利用逐步回归法建立国家财政收入的多元线性回归模型

SY2106346 郑耀彦

2021.12

摘要

在多元线性回归分析中，由于有多个自变量，存在一些在一元线性回归分析中不会遇到的问题。回归自变量的选择无疑是建立回归模型的一个极为重要的问题。在建立一个回归模型时，首先碰到的问题就是如何确定回归自变量，一般情况大都是根据所研究问题的目的，结合相关理论列出对因变量可能有影响的一些因素作为自变量。但是某些自变量数据的质量可能很差，或者和因变量之间没有较好的线性关系，如果不把它们排除在外的话会降低模型精度，直接影响到回归方程的应用。因此最优回归方程的选择是一个重要问题，目前主要采用的是全部比较法、向前法、向后法、逐步回归法。而本文选择使用了逐步回归法挑选了较为有效的几个自变量，建立了国家财政收入的多元线性回归模型。源代码位于 *https://github.com/BUAAHugeGun/regression*

关键词: 多元线性回归，逐步回归法，回归模型

Abstract

In multiple linear regression analysis, there are many independent variables, there are some problems that will not be encountered in linear regression analysis.The choice of regression independent variable is undoubtedly a very important problem in establishing regression model. In the establishment of a regression model, the first problem we encounter is how to determine the independent variables of regression. In general, some factors that may affect the dependent variables are listed as independent variables according to the purpose of the problem studied and combined with relevant theories.However, the quality of some independent variable data may be very poor, or there is no good linear relationship between the dependent variables. If they are not excluded, the model accuracy will be reduced and the application of regression equation will be directly affected. Therefore, the selection of optimal regression equation is an important problem. At present, total comparison method, forward method, backward method and stepwise regression method are mainly used. In this paper, stepwise regression method is used to select several more effective independent variables and establish a multiple linear regression model of national fiscal revenue. Source code is located in the *https://github.com/BUAAHugeGun/regression*

Keywords: Multiple Linear Regression, Stepwise Regression Method, Regression Model

目录

1 背景介绍

本文主要根据最小二乘法和逐步回归法对可能影响国家财政收入的六个因素进行了多元线性回归模型的建立。接下来主要介绍线性回归中的最小二乘法和逐步回归方法。

1.1 最小二乘法

线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。

一般在线性回归中，为了使得模型拟合的直线与样本之间更接近，人们设计了一些损失函数，比如绝对误差和 $\sum_{i=1}^{n}|y_i-\hat{y}_i|$ 和平方和误差 $Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，而平方和误差不需要考虑正负号，方便求导，因此最小化平方和误差的回归方法使用最为频繁，也即最小二乘法。

在一元线性回归中，设 $\hat{y}_i=bx_i+a$，则平方和误差 $Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-a-bx_i)^2$。为了使得 Q 尽量小，求其关于 a 和 b 的偏导并令它们为 0：

$$\begin{cases}\dfrac{\partial Q}{\partial a}=-2\displaystyle\sum_{i=1}^{n}(y_i-a-bx_i)=0\\[2ex]\dfrac{\partial Q}{\partial a}=-2\displaystyle\sum_{i=1}^{n}(y_i-a-bx_i)x_i=0\end{cases}$$

整理得正规方程组：

$$\begin{cases}na+n\overline{x}b=n\overline{y}\\[2ex]n\overline{x}a+\displaystyle\sum_{i=1}^{n}x_i^2b=\sum_{i=1}^{n}x_iy_i\end{cases}$$

解得：

$$\begin{cases}\hat{a}=\overline{y}-\hat{b}\overline{x}\\[2ex]\hat{b}=\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\end{cases}$$

这样求得的回归方程 $\hat{y}=\hat{a}+\hat{b}x$ 能让平方和误差最小，同时 $\hat{a}$ 和 $\hat{b}$ 是参数 a 和 b 的一致最小方差线性无偏估计。如图 1所示，最小二乘法能对数据做出良好的线性回归建模。

