阮敬 博士

首都经济贸易大学研究生院 副院长
首都经济贸易大学统计学院 教授

© ruanjing@msn.com

回归分析

- 在上一章的相关分析中，参与分析的变量通常是两个或者两组，而现实生活中变量之间的关系往往不仅限于这种相互影响，多个变量可能都会对所研究的因变量产生影响。

- 而本章将要介绍的回归分析不仅可以分析两个变量的影响，还可以分析多个变量之间的统计关系。回归分析是统计分析中最为重要的方法之一，本章将在结合实际问题的基础上，对一些最为常用的回归分析方法进行介绍。

线性回归分析

- 当变量之间存在统计关系时，还可以进行回归分析（Regression）。

- “回归”一词来源于高尔顿研究人类身高遗传问题的过程。1870 年，高尔顿在研究人类身高的遗传问题时，发现高个子父母的子女，其身高有低于其父母身高的趋势，而矮个子父母的子女，其身高有高于其父母的趋势，即有“退回”（即 Regression 的原意）到身高均值的趋势。高尔顿首次引入了回归直线等概念，始创了回归分析。回归分析的研究领域非常多，有线性回归、非线性回归、定性自变量回归、离散因变量回归等。在社会经济领域的实际应用中，线性回归分析应用非常广泛。

回归分析的基本原理

- 回归分析与相关分析在理论和方法上具有一致性， 变量之间没有关系， 就谈不上回归分析或建立回归方程；相关程度越高，回归效果就越好，而且相关系数和回归系数方向一致，可以互相推算。
- 相关分析中的两个变量之间的地位是对等的，即变量 A 与变量 B 相关等价于变量 B 与变量 A 相关，相关分析的两个变量均为随机变量；而回归分析中要确定自变量和因变量，通常情况下只有因变量是随机变量，人们可以利用回归分析来对研究对象进行预测或控制。
- 回归分析往往是通过一条拟合的曲线来表示模型的建立。以线性回归为例，设 y 表示因变量，x 表示自变量，则有如下线性回归模型：

$$y = \alpha + \beta x + \epsilon$$

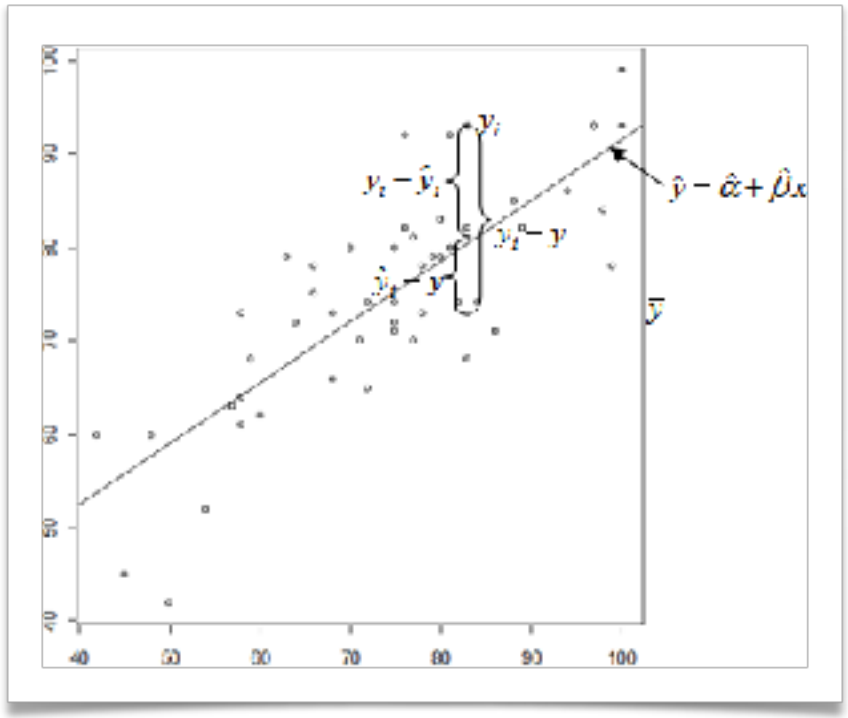
- 其中：
  - α 和 β 是回归模型的参数，称为回归系数（α 也可称为截距项）；
  - ε 是随机误差项或随机扰动项， 反映了除 x 和 y 之间线性关系之外的随机因素或不可观测的因素。
- 通常在回归分析中，对 ε 有如下最为常用的基本经典假定：
  - ε 的期望值为 0；
  - ε 对于所有 x 而言具有同方差性；
  - ε 是服从正态分布且相互独立的随机变量。

回归分析的基本原理

- 如果存在多个自变量，回归模型可以写作：
- $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon$
- 因为上述直线模型中含有随机误差项， 所以回归模型反映的直线是不确定的。 回归分析的主要目的就是要从这些不确定的直线中， 找出一条最能够代表数据原始信息的直线， 来描述因变量和自变量之间的关系，这条直线称之为回归方程，如只有一个自变量的情况下，可对模型左右两边取 x 的条件期望并根据 ε 的经典假定，可得：
- $E(y|x) = \alpha + \beta x + 0$
- 然后通过一定的参数估计方法，可得到估计的直线方程如下：
- $\hat{y} = \hat{\alpha} + \hat{\beta} x$

回归分析的基本原理

- 因为回归模型的参数往往是未知的，人们只能靠样本数据去进行参数估计，所以可用$\hat{\alpha}$和$\hat{\beta}$分别表示回归模型中 α 和 β 的参数估计值。方程$\hat{y} = \hat{\alpha} + \hat{\beta}x$也可称之为
- 估计的回归方程，如图12-1所示。
- $\hat{\alpha}$是估计的回归直线在 y 轴上的截距，$\hat{\beta}$是直线的斜率，它表示自变量 x 每变动一个单位时，因变量 y 的平均变动值，$\hat{y}$是 y 的估计值。



回归分析的基本原理

- **1．回归方程的普通最小二乘参数估计**
- 可把具体某个因变量的观测值记为y ，其期望值为 y ，把其估计值（即在回归直线上的值）记为$\hat{y}_i$。那么如何在存在随机因素的回归模型中找出最能代表原始数据信息的回归直线呢？可以通过对因变量的离差入手进行分析。
- 如图所示，因变量的离差为$y_i - \bar{y}$，可以把离差分解为两个部分：即$y_i - \hat{y}_i$（称之为残差）和$y_i - \hat{y}_i$（称之为回归离差）。每个因变量的离差等于残差与回归离差之和，即$y_i - \bar{y} = y_i - \hat{y}_i + y_i - \hat{y}_i$。对于所有因变量的观测值而言，为了避免正负符号的影响，可以对上述分解出来的 3 个差值求平方得：

$$\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2$$

- 即，总离差平方和（记为 SST）等于残差平方和（SSE）与回归离差平方和（SSR）之和， 上述等式可以进行严格数学证明， 证明过程本书不予赘述， 请查阅相关统计学原理书籍。

