学校代码 **10125** 专业代码 **071400**

山西财经大学

硕 士 学 位 论 文

题目 线性回归模型中的异方差检验与估计方法研究

姓　　名 李阔辰
专　　业 统计学
研究方向 应用数理统计
所属学院 统计学院
指导教师 张晓琴

二〇二三年五月十四日

University code 10125 **Major Code** 071400

Shanxi University of Finance & Economics

Thesis for Master's Degree

Title **A Study of Heteroscedasticity Test and Estimation Method in Linear Regression Model**

Name Li Kuochen

Major Statistics

Research Orientation Apply mathematical statistics

School School of Statistics

Tutor Zhang Xiaoqin

May 14, 2023

摘　要

经典的线性回归模型中假设随机误差项具有相同的方差，然而在很多实际情况中模型随机扰动项的方差是不完全相等的，即模型具有异方差性。在此情形下，使用普通最小二乘法计算模型参数，得到的估计量将不具备有效性，变量的系数显著性检验失去意义，模型的预测也将失去意义。因此，使用合理的异方差检验方法验证回归模型是否存在异方差，并对参数向量的协方差矩阵进行更准确的估计，使得对回归参数的假设检验更加精确是非常重要的。

首先，本文提出 M-G-Q 检验。在线性回归模型中，G-Q 检验无法直接应用于多元模型，很多学者提出了改进的 G-Q 检验使其可以应用于多元线性回归模型，而这些改进的 G-Q 检验存在着检验步骤较为繁琐和准确率不足的问题。对此本文基于变量选择的思想，根据显著性检验选择出对回归扰动项方差影响最大的解释变量，然后通过该解释变量对样本观测值进行由大到小排序，完成整体的 G-Q 检验。通过数值模拟发现：相较于已有的改进的 G-Q 检验，M-G-Q 检验步骤较为简便，且相比其他学者对 G-Q 检验的改进，M-G-Q 检验的准确率更高。通过实例分析也表明了 M-G-Q 检验的可行性。

其次，提出 HCCv 估计。自 White 证明了 HC0 为异方差一致协方差阵估计量以来，很多学者针对其收敛慢、样本需求量大的问题提出了改进，这些方法统称为异方差一致协方差阵估计量（HCCMEs），已有的 HCCMEs 方法有 HC0，HC1，HC2，HC3，HC4，HC5，HC4m，HC5m 和 HC6，这些估计量都是对参数协方差矩阵的一致估计。本文在已有的 HCCMEs 的基础上，针对其适用范围不广和准确率不足的问题提出了一个新的估计量 HCCv，新估计量在多种异方差情况下的估计效果更加精确，本文通过大量的模拟实验证明了该结论，并通过实例分析验证了其可行性。

最后，总结了全文的创新与不足，并对未来的研究方向进行了展望。

关键词：线性回归模型；异方差；G-Q 检验；异方差一致性协方差矩阵估计

ABSTRACT

The classical linear regression model assumes that the random error terms have the same variance, however, in many practical situations the variances of the random perturbation terms of the model are not exactly equal, i.e., the model has heteroskedasticity. In this case, using ordinary least squares to calculate the model parameters, the estimates obtained will not have validity, the significance test of the coefficients of the variables loses its meaning, and the prediction of the model will also lose its meaning. Therefore, it is very important to use a reasonable heteroskedasticity test to verify the existence of heteroskedasticity in the regression model and to estimate the covariance matrix of the parameter vector more accurately to make the hypothesis testing of the regression parameters more precise.

First, the M-G-Q test is proposed in this paper. In the linear regression model, the G-Q test cannot be directly applied to the multivariate model, and many scholars have proposed improved G-Q tests so that they can be applied to the multivariate linear regression model, but these improved G-Q tests have the problems of cumbersome test steps and insufficient accuracy. In this paper, based on the idea of variable selection, the explanatory variable that has the greatest influence on the variance of the regression disturbance term is selected based on the significance test, and then the overall G-Q test is completed by sorting the sample observations by this explanatory variable from largest to smallest. The numerical simulation shows that the M-G-Q test is easier than the existing improved G-Q test, and the accuracy of the M-G-Q test is higher compared to other scholars' improvements of the G-Q test. The feasibility of the M-G-Q test is also demonstrated by example analysis.