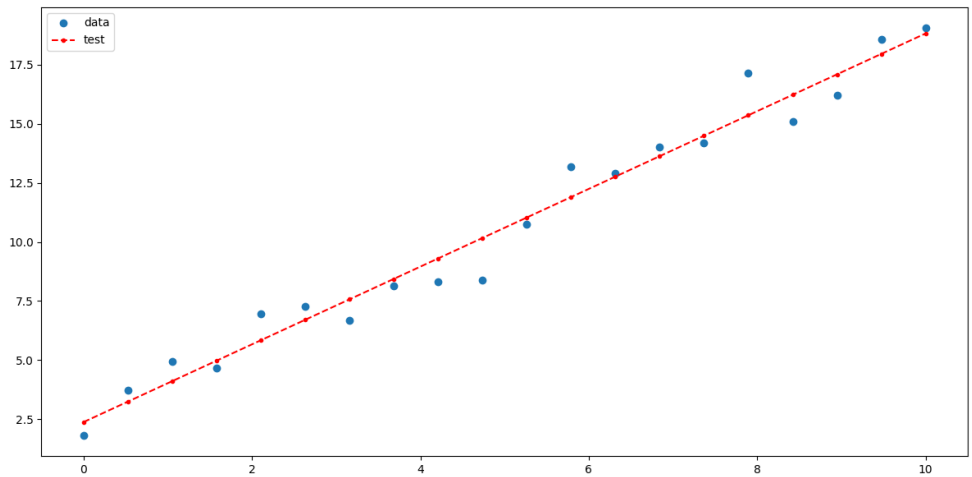


图 1: 使用最小二乘法进行一元线性回归的结果

1.2 多元线性回归

在许多实际问题中，影响事物的因素常常不止一个，要找出这些因素与事物之间的数量关系就是多元回归分析的任务。由于许多非线性的情形都可以化为线性回归来处理。因此在一般情况下，只要能处理多元线性回归问题就足够了。

对于多元线性回归而言同样可以使用最小二乘法求得 p 元线性回归方程 $\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i x_i$，其中$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}$ 使用最小二乘法解出。

然而，在多元线性回归分析中，由于有多个自变量，某些自变量数据的质量可能很差，或者和因变量之间没有较好的线性关系，如果不把它们排除在外的话会降低模型精度，直接影响到回归方程的应用。我们既希望模型中包含的自变量尽量丰富，而有需要考虑到每个自变量尽可能与因变量有较好的线性关系，所以如何建立最优回归方程是一个重要问题。

1.2.1 偏 F 检验

对于所有的自变量，我们可以使用最小二乘法求出线性回归方程，但是同时我们想要将对 y 没有显著影响的自变量排除在外，因此对于每一个自变量我们需要检验其对于多元线性回归模型的贡献是否显著。常用的检验方法为偏 F 检验，对于检验假设

$$H_0 : \beta_i = 0, \quad H_1 : \beta_i \neq 0$$

当 H_0 为真时，检验统计量为:

$$F_i = \frac{U_i}{\hat{\sigma}^2} = \frac{U_i}{Q/(n-p-1)} \sim F(1, n-p-1)$$

其中 $U_i = U - U_i'$，U 和 U_i' 分别为排除第 i 个自变量前后的两个回归模型的回归平方和，并称 u_i 为偏回归平方，而 Q 则为平方和误差。对于给定的显著性水平 α，由样本值计算偏 F 检验统计量 F_i 的值，若 $F_i \geq F_{1-\alpha}(1, n-p-1)$，则拒绝 H_0，认为第 i 个自变量对 y 有显著影响，不应该排除 x_i。反之若 $F_i < F_{1-\alpha}(1, n-p-1)$ 则认为第 i 个自变量对 y 的影响不显著，将其剔除。

1.2.2 逐步回归法

逐步回归法可以从很多候选自变量中将对 y 有显著影响的自变量选择出来并建立多元线性回归模型。其操作步骤如下:

(1) 设当前模型为 M，在备选自变量中对每个自变量 x_i，加入当前模型形成 M_i 并对 x_i 做显著性检验 (偏 F 检验)。将偏 F 检验统计量最大的显著自变量加入到模型当中，若没有这样的自变量则不加入。

