Advances in Education 教育进展, 2022, 12(8), 2876-2883
Published Online August 2022 in Hans. http://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2022.128435

Hans汉斯

同时处理多种回归异常的一般方法

吕林黛[1]，赵胜利[1]，汪　欣[1]，钟妤玥[2]

[1]重庆理工大学，重庆
[2]重庆渝高中学，重庆

收稿日期：2022年7月11日；录用日期：2022年8月11日；发布日期：2022年8月17日

摘　要

回归分析是以概率论与数理统计为基础，主要对随机现象统计资料进行分析和推断。而在实际的应用中发现同时满足基本假设的数据是非常少的，通常会出现多重共线性，异方差性和自相关性的问题。然而大多数的教材都只是给出了出现单一的问题时的解决办法，因此本文为了解决当一个回归分析中出现上述三个问题中的两个或者三个的时候应当以何种顺序解决的问题，采用了理论分析和实例验证的方法。从实例和分析结果来看当同时出现违背基本假设的多种情况下，如果数据为截面数据时的处理顺序是多重共线性－异方差性－自相关性，当数据为时间系列数据时的处理顺序为多重共线性－自相关性－异方差性。

关键词

多重共线性，异方差性，自相关性

A General Approach to Handling Multiple Regression Exceptions at the Same Time

Lindai Lyu[1], Shengli Zhao[1], Xin Wang[1], Yuyue Zhong[2]

[1]Chongqing University of Technology, Chongqing
[2]Chongqing Yugao Middle School, Chongqing

Received: Jul. 11th, 2022; accepted: Aug. 11th, 2022; published: Aug. 17th, 2022

Abstract

Regression analysis is based on probability theory and mathematical statistics, and mainly analyzes and infers the statistical data of random phenomena. In practical applications, very few data are found to meet the basic assumptions at the same time, and there are often problems of multicollinearity, heteroscedasticity and autocorrelation. However, most textbooks only give solutions to a

文章引用: 吕林黛, 赵胜利, 汪欣, 钟妤玥. 同时处理多种回归异常的一般方法[J]. 教育进展, 2022, 12(8): 2876-2883.
DOI: 10.12677/ae.2022.128435

single problem, so this paper uses theoretical analysis and example verification methods to solve the problems in what order two or three of the above three problems should be solved in a regression analysis. From the example and analysis results, when there are multiple cases that violate the basic assumptions at the same time, the processing order is multicollinear-heteroscedasticity-autocorrelation when the data are cross-sectional data, and the processing order is multicollinear-autocorrelation-heteroscedasticity when the data are time series data.

Keywords

Multicollinearity, Heteroscedasticity, Autocorrelation

1. 引言

《应用回归分析》是一门在自然科学、管理科学和社会、经济等领域应用十分广泛的统计类课程。回归分析是以概率论与数理统计为基础，主要对随机现象统计资料进行分析和推断。在“大数据”时代背景下，学习和掌握应用回归分析理论，对于提高分析和解决实际问题的能力具有重大的意义。通过对课本的学习了解到回归模型的建立过程如图 1 所示。

