数学实践——

二、线性回归分析

回归分析概述

(一)回归分析理解

(1)“回归”的含义：galton研究父亲身高和儿子身高的关系时的独特发现.

(2)回归线的获得方式一:局部平均

- **回归曲线上的点给出了相应于每一个x(父亲)值的y(儿子)平均数的估计**

(3)回归线的获得方式二:拟和函数

- **使数据拟和于某条曲线;**
- **通过若干参数描述该曲线;**
- **利用已知数据在一定的统计准则下找出参数的估计值(得到回归曲线的近似);**

回归分析概述

(二)回归分析的基本步骤

- **(1)确定自变量和因变量(父亲身高关于儿子身高的回归与儿子身高关于父亲身高的回归是不同的).**
- **(2)从样本数据出发确定变量之间的数学关系式,并对回归方程的各个参数进行估计.**
- **(3)对回归方程进行各种统计检验.**
- **(4)利用回归方程进行预测.**

回归分析概述

(三)参数估计的准则

- **目标:观察值与回归线上的预测值之间的距离总和达到最小**
- **最小二乘法(利用最小二乘法拟和的回归直线与样本数据点在垂直方向上的偏离程度最低)**

一元线性回归分析

(一)一元线性回归模型:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

β_0 和 β_1 都是模型中的未知参数，β_0 和 β_1 分别称为回归常数和回归系数。

ε 称为随机误差，是一个随机变量，应当满足

$$\begin{cases} E(\varepsilon) = 0 \\ \mathrm{Var}(\varepsilon) = \sigma^2 \end{cases}$$

一元线性回归分析

一元线性回归方程

$$E(y) = \beta_0 + \beta_1 x$$

β_1:x每变动一个单位所引起的y的平均变动

(二)一元回归分析的步骤

- **利用样本数据建立回归方程(最小二乘估计)**
- **回归方程的拟和优度检验**
- **回归方程的显著性检验(t检验和F检验)**
- **残差分析**
- **预测**

回归参数的普通最小二乘估计

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}[y_i-E(y_i)]^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

最小二乘估计是寻找参数 β_0、β_1 的估计值 $\hat{\beta}_0$、$\hat{\beta}_1$，

$$Q(\hat{\beta}_0,\hat{\beta}_1)=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2=\min_{\beta_0,\beta_1}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

用最小二乘法(Ordinary Least Square, OLS)求解方程中的两个参数，得到：

$$\hat{\beta}_1=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

$$\hat{\beta}_0=\bar{y}-b\bar{x}$$

一元线性回归方程的检验

(一)拟和优度检验:

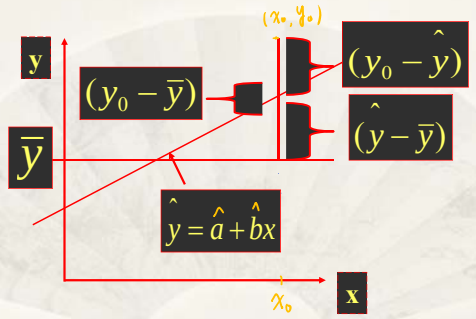
(1)目的:检验样本观察点聚集在回归直线周围的密集程度，评价回归方程对样本数据点的拟和程度

(2)思路:

- **因为: 因变量取值的变化受两个因素的影响**
  - **由于x的取值不同，使得与x有线性关系的y值不同；随机因素的影响**
- **于是: 因变量总变差=自变量引起的+其他因素引起的**

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i,\quad i=1,2,\cdots,n$$

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



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

即：总离差平方和（SST)=剩余离差平方和(SSE) +回归离差平方和（SSR)

其中；SSR是由x和y的直线回归关系引起的，可以由回归直线做出解释；SSE是除了x对y的线性影响之外的随机因素所引起的Y的变动，是回归直线所不能解释的。

回归平方和在总离差平方和中所占的比例可以作为一个统计指标，用来衡量X与Y 的关系密切程度以及回归直线的代表性好坏，称为可决系数（判定系数、决定系数） 。

■对于一元线性回归方程：

$$R^2=\frac{SSR}{SST}=\frac{SST-SSE}{SST}=1-\frac{SSE}{SST}$$

$$R^2=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}=1-\frac{\sum(y_i-\hat{y})^2}{\sum(y_i-\bar{y})^2}$$