回归分析的基本原理

- 从图 12-1 可以看出， 如果残差越小，y_i就越往回归直线靠近。 对于所有的因变量而言，残差平方和越小，观测值就越往回归直线靠近。当残差平方和（SSE）达到极小值时，即$\sum(y_i-\hat{y}_i)^2$→最小值时，估计出来的回归直线能在最大程度上代表原始数据的信息。
- 当$\sum(y_i-\hat{y}_i)^2=\sum[y_i-(\hat{\alpha}+\hat{\beta}x)]^2\rightarrow min$时，可以利用微分求极值的方法，分别对$\sum[y_i-(\hat{\alpha}+\hat{\beta}x)]^2$求$\hat{\alpha}$和$\hat{\beta}$的偏微分， 并使之同时为 0， 然后求解联立方程组便可计算$\hat{\alpha}$和$\hat{\beta}$的具体数值，作为回归模型中 α 和 β 的参数估计值如下：

$$\begin{cases}\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}\\ \hat{\beta}=\dfrac{n\sum y_i x_i-\sum y_i\sum x_i}{n\sum x_i^2-\left(\sum x_i\right)^2}\end{cases}$$

回归分析的基本原理

- 这种参数估计的方法和过程是在随机误差项 ε 是一个期望值为 0，且对于所有 x 而言具有同方差性，服从正态分布且相互独立的假定下进行的，通常被称之为“普通最小二乘法”（Ordinary Least Squares）。由上述通过条件期望形式使用普通最小二乘估计出来的参数估计值，代表了自变量变动对因变量期望变动的影响。
- 普通最小二乘法同样适用于多个自变量和因变量之间模型的参数估计，在 SAS 系统的回归分析功能中，系统默认使用普通最小二乘法对线性回归模型进行参数估计。
- 回归分析的主要内容之一便是利用上述方法对模型的参数进行估计， 对模型进行参数估计之后， 还应当对模型的拟合程度和回归系数的显著性进行检验， 经过检验和评估的模型便可用来对因变量进行预测。

回归分析的基本原理

- **2．回归方程的检验**
- 对于估计出来的回归方程， 可以从模型的解释程度、 回归方程总体显著性以及回归系数的显著性等几个方面进行检验。
- **（1）回归方程的拟合优度判定**
- 回归方程的拟合优度主要用于判定$\hat{y}_i$估计 y 的可靠性问题， 可用来衡量模型的解释程度。拟合优度判定是建立在模型参数估计时对总离差平方和（SST）分解的基础上的，SST 可以分解为残差平方和（SSE）和回归离差平方和（SSR），通常使用回归离差平方和（SSR）占总离差平方和（SST）的比重来判断模型的解释能力，即：

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

回归分析的基本原理

- 其中R^2表示拟合优度的判定系数或决定系数，其取值范围为[0，1]。R^2越接近于 1，说明变量之间的相互依存关系越密切， 其相互依存关系就越接近于函数关系， 两变量之间的相关程度越高，回归方程的拟合程度越好。所以，R^2越接近 1，模型的解释程度越好，模型越精确。
- 但是，R^2的数值与自变量的数目有关，即自变量的个数越多越大，这在一定程度上削弱了R^2的评价能力，因此可考虑剔除自变量数目影响的修正R^2。

回归分析的基本原理

- （2）回归方程总体显著性检验
- 利用普通最小二乘法拟合出来的回归方程， 都是由样本数据进行的， 那么用它来对总体进行推断是否显著呢？可以对回归方程总体的显著性进行检验。
- 回归方程总体显著性检验主要是检验因变量和自变量之间的线性关系是否显著。 其原假设和备择假设如下：

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_i = 0 \qquad H_1: \beta_i \text{不全为} 0$$

- 对于回归方程总体显著性检验可用 F 检验并计算出用于检验判定的 P 值来进行。 如果 P值小于理论显著性水平 α 值，可认为在显著性水平 α 条件下，回归方程总体显著。

回归分析的基本原理

- （3）回归方程系数显著性检验
- 如果模型的线性关系显著， 还应对模型参数估计的结果即回归方程的系数进行显著性检验，用于考察单个自变量与因变量的线性关系是否成立。其原假设和备择假设如下：

$$H_0: \beta_i = 0 (i = 1,2,\dots,k) \qquad H_1: \beta_i \neq 0 (i = 1,2,\dots,k)$$

- 回归方程系数显著性检验要求对所有估计出来的回归系数分别进行检验 （截距项通常不进行显著性检验） ，可以利用 t 检验进行。
- 在 SAS 系统中，可以计算出每个回归系数所对应的 P 值。如果某个系数对应的 P 值小于理论显著性水平 α 值，可认为在显著性水平 α 条件下，该回归系数是显著的。
- 有些情况下， 没有任何关联的变量之间进行回归分析也可能得到显著的检验结果， 会对分析过程造成不良的影响。因此，在进行回归分析之前，必须考虑好变量之间的关系及其所代表的经济含义。

回归分析的基本原理

- 3．回归方程的预测
- 回归预测是一种有条件的预测，依据估计出来的回归方程，在给定自变量数值的条件下，对因变量进行预测，其预测的基本公式为：

$$\hat{y}_f = \hat{\alpha} + \hat{\beta} x_f$$

- 其中x_f是另外给定的自变量的值，$\hat{y}_f$为根据回归方程计算出来的预测值。

一元线性回归分析

- 一元线性回归是回归分析中最简单的一种形式， 主要考察单独 1 个自变量对因变量的影响。其模型形如：

$$y = \alpha + \beta x + \varepsilon$$

- 一元线性回归分析的基本步骤如下：
- 依据变量之间的关系，判断其是否是线性关系。如果是线性关系，可以利用第 12.1.1小节 中介绍的方法进行回归模型的参数估计，然后根据参数估计的结果进行检验。
- 在检验过程中， 可以先对模型的解释能力进行拟合优度判定， 拟合优度的判定系数如果非常小，说明建立的回归方程解释能力较差， 在进行回归分析的过程中可能还有其他重要因素没有加入到模型当中， 可以考虑增加有重要影响的自变量； 回归方程总体显著性如果不显著，说明变量之间的线性关系不明显，不适合做线性回归；在拟合优度判定系数比较高、方程总体显著的情况下， 对回归系数进行检验， 通过显著性检验的回归系数才对因变量有解释能力。
- 只有通过检验的模型才能够充分描述变量之间的关系，建立的模型才有现实意义。

一元线性回归分析

- 例12-1：通常情况下， 一个国家或地区的犯罪率在很大程度上受到国民素质的影响， 而反映国民素质的一个重要指标便是文盲率（或识字率） 。在正常逻辑思维当中，文盲率越低，其普法程度就越低，可能会对社会造成一定的危害。为了研究文盲率与犯罪率之间的关系，现在全世界范围内收集到来自于不同区域的 50 个地区的文盲率与谋杀犯罪率的数据如图 12-2 所示。试对文盲率与谋杀犯罪率进行回归分析。
- 本例所使用的数据值标签如下：

```
proc format;
  value region_fmt  1='East North Central‘
                    2='East South Central'
                    3='Middle Atlantic'
                    4='Mountain'
                    5='New England'
                    6='Pacific'
                    7='South Atlantic'
                    8='West North Central'
                    9='West South Central';
run;
```

Dataset of SASUSER.MURDER

Obs	Region 区域	Illiteracy 文盲率%	Murder 谋杀犯罪率%
1	4	0.5	11.5
2	8	0.5	2.3
3	8	0.5	1.7
4	4	0.6	5.3
		……	
47	9	2.2	12.2
48	7	2.3	11.6
49	2	2.4	12.5
50	9	2.8	13.2