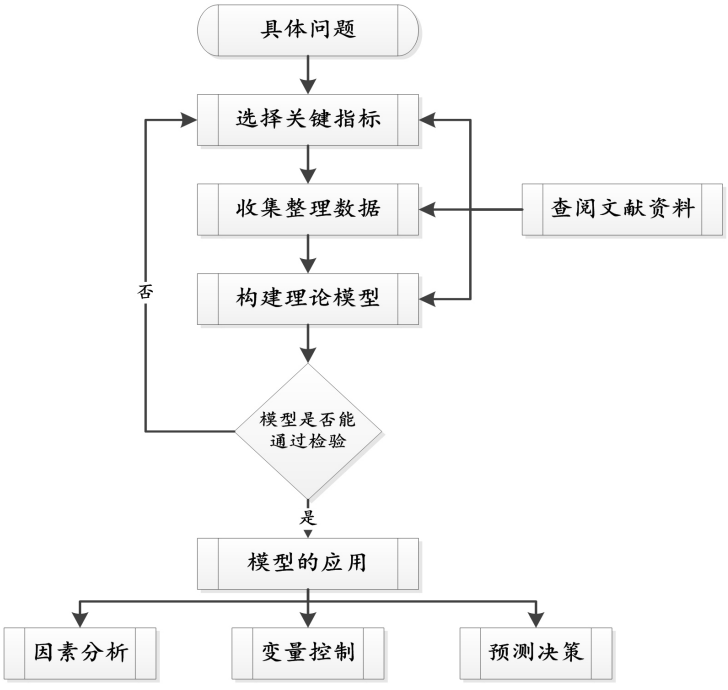

Figure 1. Flow chart of regression model establishment
图 1. 回归模型建立流程图

根据课本内容建立多元线性回归模型 $Y=X\beta+\varepsilon$ 的基本假设有：$x_1,x_2,\cdots,x_p$ 是确定性变量；$E(\varepsilon_i)=0$，当 $i=j$ 时 $cor(\varepsilon_i,\varepsilon_j)$ 为 δ^2 或 $i\neq j$ 为 0；$\varepsilon_i\sim N(0,\delta^2)$，$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 相互独立，$Y\sim N(X\beta,\delta^2 I_n)$。而在实

际的应用中发现同时满足基本假设的数据是非常少的，通常会出现多重共线性，异方差性和自相关性的问题。然而大多数的教材都只是给出了出现单一的问题时的解决办法，比如当出现多重共线性时通常采用逐步回归法和经验法；当出现异方差性时通常采用模型变换法、加权最小二乘法和模型的对数变换法；当出现自相关性时通常采用广义差分法和科克伦－奥科特迭代法[1]，并未考虑三种情况同时出现或其中两个同时出现时的情况。然而三种回归异常现象形成的原因和所造成的回归异常皆有所不同，若随意的进行处理，可能将使回归方程的拟合度达不到最优，更有甚者使得参数估计量经济含义不合理，从而失去实际应用价值。因此本文为了研究当一个回归问题中出现了上述三个回归异常中的两个或者三个的时候该如何解决的问题，下文将从理论分析的角度进行结论说明，再进行实例验证本文观点。

2. 同时处理多种回归异常的一般方法

由于大部分的教材针对在进行模型回归时，存在的多重共线性、异方差和自相关问题，通常只给出针对单一问题的解决方案，而在实际应用中可能出现多种异常都存在的情况，这种情况下直接线性回归会失效，而如何处理这种存在多种回归异常的情况目前还没有定论，因此本文就此进行了详细的探讨。查阅资料知当解释变量中存在多重共线性时将会导致参数估计值的方差增大，变量的显著性检验失去意义，区间估计和区间预测功能失效和参数估计量经济含义不合理[2]。其中参数估计值的方差增大指的时虽然有 OLS 得出的 β 任然是线性无偏的，但不再是最小方差估计，不能准确的反应数据特征，同时变量的显著性检验失去意义也可能将重要的解释变量排除在回归拟合的模型之外，使得回归方程的拟合优度达不到最优[3]。也会导致参数估计值的方差增大，而变大的方差容易使得区间预测的范围变大，使估计值稳定性变得很差，从而失去精确度，预测失去意义。在实际应用中建立回归模型就是为了解决具体的社会经济问题，运用模型进行经济因素的分析，经济变量的控制和经济决策的预测，这属于建立回归模型的初心，但是解释变量之间存在多重共线性可能会导致一些回归系数通不过显著性检验，回归系数的正负号出现倒置的情况，即参数估计量的经济含义不合理，违背了建立回归方程的初心。因此当同时出现违背基本假设的多种情况下，优先需要解决解释变量之间的多重共线性问题，使得建立的回归方程有实际的应用价值，正确的经济意义和能正常的进行更深一步的分析。

当一般多重共线性不是过分严重时，是不需要进行处理的，通过调整变量即可，此时关于异方差性和自相关性的优先检测顺序则需要根据原始数据类型来选择。因为异方差性出现的原因是截面数据中总体各单位的差距，而自相关一般出现在有关时间系列数据之中即经济系统的惯性，经济活动的滞后效应和蛛网现象等[4]。通过对教材和其他资料的学习，本文认为如果是截面数据应当优先检验异方差性，如果是时间系列数据应当优先检验是否存在自相关性。

3. 案例分析

财政收入按收入形式可以分为：各项税收收入、企业收入、债务收入、国家能源交通重点建设基金收入、基本建设贷款归还收入、国家预算调节基金收入、其他收入等。从定性分析的角度来说，财政收入会受到各种不同因素的影响，如：农业增加值、工业增加值、建筑业增加值、社会总人口数、社会消费额总额、国土受灾面积等等[5]。本文建立模型仅选取我国农业增加值、第二产业增加值(包括工业和建筑业)、第三产业增加值、社会从业人数，以及其他收入水平 5 个因素为解释变量，分析它们对财政收入的影响程度[6] [7]。