一元线性回归方程的检验

- **(3)统计量：判定系数** $R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$
- **R^2=SSR/SST=1-SSE/SST.**
- **R^2体现了回归方程所能解释的因变量变差的比例;1-R^2则体现了因变量总变差中，回归方程所无法解释的比例。**
- **R^2越接近于1，则说明回归平方和占了因变量总变差平方和的绝大部分比例，因变量的变差主要由自变量的不同取值造成，回归方程能够较好拟合样本数据点**
  - **在一元回归中$R^2=r^2$; 因此，从这个意义上讲，判定系数能够比较好地反映回归直线对样本数据的代表程度和线性相关性。**

一元线性回归方程的检验

(二)回归方程的显著性检验：F检验

(1)目的:检验自变量与因变量之间的线性关系是否显著,是否可用线性模型来表示.

(2)H_0:β=0 即:回归系数与0无显著差异

(3)利用F检验,构造F统计量:

- **F=平均的回归平方和/平均的剩余平方和~F(1,n-1-1)**
- **如果F值较大，则说明自变量造成的因变量的线性变动远大于随机因素对因变量的影响,自变量于因变量之间的线性关系较显著**

$$F=\frac{SSR/1}{SSE/(n-2)}=\frac{\sum(\hat{y}-\bar{y})^2/1}{\sum(y-\hat{y})^2/(n-2)}\overset{H_0}{\sim}F(1,n-2)$$

一元线性回归方程的检验

(4)计算F统计量的值和相伴概率p

(5)判断

- **p<=a:拒绝H_0,即:回归系数与0有显著差异，自变量与因变量之间存在显著的线性关系。反之，不能拒绝H_0**

一元线性回归方程的检验

(三)回归系数的显著性检验:t检验

(1)目的:检验自变量对因变量的线性影响是否显著.

(2)H_0:β=0 即:回归系数与0无显著差异

(3)利用t检验,构造t统计量:

$$t=\frac{\hat{\beta}_1}{\frac{\hat{\sigma}}{\sqrt{\sum(x_i-\bar{x})^2}}}\sim t(n-2)$$

$$\text{其中，}\hat{\sigma}=S_y=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}$$

一元线性回归方程的检验

- **其中:S_y是回归方程标准误差(Standard Error)的估计值，由均方误差开方后得到，反映了回归方程无法解释y变动的程度。**
- **如果回归系数的标准误差较小，必然得到一个相对较大的t值，表明该自变量x解释因变量线性变化的能力较强。**

(4)计算t统计量的值和相伴概率p

(5)判断

一元线性回归方程的检验

(四)t检验与F检验的关系

- **一元回归中,F检验与t检验一致,即: $F=t^2$,两种检验可以相互替代**

(五)F统计量和R^2值的关系

$$F=\frac{R^2/k}{(1-R^2)/(n-k-1)}$$

- **如果回归方程的拟合优度高，F统计量就越显著。F统计量越显著，回归方程的拟合优度就会越高。**

线性回归方程的残差分析

- 残差是指由回归方程计算得到的预测值与实际样本值之间的差距，定义为：

$$e_i=y_i-\hat{y}_i,\quad i=1,\cdots,n$$

对于线性回归分析来讲，如果方程能够较好的反映被解释变量的特征和规律性，那么残差序列中应不包含明显的规律性。残差分析包括以下内容：残差服从正态分布，其平均值等于0；残差取值与X的取值无关；残差不存在自相关；残差方差相等。

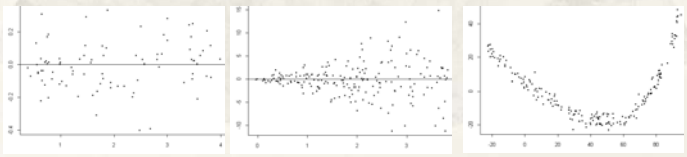
线性回归方程的残差分析

(一)残差序列的正态性检验:

