变量的任何一个线性组合，因变量*Y*均服从正态分布，反映到模型中实际上要求残差服从正态分布
- 方差齐性：就自变量的任何一个线性组合，因变量*Y*的方差均相同，反映到模型中实际上要求残差的方差是齐性的

17.2 案例

案例1

- 数据文件：CCSS_Sample.sav
- 要求：建立用年龄预测总信心指数的回归方程
- 具体过程
  - “分析”→“回归”→“线性”
  - 将“index1”选入因变量
  - 将“年龄”选入自变量
  - 其余默认

案例1


线性回归
月份 [time]
ID [id]
S0. 城市 [s0]
S2. 性别 [s2]
S3. 年龄 [s3]
S4. 学历 [s4]
S5. 职业 [s5]
S7. 婚姻状况 [s7]
S9. 家庭月收入 [s9]
C0. 请问您的家庭...
C0. 请问您的家庭...
C0. 请问您的家庭...
O1. 是否拥有家用...
A3. 首先，请问与...
A3a. 您为什么这...
A3a. 您为什么这...
A4. 那么与现在相...
A8. 那么与现在相...
因变量(D):
总指数 [index1]
块(B)1 的1
上一页(V)
下一页(N)
自变量(I):
S3. 年龄 [s3]
方法(M):
输入
选择变量(E):
规则(U)...
个案标签(C):
WLS 权重(H):
Statistics...
绘图(T)...
保存(S)...
选项(O)...
样式(L)...
Bootstrap...
确定
粘贴(P)
重置(R)
取消
帮助

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	S3. 年龄[b]	.	Enter

a. Dependent Variable: 总指数

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.219[a]	.048	.047	20.49596

a. Predictors: (Constant), S3. 年龄

相关系数为0.219，决定系数为0.048

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24249.673	1	24249.673	57.726	.000[b]
	Residual	480996.625	1145	420.084		
	Total	505246.298	1146			

a. Dependent Variable: 总指数

b. Predictors: (Constant), S3. 年龄

对回归方程显著性进行假设检验（F检验），P<0.05，则模型有统计学意义。

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	108.898	1.816		59.982	.000
	S3. 年龄	-.358	.047	-.219	-7.598	.000

a. Dependent Variable: 总指数

常数项为108.898，一次项系数为-0.358，对一次项系数进行假设检验的P<0.05，则认为总体的一次项系数非0，一元线性回归方程存在：y = -0.358x + 108.898

在“线性回归”界面，点击“保存”按钮，选择“预测值”的“未标准化”、“预测区间”的“单值”，在数据窗口会新增3列：PRE_1, LICI_1, UICI_1, 分别为每条案例的模型预测值、个体预测值95%参考值区间的下界和上界。

PRE_1	LICI_1	UICI_1	变量
101.74463	61.48488	142.00437	
91.72945	51.48370	131.97520	
87.79492	47.50922	128.08062	
101.02926	60.77602	141.28249	
100.31389	60.06632	140.56146	
90.29871	50.04140	130.55603	
101.38694	61.13056	141.64332	
102.10231	61.83900	142.36563	
101.74463	61.48488	142.00437	
99.95620	59.71115	140.20126	
86.72186	46.42083	127.02290	
99.59852	59.35576	139.84127	
85.64881	45.33054	125.96708	
91.72945	51.48370	131.97520	
101.74463	61.48488	142.00437	
98.52546	58.28834	138.76258	
94.59093	54.35814	134.82372	
98.88315	58.64436	139.12193	
100.31389	60.06632	140.56146	
92.80251	52.56321	133.04180	
91.01408	50.76297	131.26519	

17.3 多重线性回归模型入门

模型简介

回归模型：$y = a + b_1 x_1 + b_2 x_2 + ... + b_n x_n + \varepsilon$

$$\varepsilon \sim N(0, \sigma^2)$$

a：常数项或截距，即回归直线在y轴上的截距。

b_i：偏回归系数，简称回归系数。

残差ε：y的实测值与估计值的差。

自变量的筛选方法

- 进入法：所有选入“自变量”框中的变量均进入模型
- 向前法：一步一步往模型中添加自变量
- 向后法：一步一步将模型中的自变量剔除
- 逐步回归法：向前法与向后法的结合
- 删除法：强行将自变量剔除出模型

案例2

- 数据文件：CCSS_Sample.sav
- 要求：建立自变量包括年龄、性别、家庭收入的信心指数回归方程
- 具体过程
  - “分析”→“回归”→“线性”
  - 将“index1”选入因变量
  - 将“年龄”、“性别”、“Qs9”选入自变量
  - 方法下拉列表框中选择“向后法”
  - 其余默认