3.1. 提出因变量与自变量

y 表示财政收入(亿元)为因变量；五个解释变量：x_1 表示农业增加值(亿元)，x_2 表示第二产业增加值

(亿元)，x_3表示第三产业增加值(亿元)，x_4表示社会从业人数(亿人)，x_5表示其他收入水平(亿元)。(数据见表 1，来源于《中国统计年鉴 2021》)。

Table 1. Part of China statistical yearbook from 2005 to 2020
表 1. 2005~2020 年部分中国统计年鉴表

年份	财政收入(亿元)	农业增加值(亿元)	第二产业增加值(亿元)	第三产业增加值(亿元)	社会从业人数(亿人)	其他收入水平(亿元)
2020	182913.88	77754.1	384255.3	553976.8	7.5064	204145.2
2019	190390.08	70473.6	380670.6	535371.0	7.5447	179086.3
2018	183359.84	64745.2	364835.2	489700.8	7.5782	156744.3
2017	172592.77	62099.5	364835.2	438355.9	7.6058	136856.5
2016	159604.97	60139.2	331580.5	390828.1	7.6245	119618.5
2015	152269.23	57774.6	295427.8	349744.7	7.6320	105847.3
2014	140370.03	55626.3	281338.9	310654.0	7.6349	93041.6
2013	129209.64	53028.1	277282.8	277983.5	7.9301	81082.2
2012	117253.52	49084.6	261951.6	244856.2	7.8894	78564.3
2011	103874.43	44781.5	244639.1	216123.6	7.8579	75324.8
2010	83101.51	38430.8	227035.1	182061.9	7.8388	69452.1
2009	68518.30	33583.8	191626.5	154765.1	7.7510	58453.3
2008	61330.35	32464.1	160168.8	136827.5	7.7046	52651.8
2007	51321.78	27674.1	149952.9	115787.7	7.6531	49513.8
2006	38760.20	23317.0	126630.5	917662.2	7.6315	46581.7
2005	31649.29	21806.7	104359.2	77430.0	7.6120	41568.2

3.2. 作相关分析，设定理论模型

利用 Python 软件计算增广相关阵(见表 2)：从相关阵看出，所选自变量与 y 有一定的线性相关，用 y 与自变量作多元线性回归是合适的，因此可以设定理论模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \tag{1}$$

Table 2. Correlation between variables
表 2. 各变量之间相关系数表

单位：亿元	财政收入	农业增加值	第二产业增加值	第三产业增加值	社会从业人数	其他收入水平
财政收入	1.0	0.4232	0.2109	0.2755	−0.8203	0.9281
农业增加值	0.4232	1.0	−0.4609	−0.5988	−0.2542	0.6996
第二产业增加值	0.2109	−0.4609	1.0	0.7278	0.04857	−0.0316
第三产业增加值	0.2755	−0.5988	0.7278	1.0	−0.0737	−0.0908
社会从业人数	−0.8203	−0.2543	0.0486	−0.0736	1.0	−0.7699
其他收入水平	0.9281	0.6996	−0.0316	−0.0908	−0.7698	1.0

3.3. 计算结果

利用 python 软件进行计算，代码如下：

```
x = sm.add_constant(data2.iloc[:,1:])
y = data2.iloc[:, 0]
model = sm.OLS(y, x)
result = model.fit()
result.summary()
```

根据输出结果，可得：

1) 决定系数 $R_2 = 0.999$，看出回归方程高度显著，

2) 方差分析表，F = 1024，P 值 = 0.000975，说明回归方程高度显著，自变量整体上对 y 有高度显著的线性影响，

3) 回归结果为 $R^2 = 0.999$，F = 1024 回归方程为：

$$\hat{y} = 2.078e^{+05} - 0.6305x_1 - 0.0492x_2 + 0.7156x_3 - 1.731e^{+04}x_4 + 0.7428x_5$$

这里 x_1 的系数为负，显然是不符合理论常识，不具有正确的经济意义，认为可能是由于自变量之间的多重共线性导致，所以为了使之后的研究有正确的实际意义先进行多重共线性检验。

3.4. 多重共线性的诊断与处理

见上表，x_1 与 x_5 的简单相关系数为 0.699；x_2 与 x_3 的简单相关系数为 0.728.解释变量间存在一定的相关关系。所以本文运用方差扩大因子法，用 Python 诊断，代码如下：

```
x = sm.add_constant(data2.iloc[:,2:])
y = data2.iloc[:, 1]
model = sm.OLS(y, x) ; result = model.fit() ; result.summary()
x = sm.add_constant(data2.iloc[:,[1, 3, 4, 5]])
y = data2.iloc[:, 2]
model = sm.OLS(y, x) ; result = model.fit() ; result.summary()
x = sm.add_constant(data2.iloc[:,[1, 2, 4, 5]])y = data2.iloc[:, 3]
model = sm.OLS(y, x) #生成模型
result = model.fit() #模型拟合
```