- **绘制标准化残差的直方图或累计概率图**

(二)残差序列的随机性检验

- **绘制残差和预测值的散点图,应随机分布在经过零的一条直线上下**



残差均值和方差齐性检验

线性回归方程的残差分析

(三)残差序列独立性检验:

- **残差序列是否存在后期值与前期值相关的现象,利用D.W(Durbin-Watson)检验**

H_0：总体的自相关系数 ρ 与零无显著差异。

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=2}^{n} e_t^2}$$

DW 取值在 0～4 之间。

线性回归方程的残差分析

- **d-w=0:残差序列存在完全正自相关;d-w=4:残差序列存在完全负自相关;0<d-w<2:残差序列存在某种程度的正自相关;2<d-w<4:残差序列存在某种程度的负自相关;d-w=2:残差序列不存在自相关.**

- **残差序列不存在自相关,可以认为回归方程基本概括了因变量的变化;否则,认为可能一些与因变量相关的因素没有引入回归方程、回归模型不合适、不应选用线性模型或滞后性周期性的影响.**

线性回归方程的残差分析

(四)异常值(casewise或outliers)诊断

- **可以利用残差分析探测样本中的异常值。通常异常值是指那些远离均值的样本数据点，它们对回归方程的参数估计有较大影响，应尽量找出它们并加以排除。被解释变量y 和解释变量x中都有可能出现异常值。**

线性回归方程的残差分析

- **利用标准化残差不仅可以知道观察值比预测值大或小,并且还知道在绝对值上它比大多数残差是大还是小.一般标准化残差的绝对值大于3,则可认为对应的样本点为奇异值**
- **异常值并不总表现出上述特征.当剔除某观察值后，回归方程的标准差显著减小,也可以判定该观察值为异常值**

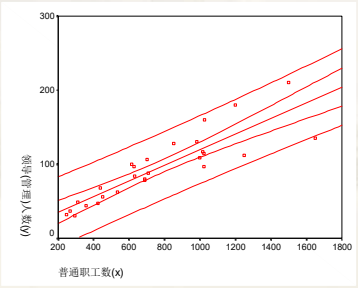
线性回归方程的预测

(一)点估计

y_0

(二)区间估计

x_0为x_i的均值时,预测区间最小,精度最高.x_0越远离均值,预测区间越大,精度越低.



多元线性回归分析

(一)多元线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

β_0，β_1，…，β_p 都是模型中的未知参数，分别称为回归常数和偏回归系数。

$$\begin{cases} E(\varepsilon) = 0 \\ \mathrm{Var}(\varepsilon) = \sigma^2 \end{cases}$$

多元线性回归方程

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

估计的多元线性回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

多元线性回归分析

最小二乘估计：

$$Q(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

$$= \min_{\beta_0, \beta_1, \beta_2, \cdots, \beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

(二)多元线性回归分析的主要问题

- **回归方程的检验**
- **自变量筛选**
- **多重共线性问题**

多元线性回归方程的检验

(一)拟和优度检验：

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSE}{SST} \qquad \overline{R}^2 = 1 - \frac{\text{均方误差}}{\text{因变量的样本方差}}$$

(1)判定系数R^2：

- **R是y和x_i的复相关系数(或观察值与预测值的相关系数),测定了因变量y与所有自变量全体之间线性相关程度**

(2)调整的R^2

- **考虑的是平均的剩余平方和,克服了因自变量增加而造成R^2也增大的弱点**
- **在某个自变量引入回归方程后，如果该自变量是理想的且对因变量变差的解释说明是有意义的，那么必然使得均方误差减少，从而使调整的R^2得到提高；反之，如果某个自变量对因变量的解释说明没有意义，那么引入它不会造成均方误差减少，从而调整的R^2也不会提高。**

多元线性回归方程的检验

(二)回归方程的显著性检验：

$$F = \frac{\sum(\hat{y}_i - \bar{y})^2 / k}{\sum(y_i - \hat{y}_i)^2 / (n-k-1)}$$

(1)目的:检验所有自变量与因变量之间的线性关系是否显著，是否可用线性模型来表示.