案例2


线性回归
C0. 请问您的家庭...
O1. 是否拥有家用...
A3. 首先，请问与...
A3a. 您为什么这...
A3a. 您为什么这...
A4. 那么与现在相...
A8. 那么与现在相...
A9. 那么您认为一...
A10. 那么与现在...
A16. 对于大宗耐...
家庭收入2级 [Ts9]
Qs9
Qa3
Qa4
Qa8
Qa9
Qa10
Qa16
因变量(D):
总指数 [index1]
块(B)1 的1
上一页(V)
下一页(N)
自变量(I):
S3. 年龄 [s3]
S2. 性别 [s2]
Qs9
方法(M):
后退
选择变量(E):
规则(U)...
个案标签(C):
WLS 权重(H):
Statistics...
绘图(T)...
保存(S)...
选项(O)...
样式(L)...
Bootstrap...
确定
粘贴(P)
重置(R)
取消
帮助

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Qs9, S2. 性别, S3. 年龄[b]	.	Enter
2	.	S2. 性别	Backward (criterion: Probability of F-to-remove >= .100).

a. Dependent Variable: 总指数

b. All requested variables entered.

向后法

模型中要剔除变量“性别”(因为P>0.1)，“性别”和“index1”在统计学上无线性相关关系。

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.231[a]	.053	.050	20.93061
2	.231[b]	.053	.051	20.92005

a. Predictors: (Constant), Qs9, S2. 性别, S3. 年龄

b. Predictors: (Constant), Qs9, S3. 年龄

c. Dependent Variable: 总指数

调整后的决定系数增大，说明模型2优于模型1.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24360.728	3	8120.243	18.536	.000[b]
	Residual	432833.227	988	438.090		
	Total	457193.955	991			
2	Regression	24359.741	2	12179.871	27.830	.000[c]
	Residual	432834.214	989	437.648		
	Total	457193.955	991			

a. Dependent Variable: 总指数

b. Predictors: (Constant), Qs9, S2. 性别, S3. 年龄

c. Predictors: (Constant), Qs9, S3. 年龄

两个模型在总体上均有统计学意义。

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	108.238	2.960		36.565	.000
	S3. 年龄	-.362	.052	-.217	-6.948	.000
	S2. 性别	.064	1.339	.001	.047	.962
	Qs9	.000	.000	.054	1.721	.086
2	(Constant)	108.330	2.238		48.409	.000
	S3. 年龄	-.362	.052	-.217	-6.952	.000
	Qs9	.000	.000	.054	1.721	.085