观察所有解释变量的方差扩大因子，发现 x_1 与 x_5 的大小大于 10，说明回归方程存在严重的多重共线性。粗略判定 x_1 与 x_5 之间存在较强的共线性。x_5 的 VIF5 = 19.6078 在所有方差扩大因子之间最大，所以剔除 x_5。再用 python 诊断，所得结果如下表 3：

Table 3. Variance expansion factor table
表 3. 方差扩大因子表

	VIF1	VIF2	VIF3	VIF4	VIF5
所有解释变量	12.6582	3.0033	3.2679	7.8125	19.6078
剔除部分解释变量	1.8116	2.1739	2.8986	1.1933	

从输出结果看，四个方差扩大因子都小于 10，回归系数也有合理的解释，说明此回归模型不存在强的多重共线性，回归方程为：

$$\hat{y} = 9.282e^{+05} + 7.1627x_1 + 0.3608x_2 + 0.9838x_3 - 1.093e^{+05}x_4 \quad (2)$$

其 $R^2 = 0.956$，$F = 16.45$。

3.5. 异方差诊断与处理

由于实际问题存在错综复杂的原因，因此在建立实际问题的回归分析模型时，经常会出现某一因素或某些因素随着解释变量观测值的变化而对被解释变量产生不同的影响，导致随机误差项产生不同方差即异方差性。异方差性出现的原因主要为以下三点：第一，模型设定误差；第二，数据的测量误差；第三，截面数据中对总体各单位的差异。当出现异方差性时参数估计式仍然具有线性性，无偏性和一致性，却不再具有最小方差性。异方差性也会使 t 统计量值变小，而且在异方差的情况下，通常由 OLS 法得到的 t 统计量不再服从 t 分布，F 统计量也不再服从 F 分布[8]。因此 t 检验和 F 检验失去存在的基础同时会扩大估计区间和预测区间，降低精度。

检验是否存在异方差性通常用图示检验法，Goldfeld-Quanadt 检验，White 检验，ARCH 检验和 Glejser 检验。本文采用等级相关系数法检验，计算随机误差项的绝对值与自变量之间的等级系数。从表 4 计算结果看出，在 0.05 的显著性水平下异方差性不明显，所以不用进行相关处理。

Table 4. Test table of rank correlation coefficient of each variable
表 4. 各变量等级相关系数检验表

	ABSE	农业增加值	第二产业增加值	第三产业增加值	社会从业人数
ABSE	1.000	0.190	0.190	0.256	0.287
农业增加值	0.190	1.000	0.999	0.814	0.764
第二产业增加值	0.190	0.999	1.000	0.815	0.825
第三产业增加值	0.256	0.814	0.815	1.000	0.778
社会从业人数	0.287	0.764	0.825	0.778	1.000

3.6. 自相关性的诊断与处理

1) DW 检验

DW 检验用于检验随机误差项具有一阶自回归形式的序列相关问题，也是就自相关检验。根据公式 DW = 2 (1 − P)计算 DW 的值，显著性水平 α，同时根据 DW 检验决策规则判断自相关状态。DW 检验法适用于解释变量 X 为非随机的小样本，并且只能用于检验随机误差项具有一阶自回归形式的自相关问题。

由 Python 计算得到 DW = 2.541，可以看出残差序列存在负自相关，代码如下所示：

```
x = sm.add_constant(data2.iloc[:,[1, 2, 3,4]])
y = data2.iloc[:, 0]
model = sm.OLS(y, x) #生成模型 ; result = model.fit() #模型拟合 ; result.summary()#模
result.summary() #由表格可以看出 DW 值为 2.541
```

2) 迭代法消除自相关

设此时回归模型为：$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \varepsilon_t$

误差项存在一阶自相关：$\varepsilon_t = p\varepsilon_{t-1} + u_t$。

且：$E(u_t)=0,\ t=1,2,\cdots,8$，

$$Cov(u_t,u_s)=\begin{cases}\sigma^2 & t=s\\ 0 & t\neq s\end{cases}\quad (t,s=1,2,\cdots,8)$$