Second, the HCCv estimation is proposed. Since White proved HC0 as the heteroskedasticity consistent covariance array estimator, many scholars have proposed improvements for its slow convergence and large sample requirement, and these methods are collectively called heteroskedasticity consistent covariance array estimators

(HCCMEs), and the existing HCCMEs methods are HC0, HC1, HC2, HC3, HC4, HC5, HC4m, HC5m, and HC6, all of which are consistent estimates of the parameter covariance matrix. In this paper, based on the existing HCCMEs, a new estimator HCCv is proposed to address the problems of its poor applicability and accuracy. The new estimator is more accurate in a variety of heteroskedasticity cases, and the paper proves the conclusion through a large number of simulation experiments and verifies its feasibility by example analysis.

Finally, the innovations and shortcomings of the whole paper are summarized, and the future research directions are prospected.

Keywords: linear regression model; heteroskedasticity; G-Q test; heteroskedasticity consistency covariance matrix estimation

目　录

第 1 章 绪 论

在线性回归模型中，通常使用普通最小二乘法（Ordinary Least Squares, 简称为 OLS）来估计未知参数。经典线性回归模型中有一个非常重要的基本假设，即随机扰动的方差是相等的。若不满足这个基本假设，则随机扰动项具有不同的方差，该回归模型为异方差模型，此时，若仍采用 OLS 对未知参数进行估计，则得到的参数估计量将不具有有效性，从而导致参数的显著性检验失去意义。所以在构建线性回归模型之前，对模型进行异方差检验非常重要。当回归模型为异方差模型时，可以使用适当的方法对回归参数来进行估计，从而构建线性回归模型。传统的异方差检验方法和回归参数的协方差矩阵估计方法有很多，但有些方法存在一定的局限性，因此分析这些异方差检验方法和回归参数的协方差矩阵估计方法的适用性有着非常重要的意义。

1.1 研究背景与意义

1.1.1 研究背景

线性回归模型在统计学中是最常见的模型[1]，通过对参数的线性模型组合来对自变量与因变量之间的关系进行研究。只有一个因变量和自变量的模型称为一元线性回归模型，有多个自变量的模型称为多元线性回归模型。无论是一元线性回归模型还是多元线性回归模型，都可以通过对模型参数进行估计并对其进行假设检验进而完成模型的构建。线性回归模型在实际生活中有着广泛的应用范围[2]，可以利用已知的数据，通过对模型中的未知参数进行估计，进而对未知的被解释变量进行预测，这已经被广泛应用在现实研究之中，例如任丹[3]在论文中利用多元线性回归模型建立了电影票房的预测模型，可以帮助人们做出更合理的决策，蔡成旺[4]等通过对各项煤指标进行线性回归分析，从而实现对焦炭相关质量指标的预测，可以用于指导实际生产。相较于非线性模型，线性回归模型具有步骤简单、结果易于解释、准确度高等优点，因此在经济学、社会学、生物学、医学等各个科学领域中广泛应用[5-8]，有着难以取代的地位。

经典的线性回归模型基于一个基本假定：模型中各个随机扰动项的方差是相等的，即同方差性。只有在同方差的条件下，使用 OLS 对回归参数协方差矩阵的估计才是无偏且一致的。然而，在实际情况中，有很多情形与上述的同方差情形是不同的[9]。比如，一家商场的规模越大，其日营业额就越大，其日营业额的波动也就越大。若以方差来刻画，则会发现它们的方差存在着较大的差别，这与经典线性回归模型中的同方差假设相悖。当线性回归模型不满足同方差性假定时，该模型就具有异方差性。在经济学的研究中，异方差现象往往出现在截面数据中[10-12]。截面数据是不同的主体在同一时间点或者同一时间段内的数据，这些主体的规模和水平不完全相同，因此容易产生不同的差异。例如，在对不同的工厂的产出和劳动力之间的关系的研究中，由于各个工厂的规模大小不相同，其劳动力也不相同，所以导致产出的差异大小不相同。劳动力数量较少的工厂，其产出的差异也较小，劳动力数量较多的工厂，其产出的差异也较大。在研究不同家庭中支出与收入之间的关系时，高收入家庭的支出变化程度大于低收入家庭。在经济领域中，这样的异方差现象很常见。当回归模型的随机扰动项不满足同方差假定时，称该回归模型具有异方差性[13]。此时，各个随机扰动项的方差不完全相同，若仍使用 OLS 对模型进行估计，则对回归参数的估计将不再有效，对回归参数的显著性检验也会失去意义，从而可能降低预测的准确性，甚至导致做出错误的判断，产生严重的影响[14]。

因此，为了得到一个更为准确的回归参数的方差估计量，在构建线性回归模型之前有必要对回归模型进行异方差检验，若模型具有异方差性，应采用适当的方法对模型中的回归参数的协方差矩阵进行估计，从而得到更准确的回归模型，进而提高模型的预测精度。所以，如何发现并解决线性回归模型中的异方差问题是建立线性回归模型过程中十分关键的问题。