(2) 设当前模型为 M，在已选自变量中对每个自变量 x_i 进行显著性检验 (偏 F 检验)。将偏 F 检验统计量最小的不显著自变量移除出当前模型，若没有这样的自变量则不移除。

(3) 重复以上步骤直到既不能在模型中加入自变量也不能在模型中移除自变量。

2 实验过程

本文在《中国统计年鉴》中查询了六种自变量和国家财政收入的近 43 年的数据；使用 mathematica 软件计算了各 F 分布的 0.95 分位数用于做逐步回归法中的在显著性水平 0.05 下的偏 F 检验；使用 python 中的 statsmodel 做多元线性回归并将结果可视化并将逐步回归法和枚举法得到的多元线性回归模型进行了对比。

2.1 F 分布分位数的计算

首先在 mathematica 中构建标准正态分布 $N(0,1)$，再使用 TransformedDistribution 和函数构建任意自由度的卡方分布，最后再用 TransformedDistribution 和函数构建任意自由度的 F 分布并使用 Quantile 求出 $F(1,41), F(1,40), F(1,39), F(1,38), F(1,37), F(1,36)$ 的 0.95 分位数：4.07855, 4.08475, 4.09128, 4.09817, 4.10546, 4.11317。

2.2 逐步回归法

本文借助了 python 中的 statsmodels 库可以对任意一组自变量进行多元线性回归模型的建立，并求出所有的系数。借助每次回归模型的系数本文完成了对其中任意一个变量的偏 F 检验统计量的计算，并将其和上文求出的 F 分布的 0.95 分位数进行比较，完成了逐步回归法，建立了多元线性回归模型。逐步回归法的伪代码如算法 1所示。

2.3 枚举法

本文同时使用枚举法，枚举了所有自变量的组合，对它们分别进行了多元线性回归之后，选出了 F 检验统计量最大的方案，和逐步回归法得出的方案作了对比。

算法 1 逐步回归法

输入: 自变量集合 $\mathbb{U}=\{X_1,X_2,...,X_p\}$, $X_i=(x_{i1},...x_{in})$，因变量向量 $Y=(y_1,...,y_n)$，F 分布 0.95 分位数 $F(1,k),k\in[36,41]$

输出: 选中的自变量集合 $\mathbb{S}$

function StepwiseRegression($\mathbb{U},Y$)
  $\mathbb{S}\leftarrow\{\}$
  repeat
    $end\leftarrow True$
    $X'\leftarrow\arg\max\{\text{PartialFTest}(\mathbb{S}\cup\{X\},X)|X\in\mathbb{U}\}$
    if PartialFTest($\mathbb{S}\cup\{X'\},X'$) $\geq F(1,n-|\mathbb{S}|-2)$ **then**
      $\mathbb{S}\leftarrow\mathbb{S}\cup X'$
      $end\leftarrow False$
    end if
    $X'\leftarrow\arg\min\{\text{PartialFTest}(\mathbb{S},X)|X\in\mathbb{S}\}$
    if PartialFTest($\mathbb{S},X'$) $< F(1,n-|\mathbb{S}|-1)$ **then**
      $\mathbb{S}\leftarrow\mathbb{S}\setminus X'$
      $end\leftarrow False$
    end if
  until $end=True$
  return $\mathbb{S}$
end function

function PartialFTest($\mathbb{X},X_i$)
  $q\leftarrow|\mathbb{X}|$
  Ordinary Least Squares: $\hat{y_{ur}}\leftarrow fit(Y,\mathbb{X})$
  Ordinary Least Squares: $\hat{y_r}\leftarrow fit(Y,\mathbb{X}\setminus\{X_i\})$
  $partialESS\leftarrow\sum_{i=1}^{n}(\hat{y_{ur}}_i-\overline{y})^2-\sum_{i=1}^{n}(\hat{y_r}_i-\overline{y})^2$
  $RSS\leftarrow\sum_{i=1}^{n}(\hat{y_{ur}}_i-y_i)$
  return $\frac{partialESS}{RSS/(n-q-1)}$
end function