(2)H_0:$\beta_1=\beta_2=\ldots=\beta_k=0$ 即:所有回归系数同时与0无显著差异

(3)利用F检验,构造F统计量:

- **F=平均的回归平方和/平均的剩余平方和~F(k,n-k-1)**
- **如果F值较大，则说明自变量造成的因变量的线性变动大于随机因素对因变量的影响,自变量于因变量之间的线性关系较显著**

(4)计算F统计量的值和相伴概率p

(5)判断：p<=a:拒绝H_0,即:所有回归系数与0有显著差异，自变量与因变量之间存在显著的线性关系。反之，不能拒绝H_0

多元线性回归方程的检验

(三)回归系数的显著性检验

(1)目的:检验每个自变量对因变量的线性影响是否显著.

(2)H_0:$\beta_i=0$ 即:第i个回归系数与0无显著差异

(3)利用t检验,构造t统计量:

$$t_i = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2}} \qquad \hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(4)逐个计算t统计量的值和相伴概率p

(5)判断

多元线性回归方程的检验

(四)t统计量与F统计量

- **一元回归中,F检验与t检验一致,即: $F=t^2$,可以相互替代**
- **在多元回归中，F检验与t检验不能相互替代**
- **$F_{change}=t_i^2$** $F_{change} = \frac{R_{ch}^2(n-k-1)}{1-R^2}$ $R_{ch}^2 = R^2 - R_i^2$
- **从F_{change}角度上讲，如果由于某个自变量x_i的引入，使得F_{change}是显著的(通过观察F_{change}的相伴概率值)，那么就可以认为该自变量对方程的贡献是显著的，它应保留在回归方程中，起到与回归系数t检验同等的作用。**

自变量筛选

(一)自变量筛选的目的

- 多元回归分析引入多个自变量. 如果引入的自变量个数较少,则不能很好的说明因变量的变化;
- 并非自变量引入越多越好.原因:
  - 有些自变量可能对因变量的解释没有贡献
  - 自变量间可能存在较强的线性关系,即:多重共线性. 因而不能全部引入回归方程.

自变量筛选

(二)自变量向前筛选法(forward):

- 即:自变量不断进入回归方程的过程.
- 首先,选择与因变量具有最高相关系数的自变量进入方程,并进行各种检验;
- 其次,在剩余的自变量中寻找偏相关系数最高的变量进入回归方程,并进行检验;
  - 默认:回归系数检验的概率值小于PIN(0.05)才可以进入方程.
- 反复上述步骤,直到没有可进入方程的自变量为止.

自变量筛选

(三)自变量向后筛选法(backward):

- 即:自变量不断剔除出回归方程的过程.
- 首先,将所有自变量全部引入回归方程;
- 其次,在一个或多个t值不显著的自变量中将t值最小的那个变量剔除出去,并重新拟和方程和进行检验;
  - 默认:回归系数检验值大于POUT(0.10),则剔除出方程
- 如果新方程中所有变量的回归系数t值都是显著的,则变量筛选过程结束.
- 否则,重复上述过程,直到无变量可剔除为止.

自变量筛选

(四)自变量逐步筛选法(stepwise):

- 即:是“向前法”和“向后法”的结合。
- 向前法只对进入方程的变量的回归系数进行显著性检验，而对已经进入方程的其他变量的回归系数不再进行显著性检验，即：变量一旦进入方程就不回被剔除
- 随着变量的逐个引进，由于变量之间存在着一定程度的相关性，使得已经进入方程的变量其回归系数不再显著，因此会造成最后的回归方程可能包含不显著的变量。
- 逐步筛选法则在变量的每一个阶段都考虑的剔除一个变量的可能性。

线性回归分析中的共线性检测

(一)共线性带来的主要问题

- 高度的多重共线会使回归系数的标准差随自变量相关性的增大而不断增大,以至使回归系数的置信区间不断增大,造成估计值精度减低.