1.1.2 研究意义

（1）理论意义

在线性回归模型中，通常采用 OLS 对回归参数来进行估计和假设检验。在同方差的条件下，使用 OLS 对回归模型进行参数估计、假设检验等一系列统计推断是合理的。当模型存在异方差时，如果仍采用 OLS 方法，则对回归参数协方差矩

阵的估计将不再一致，会造成对参数的检验无效的情况，从而可能做出错误的推断，甚至产生严重的后果。因此，为取得较好的估计，必须先对模型进行异方差检验，并且当模型存在异方差性时，通过合适的方法来消除异方差所带来的影响。本文将从异方差检验和参数协方差矩阵估计两方面展开讨论，提出更有效的异方差检验方法和参数协方差矩阵估计方法，这对回归模型的理论研究具有积极的推动作用。

（2）应用价值

线性回归模型在经济学、社会学、生物学、医学等各个科学领域中广泛应用，而在现实数据中，往往存在着大量的异方差数据。在这样的情况下，若仍使用 OLS 对回归参数的检验就无效了，从而降低了被解释变量的预测精度，更为严重可能导致其预测失去意义。提高异方差检验方法的精确性，并且通过恰当的方法消除异方差对模型造成的影响，可以提高预测的准确性，有利于更准确地指导经济决策，对实际生活中经济问题的相关研究具有十分重要的意义[15]。所以如何发现异方差，并且消除异方差对模型的影响是线性回归模型中的一个至关重要的问题。故本文将针对异方差的检验和存在异方差情况下对参数协方差矩阵的估计问题展开讨论。

1.2 国内外研究现状

在建立线性回归模型的过程中，对异方差性进行检验以及对回归参数协方差矩阵进行估计十分重要，国内外有很多学者在这两方面进行了大量的研究，提出了很多经典的检验和估计方法，还有很多学者对这些经典方法进行了很多改进。下面将对异方差检验方法和估计方法两方面的研究现状进行介绍。

1.2.1 异方差检验方法

（1）经典的异方差检验方法

为了解决回归模型中的异方差对模型带来的问题，国内外学者对此进行了很多相关的研究，取得了丰硕的成果。在异方差检验方面，已经有了很多经典的异方差检验方法，例如图示法、Park 检验、Glejser 检验、Breusch-Pagan 检验（B-P 检验）、Goldfeld-Quandt 检验（G-Q 检验）和 White 检验等。

①图示检验法

图示检验法可以通过绘制被解释变量与解释变量的散点图来大致判断是否存在异方差情况，也可以先对被解释变量与解释变量进行回归，然后绘制解释变量与残差平方的散点图来进行判断。具体判断方法如下：使用解释变量与被解释变量绘制散点图时，如果被解释变量的波动幅度随着解释变量的增加而逐渐增大或者逐渐减小时，可以粗略地认为模型存在异方差。使用解释变量与残差平方绘制散点图时，如果残差平方有随着解释变量的增加而逐渐增大或缩小的趋势，可以粗略地认为模型存在异方差。图示检验法的优点是可以较为直观地对模型的异方差性进行观察。然而，图示检验法的适用范围仅限于一元回归模型，且精度较低，仅能判断出较为明显的异方差情况。

②Park 检验[16]

Park（1966）通过使用残差平方的对数与解释变量的对数进行辅助回归，根据回归参数的显著性来判断判断回归扰动项的方差与解释变量之间是否存在较强的相关关系。在同方差的情况下，用残差平方的对数与解释变量的对数建立的回归模型应该不显著。在一元线性回归模型中使用 Park 检验能够快速判断模型是否存在异方差性，但在多元线性回归模型中，需要多次进行 Park 检验，其过程较为繁琐，且变量间的相关性可能会影响检验结果的准确性。

③Glejser 检验[17]

Glejser（1969）通过对残差的绝对值与解释变量的某种函数形式做回归，若回归参数显著，则认为模型存在异方差，若回归参数不显著，则认为模型不存在异方差。该检验在发现异方差的同时可以同时发现异方差的具体形式，但该方法与 Park 检验有着类似的问题，即在多元回归模型中需要构造多个残差绝对值与各个解释变量的某种函数形式的回归模型并进行多次 Glejser 检验，过程较为繁琐且结果准确度较低。

④B-P 检验[18]