一元线性回归分析

- 本例要研究文盲率对谋杀犯罪率的影响，因变量为谋杀犯罪率，自变量为文盲率，二者之间的实际意义明显。 在进行回归分析之前， 应当对两个变量之间的是否是线性关系进行研究，根据本例变量绘制散点图（图略） ，可发现二者之间在一定程度上呈线性关系。
- SAS 系统中提供了近 10 种回归分析的方法，本节所述的回归分析可以利用 REG 过程来实现，其主要语法如下：

```
PROC REG < 选项> ;
  <模型标签: > MODEL 因变量列表=<自变量> < /选项> ;
  BY 变量列表;
  FREQ 变量;
  ID 变量列表;
  VAR 变量列表;
  WEIGHT 变量;
  ADD 变量列表;
  DELETE 变量列表;
  <标签: > MTEST <方程, ... ,方程> < /选项> ;
  OUTPUT <OUT=输出数据集 > 统计量关键字=变量名列表 < ... 统计量关键字=变量名列表> ;
  PAINT <条件或 ALLOBS> < /选项> 或 < STATUS 或 UNDO> ;
  PLOT <y 轴变量*x 轴变量> <=绘图标记> < ...y 轴变量*x 轴变量> <=绘图标记> < /选项> ;
  PRINT <选项> < ANOVA > < MODELDATA > ;
  REFIT;
  RESTRICT 方程, ... , 方程;
  REWEIGHT <条件或 ALLOBS> < /选项> 或< STATUS 或 UNDO> ;
  <标签: > TEST 方程,<, ..., 方程> < /选项> ;
```

一元线性回归分析

- REG 过程中，BY、FREQ、VAR、WEIGHT、ID 语句功能与在前面介绍过的语句中的作用一致，其他常用语句的功能如下：
  - MODEL：指定回归模型，等号左边为因变量列表，等号右边为自变量列表。可以使用“标签： ”的形式指定模型标签；
  - ADD：在回归模型中增加自变量；
  - DELETE：从回归模型中剔除自变量；
  - MTEST：在多元因变量模型中进行多元检验；
  - OUTPUT：输出一个新的数据集，该数据集中包括预测值、残差以及其他用于诊断的统计量；
  - PAINT：在散点图中描点；
  - PLOT：绘制散点图；
  - PRINT：显示能够进行模型选项调整的信息；
  - REFIT：重新拟合模型；
  - RESTRICT：在进行模型参数估计时，设定线性约束；
  - REWEIGHT： 从分析中剔除特定观测值或改变所使用观测值的权重；
  - TEST：对参数估计的线性方程进行 F 检验。
- 上述常用语句中，MODEL 语句是 REG 过程必不可少的语句，同时也是指定回归模型的最重要语句，不仅可指定模型中的因变量、自变量，还可指定模型选项及输出结果选项。

一元线性回归分析

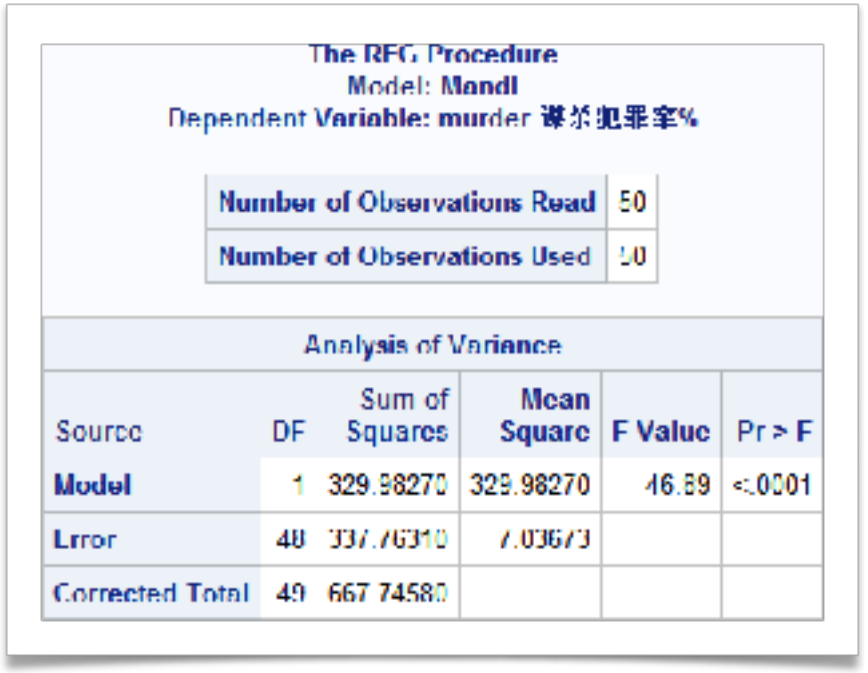
- 注意：MODEL 语句中的变量必须是所用数据集中的现有的变量，而不能是现有变量的变换变量，即如果要在模型中加入 x 变量的对数形式，必须要先在数据集中生成一个新的变量来代表 x 的对数，如生成 Log_x 变量来代表 x 变量的对数，则模型中应当引入变量 Log_x 而不是 Log(x) 函数。
- MODEL 语句常用的选项主要有：
  - NOINT：不估计模型的截距项；
  - STB：估计模型的标准化参数估计结果；
  - CLI：给出因变量预测值的 1-α 置信区间上下限（α 具体数值可在 REG 的过程选项中加入关键字 ALPHA=来指定）；
  - CLM：给出因变量期望值的置信区间上下限；
  - R：输出每个样本因变量预测值、残差及标准误差；
  - P：输出因变量观测值、预测值及残差。
- 此外，在进行多元回归分析时，MODEL语句的选项还可对变量筛选的方法进行设置（见第 12.1.3 小节）。

一元线性回归分析

- 本例利用 REG 过程进行回归分析的具体程序如下：

```
proc reg data=sasuser.murder;
 MandI: model murder=illiteracy;
      /*指定模型标签为 MandI，因变量 MURDER，自变量 ILLITERACY*/
run;
quit;
```

- 程序运行后，首先可以得到如图 12-3 所示的 F 检验结果。
- 在图 12-3 中可以看到用于回归方程总体显著性检验的 F 统计量的值 （F Value） ， 其对应的 P 值（Pr>F）<0.0001，非常显著，因此回归方程总体上是显著的。



The REG Procedure
Model: MandI
Dependent Variable: murder 谋杀犯罪率%

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	329.98270	329.98270	46.69	<.0001
Error	48	337.76310	7.03673		
Corrected Total	49	667.74580			

一元线性回归分析

- 图 12-4 所示的结果中能够进行拟合优度的判定。
- 拟合优度的判定系数 R^2（R-Square）和修正的R^2（Adj R-Sq）分别是：R^2=0.4942，修正R^2=0.4836， 拟合程度不高， 但是在本例所示的截面数据中， 该拟合程度还是可以接受的。
- 在图 12-5 所示的参数估计表中，可以看到使用普通最小二乘法（OLS）对回归模型进行参数估计的结果及对应回归系数的显著性检验过程。
- 在 Parameter Estimates 表中截距项（Intercept）的参数估计值为 2.39678，其对应显著性检验 P 值=0.0052 （通常不对截距项进行显著性检验）；自变量 Illiteracy 对应的回归系数估计值为 4.25746，其对应显著性检验的 P 值<0.0001，在给定的显著性水平条件下非常显著。

Root MSE	2.65268	R-Square	0.4942
Dependent Mean	7.37800	Adj R-Sq	0.4836
Coeff Var	35.95397		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.39678	0.81844	2.93	0.0052
illiteracy	文盲率%	1	4.25746	0.62171	6.85	<.0001