(二)共线性诊断

- 自变量的容忍度(tolerance)和方差膨胀因子
  - 容忍度:$Tol_i=1-R_i^2$. 其中: R_i^2是自变量x_i与方程中其他自变量间的复相关系数的平方.
  - 容忍度越大则与方程中其他自变量的共线性越低,应进入方程. (具有太小容忍度的变量不应进入方程,spss会给出警)(T<0.1一般认为具有多重共线性)

线性回归分析中的共线性检测

- 方差膨胀因子(VIF):容忍度的倒数

$$VIF_i=\frac{1}{1-R_i^2}$$

方差膨胀因子的取值大于等于1。

- VIF_i越大多重共线性越强，当VIF大于等于10时，说明解释变量 x_i与方程中其余解释变量之间有严重的多重共线性，且可能会过度地影响方程的最小二乘估计。
- SPSS在回归方程建立过程中不断计算待进入方程自变量的容忍度,并显示目前的最小容忍度

线性回归分析中的共线性检测

(二)共线性诊断

* 用特征根刻画自变量的方差
  * 若自变量间确实存在较强的相关关系，那么它们之间必然存在信息重叠，于是可从这些自变量中提取出既能反映自变量信息(方差)又相互独立的因素(成分)来.
  * 从自变量的相关系数矩阵出发，计算相关系数矩阵的特征根，得到相应的若干成分.
  * 若某个特征根既能够刻画某个自变量方差的较大部分比例(如大于0.7),同时又可以刻画另一个自变量方差的较大部分比例，则表明这两个自变量间存在较强的多重共线性。
* 条件指标
  * 0<k<10 无多重共线性; 10<=k<=100 较强; k>=100 严重

$$k_i = \sqrt{\frac{\lambda_m}{\lambda_i}}$$

线性回归分析中的异方差问题

(一)什么是差异方差

* 回归模型要求残差序列服从均值为0并具有相同方差的正态分布,即:残差分布幅度不应随自变量或因变量的变化而变化.否则认为出现了异方差现象

(二)差异方差诊断

* 可以通过绘制标准化残差序列和因变量预测值(或每个自变量)的散点图来识别是否存在异方差

(三)异方差处理

* 实施方差稳定性变换
  * 残差与y_i(预测值)的平方根呈正比：对y_i开平方
  * 残差与y_i(预测值)呈正比:对y_i取对数.
  * 残差与y_i(预测值)的平方呈正比,则$1/y_i$

曲线估计

(一)目的:

* 在一元回归分析或时间序列中,因变量与自变量(时间)之间的关系不呈线性关系,但通过适当处理,可以转化为线性模型.可进行曲线估计.

(二)曲线估计的常用模型:

* $y=b_0+b_1t$ (线性拟和linear)
* $y=b_0+b_1t+b_2t^2$ (二次曲线quadratic)
* $y=b_0+b_1t+b_2t^2+b_3t^3$ (三次曲线cubic)

t为时间,也可为某一自变量.

曲线估计

(三)基本操作步骤

(1)绘制散点图,观察并确定模型.

(2)菜单选项: 分析->回归->曲线估计

(3)选择因变量

(4)选择自变量或选time以时间作自变量

(5)选择模型 (R^2最高拟和效果最好)

线性回归分析的应用举例

一般线性回归分析举例

案例1：为研究腰围、体重和脂肪比重之间的关系，随机调查了20个人。现利用一般线性回归分析方法进行研究。这里，被解释变量为腰围，解释变量为体重和脂肪比重。

具体数据在可供下载的压缩包中，文件名为“腰围和体重. sav"。

SPSS绘制散点图的基本操作

选择菜单：

【分析(A)】-【回归(R)】-【线性(L)】

选择被解释变量到【因变量(D) 】框中。

选择一个或多个变量到【自变量(I)】框中

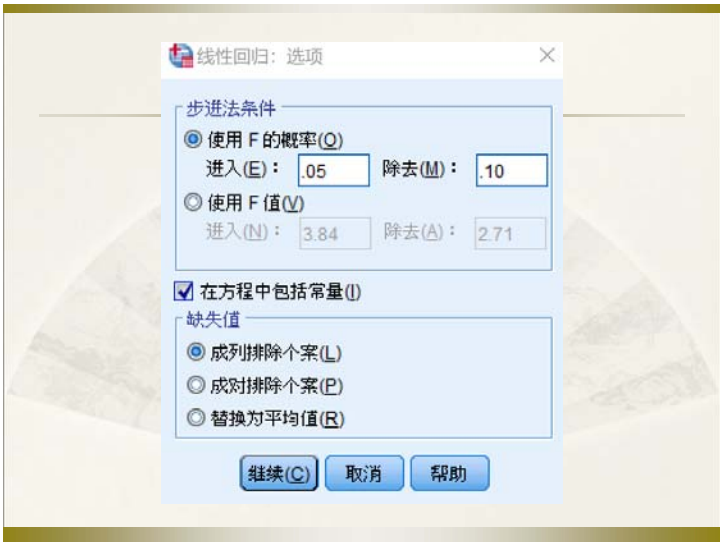


表 9—1（a） 关于腰围的线性回归分析结果（一）

模型汇总[c]