3 实验结果

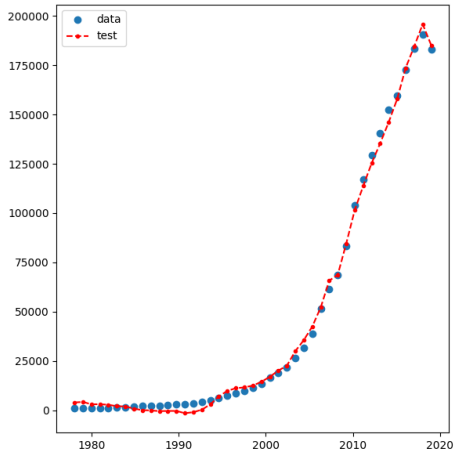
3.1 所有自变量

本文先使用对所有自变量做了一次多元线性回归以作对比，如图 2所示，虽然从预测值图像上看效果不错，但是实际上并不是所有自变量都和因变量紧密相关，F 检验统计量的值为 2900。

Dep. Variable:	y	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	2900.
Date:	Thu, 09 Dec 2021	Prob (F-statistic):	8.80e-47
Time:	20:15:28	Log-Likelihood:	-402.98
No. Observations:	43	AIC:	820.0
Df Residuals:	36	BIC:	832.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.354e+04	1.25e+04	5.094	0.000	3.82e+04	8.88e+04
x1	0.3880	0.334	1.163	0.252	-0.289	1.065
x2	0.4089	0.060	6.792	0.000	0.287	0.531
x3	-2.1990	0.513	-4.287	0.000	-3.239	-1.159
x4	0.5359	0.077	6.930	0.000	0.379	0.693
x5	-0.6001	0.135	-4.461	0.000	-0.873	-0.327
x6	-0.6704	0.907	-0.739	0.465	-2.511	1.170

(a) 分析结果与指标



(b) 预测结果与真实值比较

图 2: 所有自变量多元回归分析结果

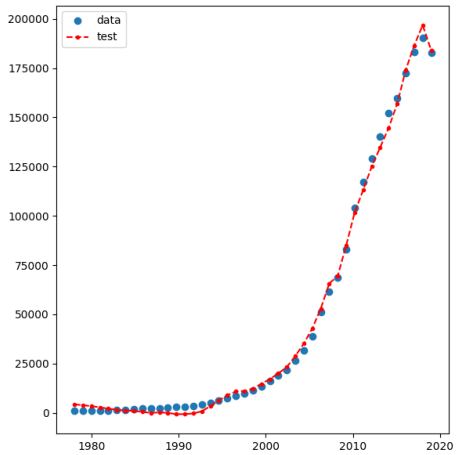
3.2 逐步回归法

从结果来看，使用逐步回归法在原来的六个自变量中选择了四个比较好的自变量，分别是：社会商品零售总额、工业总产值、人口、建筑业总产值。如图 3所示，逐步回归法选择的四个自变量构成的模型性能比所有自变量更好，其 F 检验统计量值为 4337，远大于使用所有自变量得出的结果。

Dep. Variable:	y	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	4337.
Date:	Thu, 09 Dec 2021	Prob (F-statistic):	5.66e-50
Time:	20:15:28	Log-Likelihood:	-404.20
No. Observations:	43	AIC:	818.4
Df Residuals:	38	BIC:	827.2
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.412e+04	8396.708	6.445	0.000	3.71e+04	7.11e+04
x1	0.5246	0.074	7.096	0.000	0.375	0.674
x2	0.4692	0.042	11.289	0.000	0.385	0.553
x3	-0.5305	0.076	-6.935	0.000	-0.685	-0.376
x4	-2.0233	0.464	-4.361	0.000	-2.962	-1.084

(a) 分析结果与指标



(b) 预测结果与真实值比较

图 3: 逐步回归法多元回归分析结果

3.3 枚举法

本文使用枚举法枚举了 31 种自变量的选法，选择了 F 检验统计量最大的方案，和逐步回归法对比发现两种方法得到的结果完全一样，这既说明了逐步回归法在大大降低计算量的同时，得出的结果也非常好。