一元线性回归分析

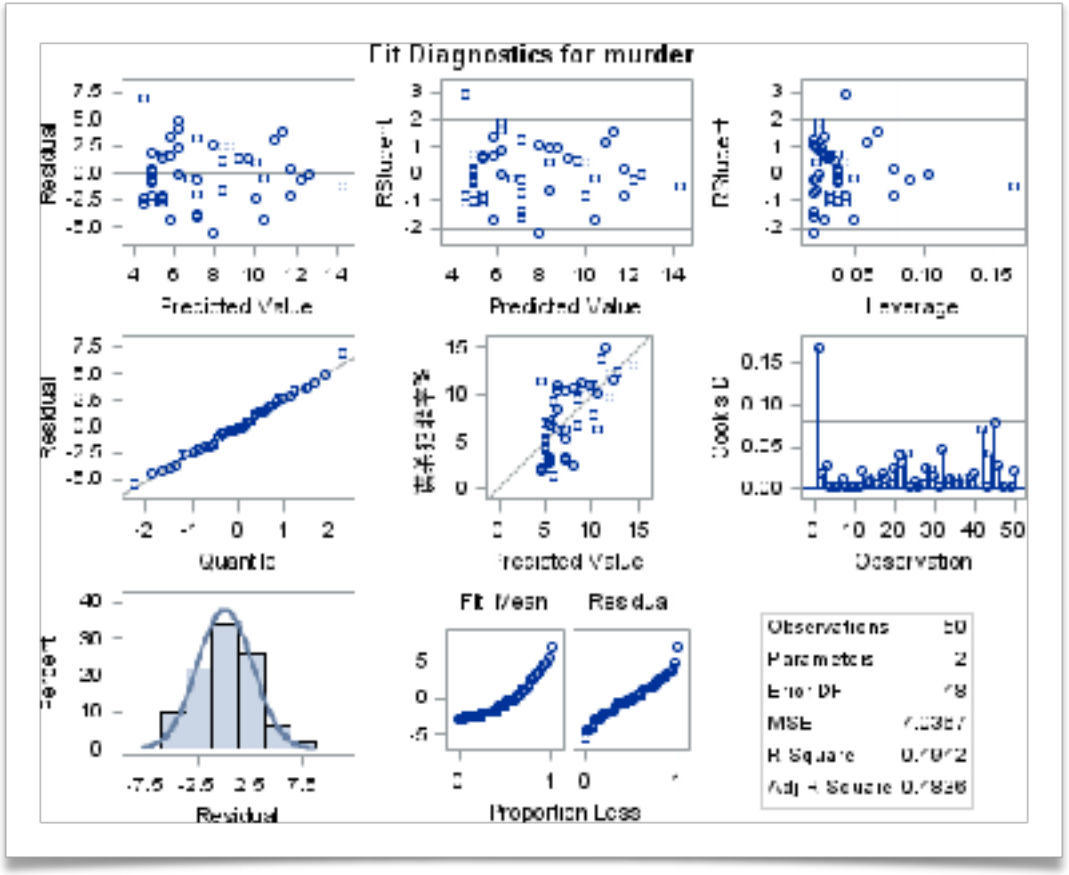
- 综上所述，本例所建立的回归模型拟合程度尚好，模型总体上是显著性的，回归系数也是显著的。因此，可以依据图 12-5 所示参数估计表的数值写出回归方程式，并根据此方程式对自变量与因变量之间的关系进行分析：

$$Murder = 2.39678 + 4.25746 \times Illiteracy$$

- 从上述通过拟合优度判定和显著性检验的方程中可知，当文盲率每增加/降低 1 个单位时，谋杀犯罪率会平均增加/降低 4.25746 个单位。具体而言，文盲率每增加/降低 1 个百分点时，谋杀犯罪率平均增加/降低 4.25746 个百分点。
- 此外， 根据回归方程估计出来的残差项 ε 也应当符合经典假定 （见第 12.1.1 小节） 。 因此，在进行回归分析的过程当中， 还应当对残差项是否符合假定进行判定， 只有在符合假定的前提条件下，上述用 OLS 方法估计出来的回归方程才有解释能力。在一元线性回归模型中，通常可用残差图来判定残差是否与变量相关，用 P-P 图或 Q-Q 图来判定残差项是否符合正态分布。

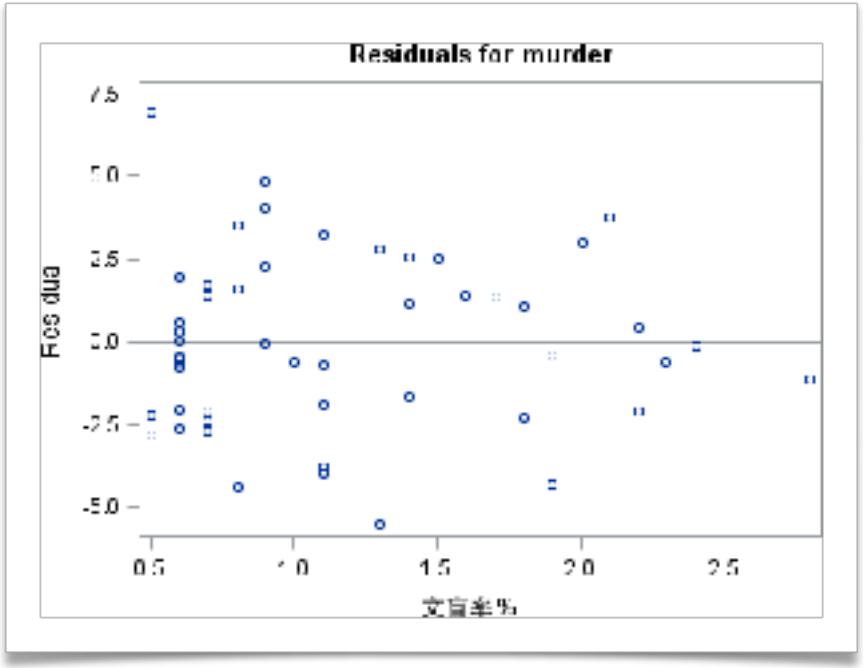
一元线性回归分析

- 在 SAS9.3 的 REG 过程中的 MODEL 语句会自定输出上述有关模型诊断的图形，本例模型拟合的诊断图形如图 12-6 所示。
- 从图 12-6 的第一行图形可以看出，因变量的预测值与残差项没有什么关系，对本例数据采用线性回归进行建模没有问题。 从第二行和第三行的第一个图形均可以大致看出， 模型残差项服从正态分布。



一元线性回归分析

- 残差项与自变量之间的关系可以用图 12-7 来描述。
- 从图 12-7 可明显看出，自变量与残差之间的关系不明显，基本上无关，符合 ε 对于所有 x 而言具有同方差性的假定； 而残差大体均匀的分布在[-6， 6]之间， 其均值与 0 非常接近，故符合 ε 零均值的假定。