模型	R	R 方	调整 R 方	标准 估计的误差
1	0.887[a]	0.787	0.775	1.812
2	0.945[b]	0.894	0.881	1.315

a. 预测变量：（常量），%脂肪比重。
b. 预测变量：（常量），%脂肪比重，体重（磅）。
c. 因变量：腰围（英寸）。

从拟合优度角度看，第二个模型的拟合效果更佳。

表 9—1（b）　关于腰围的线性回归分析结果（二）

Anova[c]

模型		平方和	df	均方	F	Sig.
1	回归	217.829	1	217.829	66.320	0.000[a]
	残差	59.121	18	3.284		
	总计	276.950	19			
2	回归	247.541	2	123.770	71.545	0.000[b]
	残差	29.409	17	1.730		
	总计	276.950	19			

a. 预测变量：（常量），%脂肪比重。
b. 预测变量：（常量），%脂肪比重，体重（磅）。
c. 因变量：腰围（英寸）。

解释变量全体与被解释变量间存在显著的线性关系，选择线性模型具有合理性。

表 9—1（c）　关于腰围的线性回归分析结果（三）

系数[a]

模型		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准 误差	试用版			容差	VIF
1	（常量）	30.058	0.949		31.657	0.000		
	%脂肪比重	0.354	0.043	0.887	8.144	0.000	1.000	1.000
2	（常量）	20.236	2.468		8.199	0.000		
	%脂肪比重	0.227	0.044	0.569	5.163	0.000	0.515	1.943
	体重（磅）	0.065	0.016	0.457	4.144	0.001	0.515	1.943

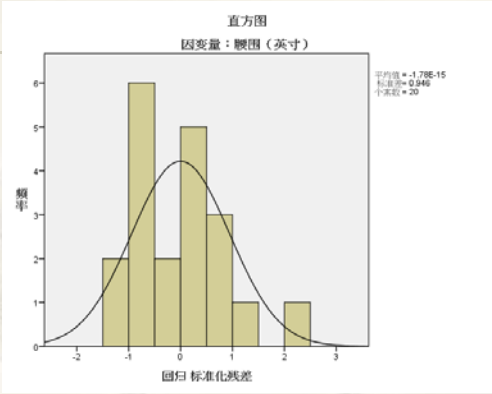
a. 因变量：腰围（英寸）。

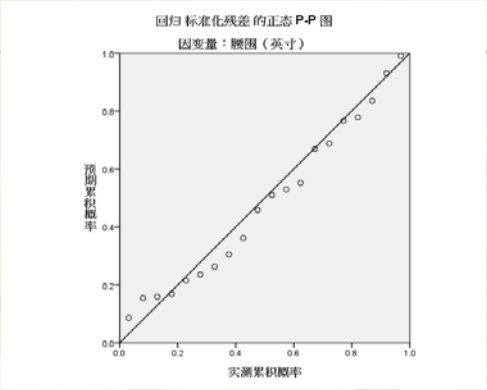
表 9—1（d）　关于腰围的线性回归分析结果（四）

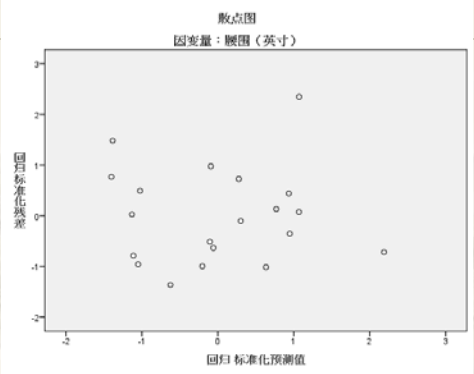
已排除的变量[b]