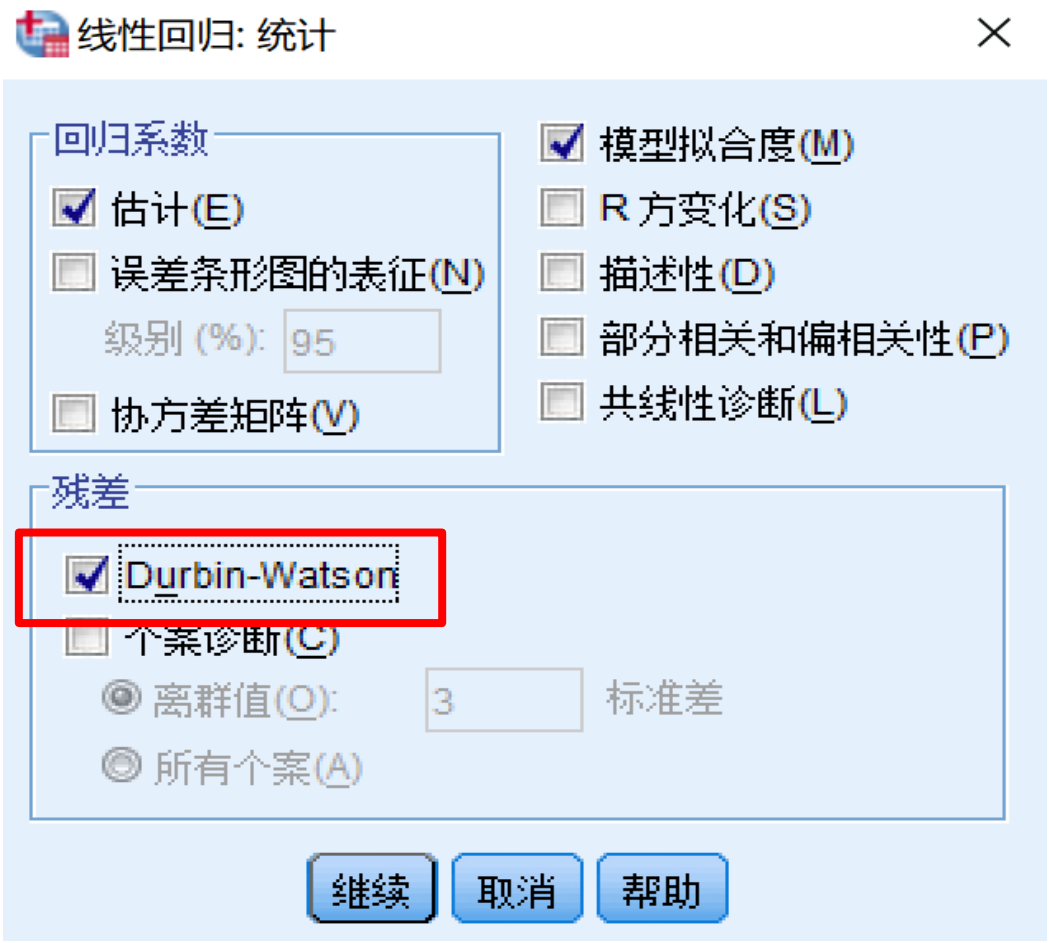
a. Dependent Variable: 总指数

常数项为108.330，“年龄”的系数为-0.362，“家庭收入”的系数为0.000165。本次检验的显著水平为0.1，因此保留了“家庭收入”这个变量。

案例2

- 上述过程得到了回归方程，并进行了相应的假设检验。
- 以下对模型的适用条件进行检验：
  - 残差独立性：通过“线性回归”界面“统计量”中的“Durbin-Watson检验”来完成
  - 残差正态性：通过“线性回归”界面“绘图”中的“直方图”和“正态概率图”来完成
  - 残差方差齐性：通过“线性回归”界面“绘图”中的散点图来完成

案例2 — 残差独立性检验


线性回归: 统计
回归系数
估计(E)
误差条形图的表征(N)
级别 (%): 95
协方差矩阵(V)
模型拟合度(M)
R 方变化(S)
描述性(D)
部分相关和偏相关性(P)
共线性诊断(L)
残差
Durbin-Watson
个案诊断(C)
离群值(O): 3
标准差
所有个案(A)
继续
取消
帮助

案例2 — 残差独立性检验

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.231[a]	.053	.050	20.93061	
2	.231[b]	.053	.051	20.92005	2.016