一元线性回归分析

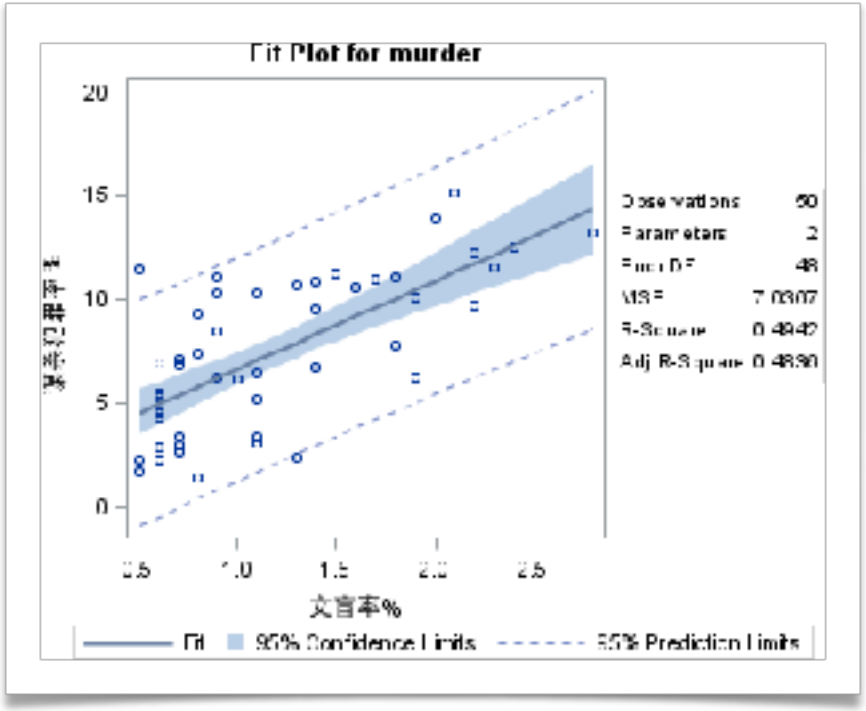
- REG 过程运行结果还会给出如图 12-8 所示的拟合曲线和因变量预测置信区间。
- 此外，一些用于诊断的图形在 REG 过程中也可以使用 PLOT 语句来绘制，如：

  plot r.*illiteracy; /*绘制残差与自变量的散点图，关键字 R.表示残差*/

  plot npp.*r.; /*绘制 P-P 图，NPP.表示正态累积分布*/

  plot nqq.*r.; /*绘制 Q-Q 图*/

- 请读者自行在 SAS 系统中查看本段程序的输出结果， 这些图形结果有助于判定模型残差项是否符合经典假定。



多元线性回归分析

- 对因变量产生影响的自变量可能不止单独一个， 有可能有多个。 如一个人的体重可能会受到其身高、血型、生活习惯、收入水平等变量的影响。对于多个变量对因变量的影响，可以考虑利用多元线性回归分析的方法进行分析。多元线性回归模型如下：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

- ε仍然服从零均值、 相互独立且同方差服从正态分布等经典假定。 同一元线性回归一样，可以对上述模型左右两边同时取条件期望，可得：

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + 0$$

- 仍然利用普通最小二乘法，可估计出参数$\beta_0, \beta_1, \beta_2, \ldots, \beta_n$的估计值$\hat{\beta}_0, \hat{\beta}_1$， $\hat{\beta}_2, \cdots$， $\hat{\beta}_n$，即可得到多元回归方程：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_n x_n$$

- 对于多元回归方程的拟合优度判定、 方程总体显著性检验与回归系数显著性检验过程与一元回归方程的检验过程一样。

多元线性回归分析

- 例12-2：公司管理层往往会根据公司员工的起薪、年龄大小、工作经验、职位以及学历等诸多因素来决定其当前薪酬。 为了考察某集团公司员工当前薪酬水平的影响因素， 现收集了471 名公司雇员的背景信息，具体信息如图 12-9 所示。试对该公司员工的当前薪酬及其影响因素进行回归分析（设显著性水平 α=0.1） 。
- 本例使用的数据值标签如下：

```
proc format;
  value gender_fmt  0='女性'
                    1='男性';
  value position_fmt  1='经理'
                      2='主管'
                      3='普通员工';
run;
```

Dataset of SASUSER.SALARY

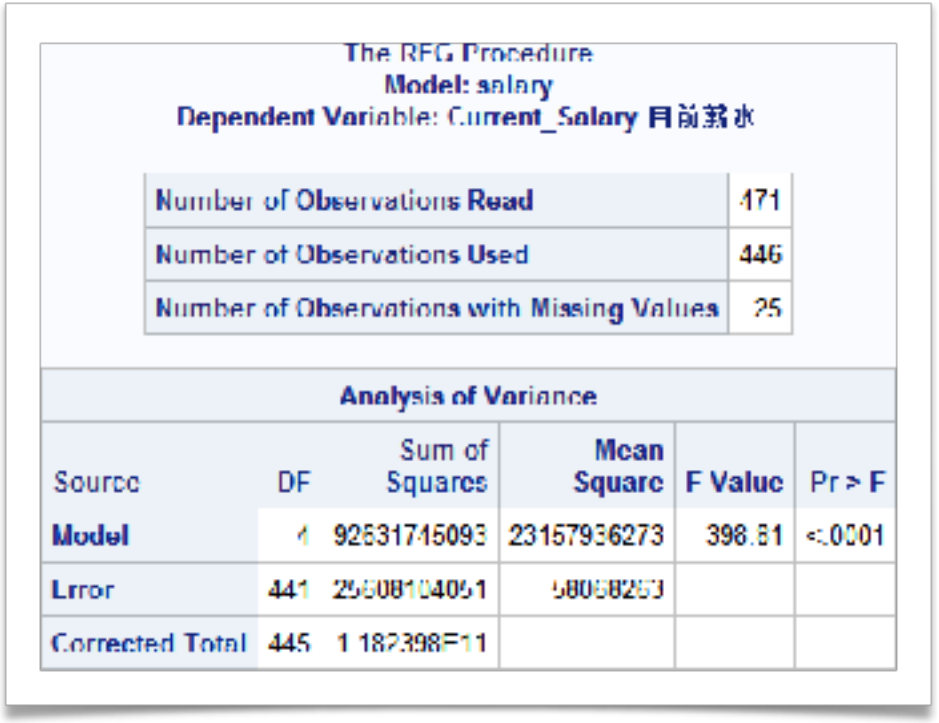
Obs	ID 编号	Gender 性别	Education 受教育年限	Position 职位	Current_Salary 当前薪水	Begin_Salary 起薪	Experience 已工作周数	Age 年龄
1	1	1	15	1	57000	27000	144	55
2	34	1	19	1	92000	39990	175	58
3	18	1	16	1	103750	27510	70	51
4	200	1	17	1	67500	34980	9	44
5	199	1	16	1	51250	27480	69	49
					.....			
468	326	1	8	2	29550	15750	144	48
469	48	1	12	2	30750	14100	240	59
470	98	1	8	2	30000	15000	144	50
471	126	1	15	2	24300	15000	191	56

多元线性回归分析

- 在图 12-9 所示的背景资料中，性别变量 Gender 和职位变量 Position 都是定性变量，对于定性自变量的回归有其特殊的处理方法（见第 12.2 节），本例的分析暂不考虑定性变量对因变量的影响。现考虑其余定量变量对当前薪酬的影响。
- 对于多元回归分析，同样可以利用 REG 过程进行分析，本例的程序如下：

```
proc reg data=sasuser.salary;
    salary: model current_salary =education begin_salary experience age;
run;
quit;
```

- 程序运行之后，首先可得到如图 12-10 所示的回归方程总体显著性检验结果。



The REG Procedure
Model: salary
Dependent Variable: Current_Salary 目前薪水

Number of Observations Read	471
Number of Observations Used	446
Number of Observations with Missing Values	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	92631745093	23157936273	398.81	<.0001
Error	441	25508104051	58068263		
Corrected Total	445	1.182398E11			