模型		Beta In	t	Sig.	偏相关	共线性统计量		
						容差	VIF	最小容差
1	体重（磅）	0.457[a]	4.144	0.001	0.709	0.515	1.943	0.515

a. 模型中的预测变量：（常量），%脂肪比重。
b. 因变量：腰围（英寸）。







最终的回归方程为：

腰围＝20.2＋0.23×脂肪比重＋0.07×体重

方程表明，当体重保持不变时，脂肪比重提高一个百分点，腰围平均增加 0.23 英寸。

当脂肪比重保持不变时，体重增加 1 磅，腰围平均增加 0.07 英寸。

带虚拟解释变量的回归分析举例

- 为研究工龄和性别对月工资收入产生的影响，随机调查了30名职工得到月基本工资，工龄(月)和性别数据。数据所在文件名”工资收入的影响因素.sav”
- 本例中的性别有两个分类值， 可生成为两个虚拟变量， 分别表示 “是男性吗”“是女性吗＂ 。 如果是男性， 则第一个虚拟变量取值为1, 第二个虚拟变量取值为0。 反之． 类似。

由于多个虚拟变批间存在完全的线性关系（之和为 1), 如果都参与回归建模，会导致完全共线性问题。

所以， 若分类型解释变量有k个类别值， 仅需引入前k- 1个虚拟变量到回归模型中。

例如， 本例仅引入 “是男性吗”这一个虚拟变量即可。

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	43730535.43	2	21865267.72	43.796	.000[b]
	残差	13479734.57	27	499249.428		
	总计	57210270.00	29			

a. 因变量：月基本工资

b. 预测变量：(常量), 工龄(月), 是男性吗？

分析结果表明， 本例采用线性回归模型是合理的。

系数[a]

模型		未标准化系数		标准化系数	t	显著性
		B	标准误差	Beta		
1	(常量)	2403.834	416.348		5.774	.000
	是男性吗？	1377.873	280.337	.489	4.915	.000
	工龄(月)	42.659	4.653	.912	9.167	.000

a. 因变量：月基本工资

两个解释变量的回归系数的显著性检验均显著， 它们均应保留在回归方程中。

最终的回归方程为：

月基本工资=2403. 8+1377. 9*是男性吗+42.7*工龄

##表示性别相同的条件下，工龄增加1个月， 月基本工资将平均增加42. 7元。

##为解释1377. 9的含义， 可参考以下两个回归方程：

当“是男性吗” 取0时有：月基本工资=2403. 8+42.7*工龄

该方程为女性的工资方程。

当”是男性吗” 取1时有：

月基本工资=2403. 8+1377.9+42.7*工龄

该方程为男性的工资方程。

可见， 上述两个方程所代表的回归直线是两条平行线，1377. 9是两直线的截距之差，反映了当工龄相同时男性与女性月基本工资的平均差异。

本例中， 相同工龄下男性的月基本工资的平均值比女性高1377. 9元。

本例默认男性的回归线与女性的平行， 即不同性别的工龄的月工资回报是相同的。

当无法确认这种默认是否合理时， 可建立回归方程

$$月基本工资=\beta_0+\beta_1 工龄+\beta_2 是男性吗+\beta_3 工龄\times 是男性吗$$

方程中的第四项称为工龄和性别的交互项。

当“是男性吗” 取0时有：

$$月基本工资=\beta_0+\beta_1\times 工龄。$$

当“是男性吗” 取1时有：

$$月基本工资=(\beta_0+\beta_2)+(\beta_1+\beta_3)\times 工龄。$$

β_3 是两直线斜率之差， 反映不同性别工龄的月工资回报的差异。

系数[a]

模型		未标准化系数		标准化系数		
		B	标准误差	Beta	t	显著性
1	(常量)	2446.429	625.113		3.914	.001
	是男性吗？	1314.967	735.791	.466	1.787	.086
	工龄(月)	42.113	7.558	.900	5.572	.000
	工龄和性别的交互项	.900	9.705	.023	.093	.927

a. 因变量：月基本工资

回归方程中交互项的回归系数检验不显著， 所以本例默认男性和女性的回归直线平行具有合理性。