a. Predictors: (Constant), Qs9, S2. 性别, S3. 年龄

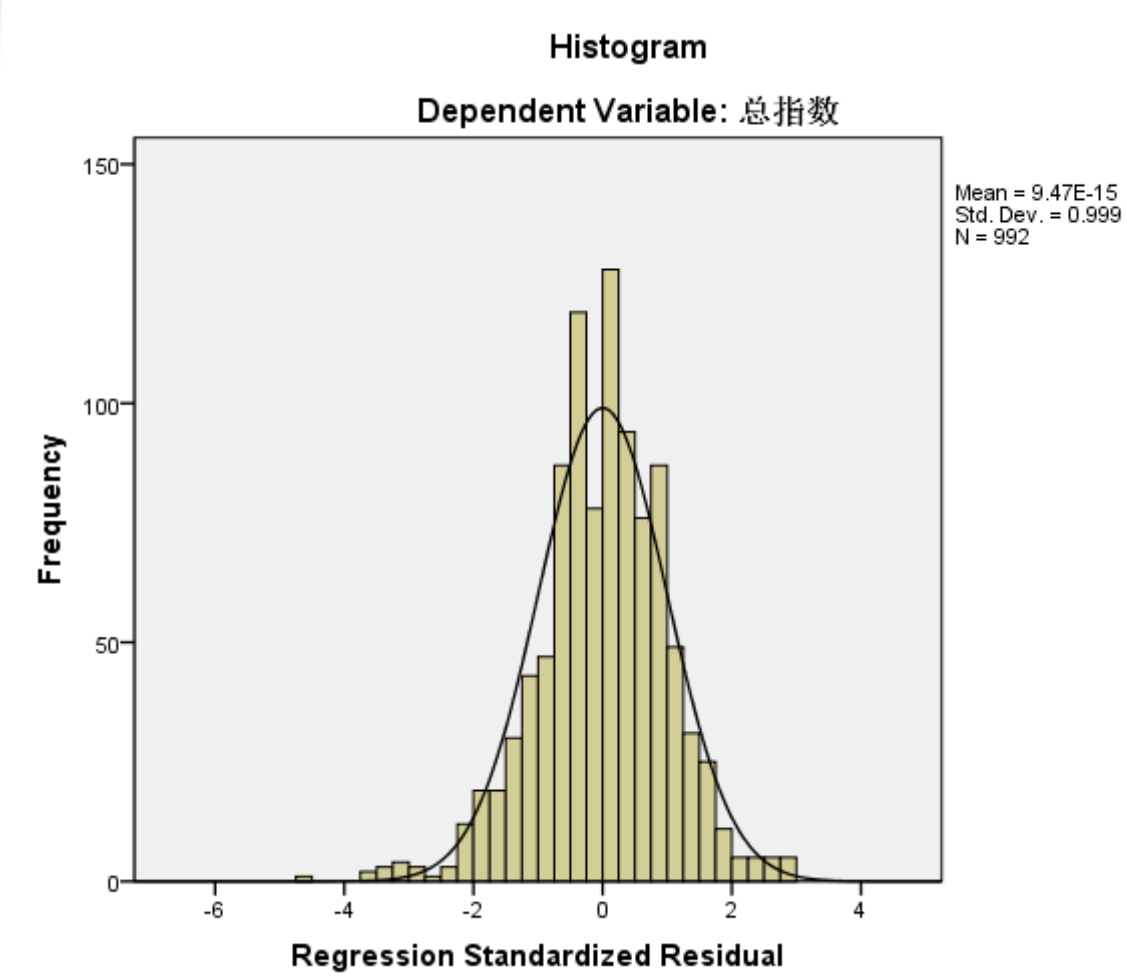
b. Predictors: (Constant), Qs9, S3. 年龄

c. Dependent Variable: 总指数

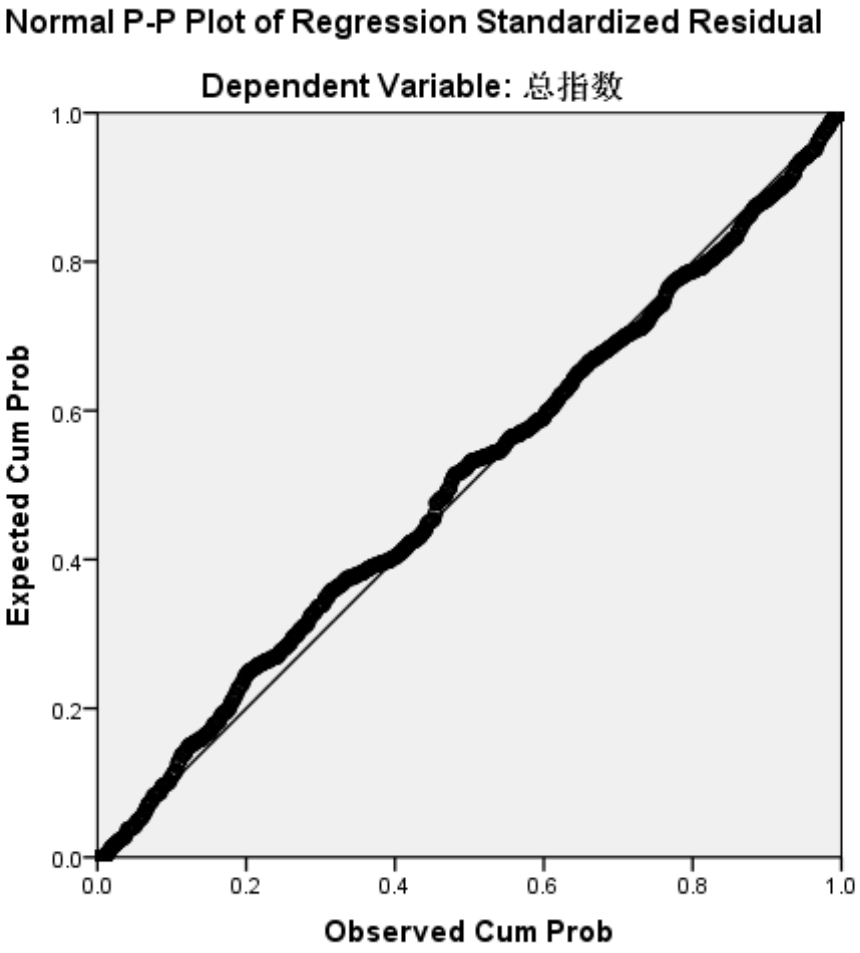
D-K统计量的取值在0~4之间，若>4,残差间相互独立，若<0,残差间存在自相关性。当统计量在1-3之间取值时，基本能判断残差间相互独立。