多元线性回归分析

- 回归方程总体显著性检验结果显示，F 值=398.81，其对应 P 值<0.0001，非常显著。同时，图 12-11 中的拟合优度判定系数R^2和调整的R^2也较高。
- 参数估计的结果显示在图 12-12 中。
- 通过各自变量对应回归系数的 t 检验 P 值（P>|t|）的大小，可以判定各个回归系数是否显著（截距项通常情况下不用检验）。在图 12-12 中，如果给定显著性水平 α=0.1，Education、Experience 和 Begin_Salary 的 P 值均小于 0.1，因而在模型中影响显著；而 Age 在显著性水平 α=0.1 下不显著，说明年龄对当前薪酬的影响不显著。

Root MSE	7620.25340	R-Square	0.7834
Dependent Mean	34592	Adj R-Sq	0.7815
Coeff Var	22.02910		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	703.31182	3147.80447	0.22	0.8233
Education	教育年限	1	490.50221	180.72218	2.71	0.0069
Begin_Salary	起始薪水	1	1.89364	0.07222	26.22	<.0001
Experience	工作经历(月)	1	-11.62766	5.97644	-1.95	0.0523
Age	年龄	1	-74.92635	53.23112	-1.41	0.1600

多元线性回归分析

- 对于回归系数不显著的变量，应当在模型中剔除。在 REG 过程中可以配合 MODEL 语句的选项关键字 SELECTION=按照设定的显著性水平自动剔除回归系数不显著的变量，其具体语句如下：

  MODEL 因变量列表=自变量列表 /SELECTION=变量筛选方法 SLE=进入阈值 SLS=剔除或保留阈值

- MODEL 语句的 SELECTION=选项提供了 8 种建模过程中自动剔除变量的方法， 具体关键字及其功能如下：

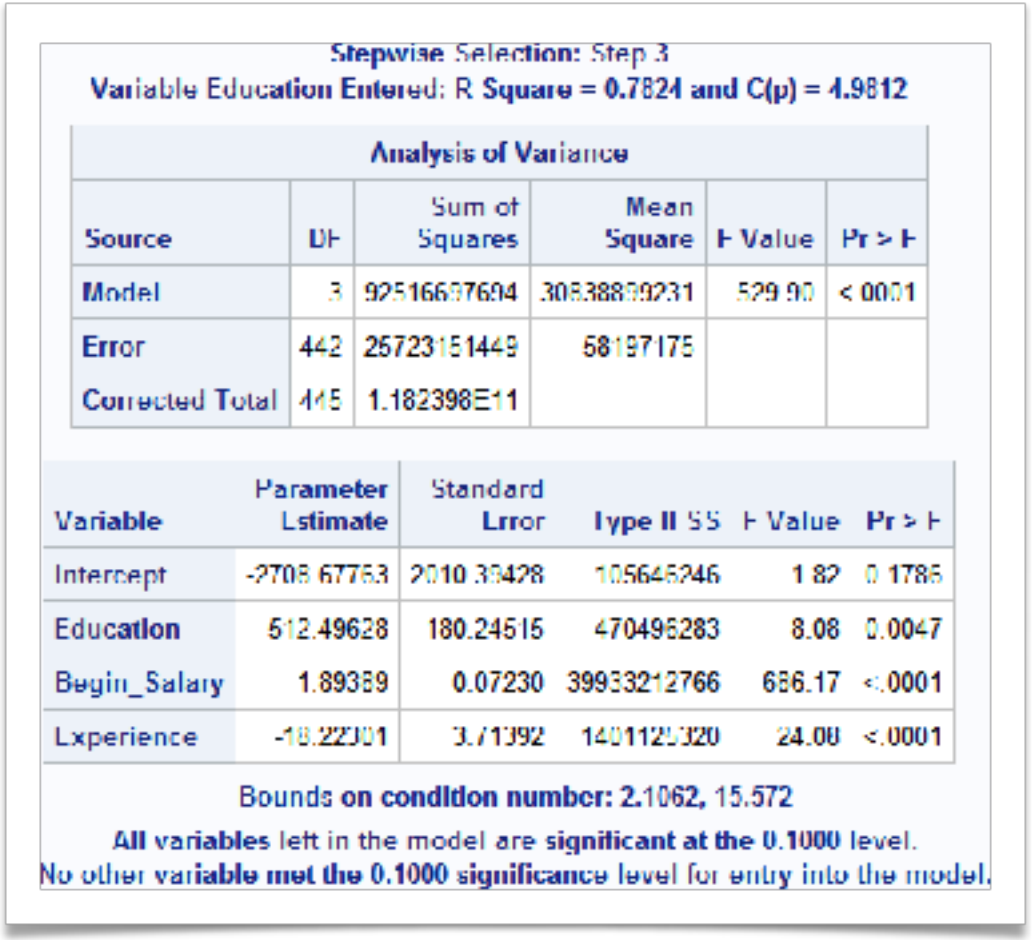
  - FORWARD 或 F：向前引入法。即向模型中逐个引入变量，建模伊始，模型中没有自变量， 每当引入一个自变量之后， 便计算回归方程 F 统计量的值及其对应的 P 值，系统根据用户设定的 P 值阈值决定该变量是否应当引入；
  - BACKWARD 或 B，向后剔除法。即从模型中逐个剔除变量，建模伊始，模型中包含所有的自变量， 每当剔除一个自变量之后， 便计算回归方程 F 统计量的值及其对应的 P 值， 系统根据用户设定的 P 值阈值决定该变量是否应当剔除；
  - STEPWISE：逐步回归法。即根据用户设定 P 值的引入阈值，逐个向模型引入自变量；然后重新计算模型 P 值，根据用户设定的剔除阈值进行变量筛选；
  - MAXR：最大R^2增量法。即穷尽所有变量组合所构成的模型，找出使得模型拟合优度判定系数R^2增加最大的模型；
  - MINR：最小R^2增量法。类似于最大 R 2 增量法，不同的是最后选择的模型是使得R^2增加最小的模型；
  - CP：Mallow Cp 统计量法，即根据 Mallow Cp 统计量进行模型变量选择；
  - RSQUARE：R2 选择法。即根据R^2进行模型变量选择；
  - ADJRSQ：修正 R2 选择法，即根据修正R^2进行模型变量选择；
  - NONE：全变量模型。即不指定任何变量筛选的方法，把 MODEL 语句等号右边的所有自变量均放入模型中进行参数估计，这是 REG 过程系统默认情况下采用的方法。

多元线性回归分析

- 变量筛选在统计分析和机器学习中是非常重要的， 在 SAS 系统中不光是 REG 过程可以指定变量筛选， 其他很多过程的统计建模均可以进行变量筛选， 如当变量数目有成千上万时，可在使用 GLMSELECT 过程进行分析时，可以在 MODEL 语句的 SELECTION=指定 LAR（最小角回归）或 LASSO 等较为前沿的变量选择方法。
- 本例拟采用逐步回归法进行变量筛选，并设置变量进入模型的显著性水平为 0.1，被剔除出模型的显著性水平也为 0.1，程序如下：

```
proc reg data=sasuser.salary;
    salary_selection: model current_salary =education begin_salary experience age /selection=stepwise
  sle=0.1 sls=0.1;
run;
quit;
```

- 程序运行之后， 在输出窗口中会看到对于本例数据的模型选择问题一共经历了 3 次回归建模之后最终得到通过检验的模型，3 次回归过程均会详细的在结果中显示出来。本书节选出第 3 次建模的结果如图 12-13 所示。



Stepwise Selection: Step 3

Variable Education Entered: R Square = 0.7824 and C(p) = 4.9812

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	92516697694	30838899231	529.90	< 0001
Error	442	25723151449	58197175		
Corrected Total	445	1.182398E11			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-2708.67753	2010.39428	105646246	1.82	0.1785
Education	512.49628	180.24515	470496283	8.08	0.0047
Begin_Salary	1.89389	0.07230	39933212766	686.17	<.0001
Experience	-18.22301	3.71392	1401125320	24.08	<.0001

Bounds on condition number: 2.1062, 15.572

All variables left in the model are significant at the 0.1000 level.