则：$y_{t-1}=\beta_0+\beta_1x_{t-1,1}+\beta_2x_{t-1,2}+\beta_3x_{t-1,3}+\beta_4x_{t-1,4}+\varepsilon_{t-1}$

$$y_t-py_{t-1}=\beta_0-p\beta_0+\beta_1(x_{t1}-px_{t-1})+\beta_2(x_{t2}-px_{t-1,2})$$
$$+\beta_3(x_{t3}-px_{t-1,3})+\beta_4(x_{t4}-px_{t-1,4})+\varepsilon_t-p\varepsilon_{t-1}$$

令：$$\begin{cases}y_t=y_t-py_{t-1} & \beta_0=\beta_0-p\beta_0\\ x_{t1}=x_{t1}-px_{t-1,1} & \beta_1=\beta_1\\ x_{t2}=x_{t2}-px_{t-1,2} & \beta_2=\beta_2\\ x_{t3}=x_{t3}-px_{t-1,3} & \beta_3=\beta_3\end{cases}\qquad u_t=\varepsilon_t-p\varepsilon_{t-1}$$

即：$y_t=\beta_0+\beta_1x_{t1}+\beta_2x_{t2}+\beta_3x_{t3}+\beta_4x_{t4}+u_t$

然后，用 Python 计算输出结果，得到新回归残差的 DW 为 1.768，查表，n = 7，k = 5，显著性水平为 α = 0.05，得 $d_L=0.56$，$d_U=2.21$，DW 检验仍然落在不确定区域。但一步迭代得误差项的标准差小于原来的标准差，所以进一步迭代。重复上述过程，再用 Python 输出结果得 DW = 2.379，查表得 DW 检验基本落入无自相关区。且进一步迭代的误差项小于一步迭代的误差项，所以最后还原的原始方程为：

$$\begin{aligned}&\hat{y}_t+0.2705y_{t-1}-0.116y_{t-2}=9.4386e^{+05}+(7.1627x_{t,1}+0.8994x_{t-1,1}-0.2797x_{t-2,1})\\&+(0.3608x_{t,2}+0.3688x_{t-1,2}-0.0317x_{t-2,2})+(0.9838x_{t,3}+0.2410x_{t-1,3}-0.0752x_{t-2,3})\\&-(1.093e^{+05}x_{t,4}+1.395e^{+05}x_{t-1,4}-0.289e^{+05}x_{t-2,4})\end{aligned}\tag{3}$$

4. 结论

综上所述，本文介绍了一种处理回归分析中同时出现多种回归异常时的一般方法，并以一个实例来验证这个方法和技巧的具体过程。本文认为当同时出现违背基本假设的多种情况下，优先需要解决解释变量之间的多重共线性问题，需要使得建立的回归方程有实际的应用价值，正确的经济意义和能正常的进行更深一步的分析。而当一般多重共线性不是过分严重时，不需要进行处理，通过调整变量即可，此时关于异方差性和自相关性的优先检测顺序需要根据原始数据类型来选择[9]。因为异方差性出现的原因是截面数据中总体各单位的差距，而自相关一般出现在有关时间系列数据之中即经济系统的惯性，经济活动的滞后效应和蛛网现象等，所以当数据为截面数据时的处理顺序是多重共线性－异方差性－自相关性，当数据为时间系列数据时的处理顺序为多重共线性－自相关性－异方差性。实际上，为了便于解释相关理论结果，本文所举的例子是选取一个比较简单的数据结构进行回归分析，而在实际应用中将会有更复杂的数据结构出现，本文观点也可在其回归分析中进行论证，而本文对于对比论证就不做过多论述。

参考文献

[1] 何晓群, 刘文卿. 应用回归分析[M]. 北京: 中国人民大学出版社, 2001.

[2] 林乐义. 岭回归在消除多重共线性中的应用[J]. 辽东学院学报, 2020(2): 274-278.

[3] 庞皓. 计量经济学[M]. 第三版. 北京: 科学出版社, 2018.

[4] 徐生霞, 潘海涛. 截面数据异方差问题检验技术的比较[J]. 统计与决策, 2017(5): 16-20.

[5] [美]罗伯特 S.平狄克, 平代克, 鲁宾费尔德. 计量经济模型与经济预测[M]. 钱小军, 译. 北京: 机械工业出版社, 1999.

[6] 白萍. 影响我国财政收入的多元线性回归模型[J]. 统计与决策, 2005(10): 92-94.

[7] 赵慧江. 基于回归分析的粮食产量影响因素分析[J]. 内蒙古农业科技, 2009(2): 31-35.

[8] 王义闹, 卢庆华. 关于多重共线性的三个知识点的准确表述[J]. 温州大学学报, 2019, 40(3): 7-12.

[9] Rathowsky, D. (1990) Handbook of Nonlinear Regression Models. Marcel Dekker, New York and Basel.