案例2 — 残差正态性检验


线性回归: 图
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
散点 1 的 1
上一页(V)
下一页(N)
Y:
X:
标准化残差图
直方图(H)
正态概率图(R)
产生所有部分图(P)
继续
取消
帮助

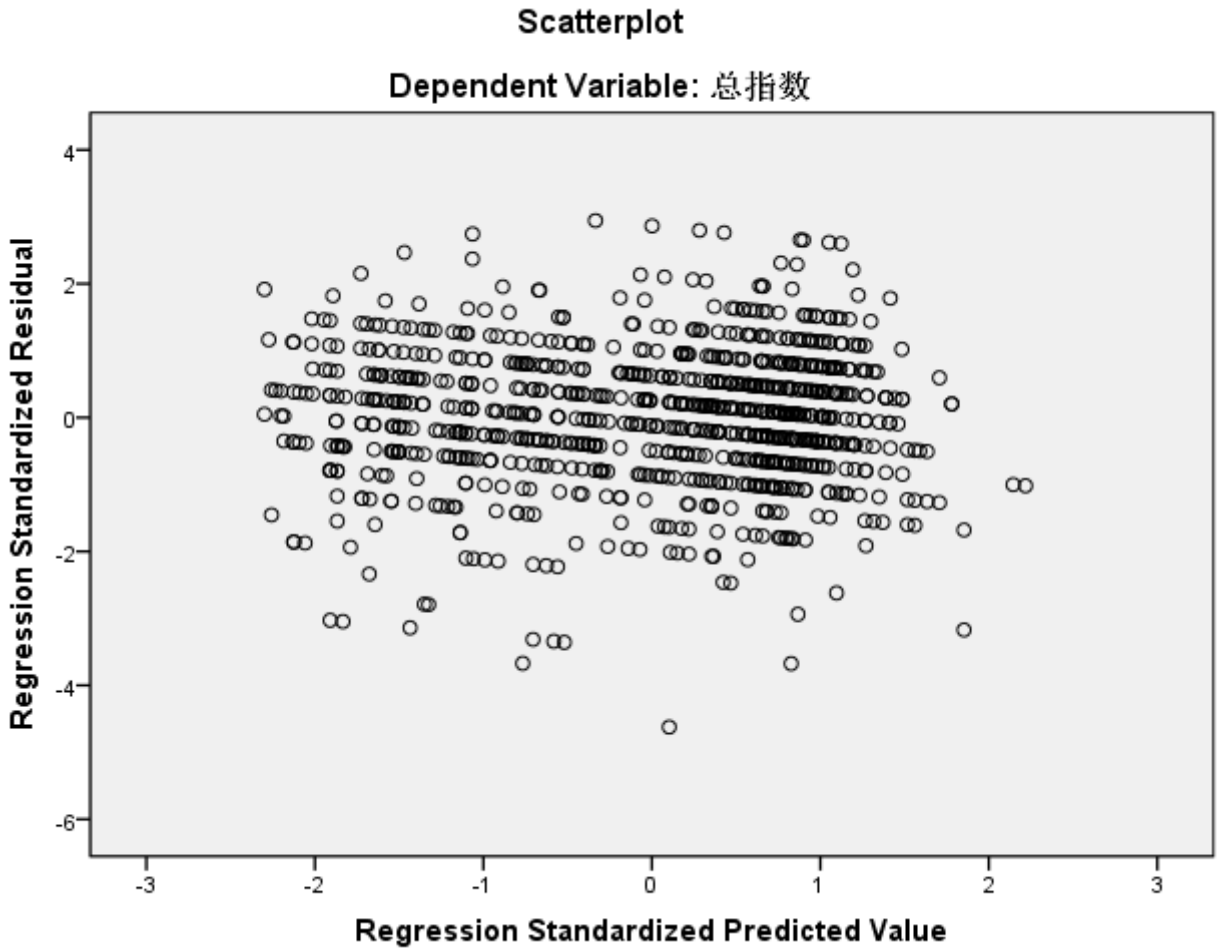


上图是标准化残差的直方图，观察其与正太曲线的拟合程度，从而判断残差是否正态分布。



上图是标准化残差的P-P图，观察两条轨迹的拟合程度，从而判断残差是否正态分布。

案例2 — 残差的方差齐性检验



在“绘图”界面，将“标准化残差”选入Y轴，将标准化预测值选入X轴，绘制散点图。理想状态是残差的取值介于-3~3之间，且在X轴和Y轴方向分布均匀。

THE END