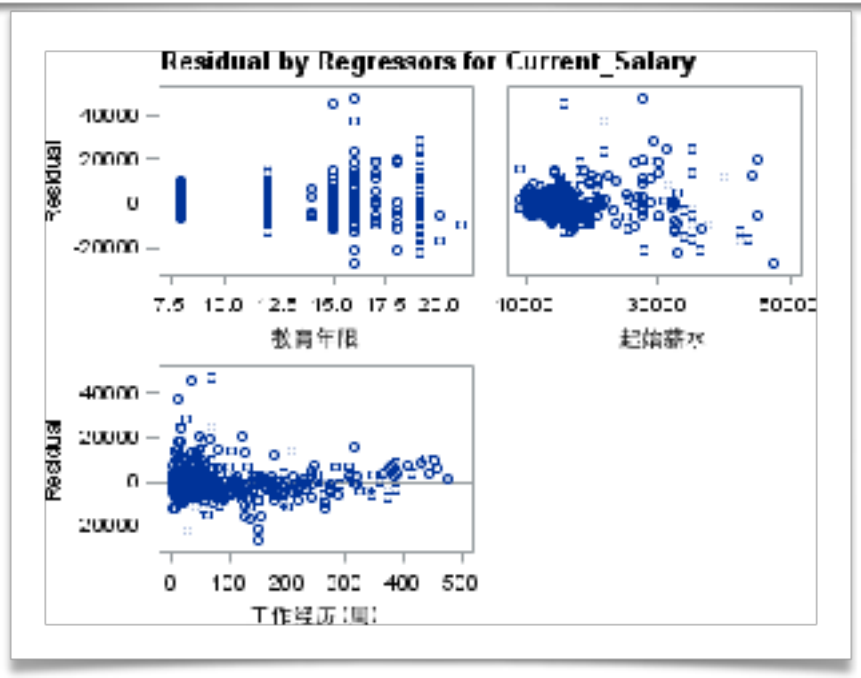
No other variable met the 0.1000 significance level for entry into the model.

多元线性回归分析

- 图 12-13 显示， 最终模型当中含有 Education、 Begin_Salary和 Experience 等 3 个自变量，且均通过 F 检验和回归系数显著性检验。而且系统最终还会提示，在模型中的所有自变量回归系数的显著性（即 P 值）都小于 0.1，且在 0.1 的条件下没有其他的自变量可以引入回归模型当中。此外，系统还给出了个变量筛选步骤的简要总结，如图 12-14 所示。
- 图 12-15 给出了残差项与各自变量之间的散点图。从这些图形来看，可以认为模型随机误差项与各自变量均没有显著影响关系，且均值趋近于 0。

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Begin_Salary		起始薪水	1	0.7581	0.7581	50.4818	1391.77	<.0001
2	Experience		工作经历(周)	2	0.0203	0.7785	11.0837	40.66	<.0001
3	Education		教育年限	3	0.0040	0.7824	4.9812	8.08	0.0047

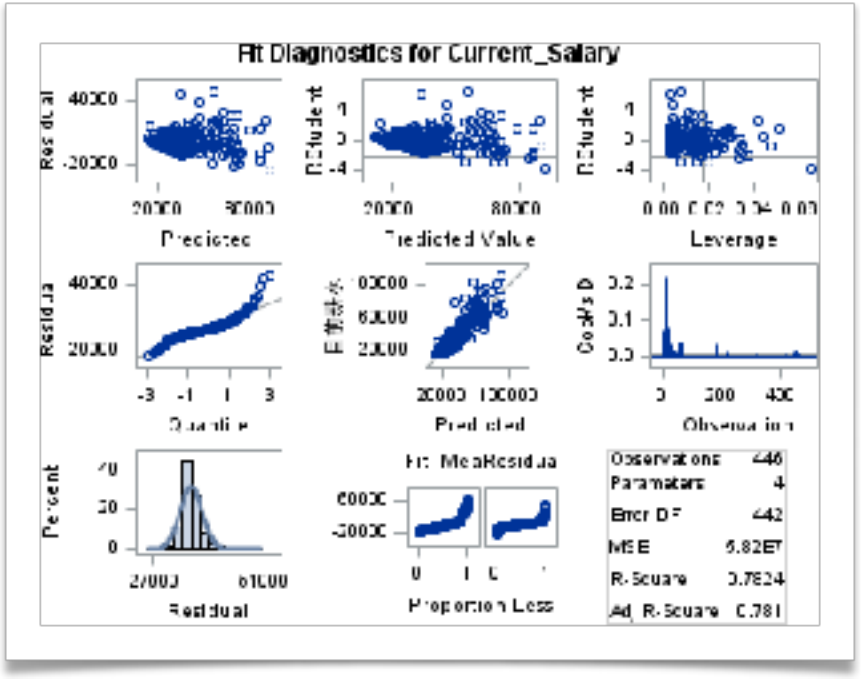


多元线性回归分析

- 从图 12-16 所示的模型诊断图形来看，残差项基本符合正态分布，本例数据分布特征符合线性回归建模的基本要求。
- 至此，根据逐步回归法，可以得到如下的回归方程：

Current_Salary = -2708.68 + 512.50 × Education + 1.89 × Begin - 18.22 × *Experience*

- 从回归方程中可以得知，Education 变量对因变量的平均影响最大，当学历每增加一个单位时，当前薪酬平均增加 512.50 个单位；起始薪酬状况对当前薪酬的影响不大，说明公司在考虑当前薪酬的时候主要考虑个人素质， 即学历以及工作经验等主观因素； 而年龄对当前薪酬的影响不显著，不能作为制定薪金的主要考虑因素。



定性自变量回归分析

- 在影响因变量的诸多因素中， 除了定量变量的影响之外， 有些时候还有一些定性因素的影响，如例 12-2 中的性别、职位等对当前薪酬的影响。定性因素对因变量的影响在进行回归分析的过程中，需要进行特殊的处理，即应当把定性变量转化为虚拟变量（或哑变量）之后再引入回归模型中进行分析。

虚拟变量的设定

- 虚拟变量的设定即是把对变量的定性描述转化成定量数据来进行描述， 如性别定性变量有“男”和“女”2 种表现，在设定虚拟变量时，可考虑用“0”、“1”数字分别代表“男”、“女”，则性别在 SAS 系统中便可转化为数值型变量进行分析了。
- 设定虚拟变量时应当遵循如下原则：
  - 对于有 k 个表现值的定性变量，只设定（k-1）个虚拟变量；
  - 虚拟变量的值通常用“0”或“1”来表示；
  - 对于每个样本而言，同一个定性变量对应虚拟变量的值之和不超过 1。
- 如性别变量，有 2 个表现值，即“男”和“女”（即 k=2），因此只需设定 1 个虚拟变量即可，可以考虑用“0”代表女性，“1”代表男性。

虚拟变量的设定

- 而例 12-2 中的职位变量，有 3 个表现值（即 k=3），因此需要设定 2 个虚拟变量来进行分析，如表 12-1 所示。

表 12-1 虚拟变量的设定

	虚拟变量		含义
	Position1	Position2	
“职位”定性变量	0	0	普通员工
	0	1	主管
	1	0	经理

- 在表 12-1 所示的虚拟变量中，其具体数值表示何种含义，用户可以根据自身需求进行指定。 本例按照表 12-1 的关系指定 “职位” 的虚拟变量， 如果考虑例 12-1 中的所有自变量，建立的回归模型为：

$$Current\_Salary = \alpha + \beta_1 \times Gender + \beta_2 \times Education + \beta_3 \times Position1 + \beta_4 \times Position2 + \beta_5 \times Begin\_Salary + \beta_6 \times Experience + \beta_7 \times Age + \varepsilon$$

虚拟变量的设定

- 当虚拟变量 Gender 为 0 时，回归方程中不含 Gender 变量，表示女性职员当前薪酬的影响状况，而 Gender 为 1 时，回归方程中含有 Gender 变量，表示男性职员当前薪酬的影响状况；同理，当 Position1 和 Position2 同时为 0 时，表示普通员工的当前薪酬影响状况。
- 对于本例数据，可以按照上述理论分析过程，使用如下程序生成虚拟变量 POSITION1和POSITION2：

```
data sasuser.salary_dummy;
  set sasuser.salary;
  if position=1 then do;position1=1;position2=0;end;
  if position=2 then do;position1=0;position2=1;end;
  if position=3 then do;position1=0;position2=0;end;
run;
```

含有虚拟变量的回归分析

- 在对含有虚拟变量的实际问题进行线性回归分析时，仍然可调用 REG 过程，其语法与定量变量回归分析一样，只要把虚拟变量当作解释变量加入模型即可，如本例的程序如下：

```
proc reg data=sasuser.salary_dummy;
    model current_salary =gender education begin_salary experience position1 position2;
run;
quit;
```

- 本段程序已经剔除了影响不显著的 AGE 变量，程序运行之后可以得到如图 12-17 所示的系列结果。
- 在图 12-17 中可以看到所有变量的回归系数均在 α=0.1 条件下显著，且回归方程拟合优度的判定系数 R 2 =0.8155，其总体显著性检验的 F 统计量为 324.13，对应 P 值<0.0001，非常显著。

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	96424450963	16070741826	324.13	<.0001
Error	440	21816526239	49583741		
Corrected Total	446	1.1824E11			

Root MSE	7041.35035	R Square	0.8155
Dependent Mean	34090	Adj R Sq	0.8130
Coeff Var	20.65613		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3676.08499	2151.25486	1.85	0.0652
Gender	性别	1	1958.21124	812.92852	2.41	0.0164
Education	教育年限	1	503.10106	172.31207	2.92	0.0037
Begin_Salary	工资基本	1	1.01960	0.09651	10.56	<.0001
Experience	工作经历(月)	1	22.52242	3.77511	5.97	<.0001
position1		1	11573	1538.71385	7.52	<.0001
position2		1	6635.72266	1691.17205	3.96	<.0001

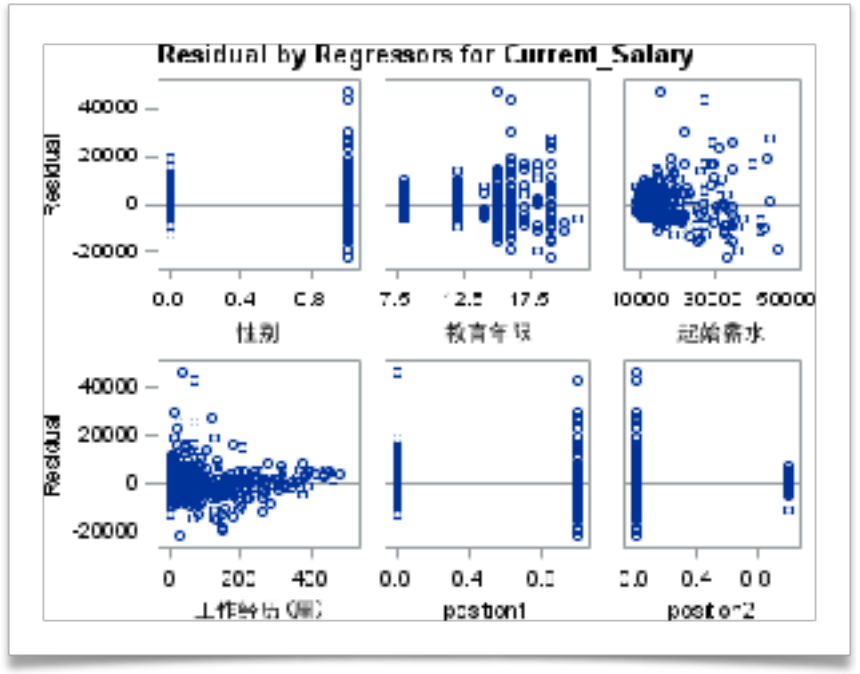
含有虚拟变量的回归分析

- 而根据图 12-18 所示的残差与所有自变量的散点图（限于篇幅，本例省略了用于回归模型诊断的图形面板）显示，残差项是随机的，其期望趋于 0，且与自变量均无显著关系。
- 根据参数估计结果，可写出回归方程如下：

$$Current\_Salary = 3976.08 + 1958.24 \times Gender + 503.46 \times Education + 11573 \times Position1 + 6695.72 \times Position2 + 1.32 \times Begin\_Salary - 22.52 \times Experience$$

- 对含有虚拟变量的回归方程进行分析， 应当先确定分析的参照方程。 参照方程就是指当所有虚拟变量为 0 时的方程，本例的参照方程为：

$$Current\_Salary = 3976.08 + 503.46 \times Education + 1.32 \times Begin\_Salary - 22.52 \times Experience$$



含有虚拟变量的回归分析

- 因本例中有两个虚拟变量， 故所有虚拟变量均为 0 时的参照方程表示女性 （Gender=0） 、且职位为普通员工（Position1=Position2=0）的当前薪酬影响关系。参照方程的具体含义即女性普通员工的学历每增加 1 年，当前薪酬平均增加 503.46 元；工作经验增加 1 周，则当前薪酬反而平均减少 22.52 元；而起薪增加 1 元，则当前薪酬平均增加 1.32 元。
- 对于不同职位的男女性别员工的薪酬影响，可以根据对应虚拟变量的取值来进行分析。如要分析职位为经理的男性薪酬状况，即把虚拟变量 Gender=1，Position1=1，Position2=0代入估计方程中，得到：

$$\begin{aligned} Current\_Salary &= 3976.08 + 1958.24 + 503.46 \times Education + 11573 \\ &\quad + 1.32 \times Begin\_Salary - 22.52 \times Experience \\ &= 3976.08 + 13531.24 + 503.46 \times Education \\ &\quad + 1.32 \times Begin\_Salary - 22.52 \times Experience \end{aligned}$$