Breusch 和 Pangan（1979）假设随机扰动项方差为各解释变量的线性函数，通过构造随机扰动项与各解释变量的线性回归模型并对回归系数进行检验。Breusch 和 Pagan 证明了模型的残差平方和渐近服从卡方分布，通过对回归系数的显著性检验来判断模型是否具有异方差性。B-P 检验法适用于样本容量较大的情况，该

检验不适用于小样本检验。

⑤G-Q 检验[19]

Goldfeld 和 Quandt（1965）提出了先将样本按照解释变量由小到大进行排序并分为三个样本组，去掉排在中间的样本组，然后使用剩下的两个样本组构造 F 统计量，进而通过比较该统计量与 F 统计量的临界值来判断数据是否存在异方差性。G-Q 检验需要将样本按照解释变量的大小进行排序，故该方法仅适用于一元线性回归模型，无法直接应用于多元的情况。

⑥White 检验[20]

White（1980）证明了参数协方差矩阵的一致估计量，并由此得到一个服从卡方分布的统计量，进而判断模型是否具有异方差性。White 使用模型的残差平方来代替随机扰动项的方差来对解释变量及其平方项和交叉项乘积做回归，并证明了这样不影响统计量的分布。White 检验使用条件较为宽松，不依赖于正态假设，但由于其需要使用解释变量的平方以及它们的交叉项乘积来构造辅助回归模型，造成了很多自由度的损失，所以该检验方法需要较大的样本量才能取得较为准确的结果，不适合在样本较小的情况中使用。

（2）改进的 G-Q 检验方法

G-Q 检验是异方差检验中常用的方法，它有着方便快捷的优点，但也存在一定的局限性，其无法直接应用于多元线性回归模型，为了解决这一问题，近年来有很多学者对 G-Q 检验做出了改进，通过寻找一个合适的排序标准，从而使 G-Q 检验可以应用于多元线性回归模型中。

龚秀芳[21]（2005）提出了一种改进的 G-Q 检验，使其可以应于多元线性回归模型的异方差检验问题。这种方法的思想是找出一个变量来尽可能多地反映多个解释变量的信息，从而可以根据该变量对样本进行排序，然后再进行 G-Q 检验。龚秀芳提出若样本的第一主成分的贡献率达到 70%以上，则可以使用样本的第一主成分来反映多个解释变量的信息，然后将样本按照第一主成分排序，继而可以进行推广的 G-Q 检验。

金蛟[22]（2008）基于统计深度函数对 G-Q 检验进行了改进，提高了其适用范围，使其可应用于多元回归模型。金蛟指出使用第一主成分来对样本进行排序可能会使样本信息有较多的损失，尤其在第一主成分的贡献率较低的情况下，统计

深度函数可以作为 G-Q 检验中数据排序的有效工具，金蛟提出使用统计深度函数对样本来进行排序，可以解决使用第一主成分可能导致的信息损失的问题，从而得到更为准确的异方差 G-Q 检验，但该方法计算过程较为复杂。

郑红艳[23]（2010）等提出将多元线性回归模型的 G-Q 检验转化为各个解释变量和被解释变量的一元线性回归模型的 G-Q 检验，如果所有的一元线性回归模型都不存在异方差，则可以认为这个多元线性回归模型不存在异方差，如果有一个一元线性回归模型出现异方差，则可以认为此多元线性回归模型存在异方差，这样就可以得出一种适用于多元线性回归模型的 G-Q 检验。该方法将多元线性回归模型的 G-Q 检验分解为多个一元线性回归模型的 G-Q 检验，计算十分繁琐，而且存在着误检率较高的问题。

李俊领[24]（2012）将传统 G-Q 检验中去除样本序列中间 $c = n/4$ 个观察值改为去除样本序列中间 $c = n/7$ 个观察值，从而提出了一种修正的 G-Q 检验，并通过实例分析论证了其方法在样本容量较小时具有较好的效果。

刘明[25]（2018）等在基于随机干扰项的方差和解释变量具有线性关系的假设下，提出以被解释变量的拟合值的大小为标准对样本点进行排序，使得 G-Q 检验可以适用于多元回归模型，该方法体现了每个解释变量的作用，提出了可应用于多元线性回归模型且实施较为简便的改进的 G-Q 检验，但该方法仅适用于异方差的变动方向与各个解释变量的变动方向相同或相反的情况，如果各个解释变量变动方向不完全相同或相反，则该检验方法的效果会明显降低。

张晓琴[26]（2019）等将 G-Q 检验的思想与非参数 Kolmogorov-Smirnov 检验（K-S 检验）结合，提出了基于 G-Q 检验的 K-S 检验方法。该方法通过将样本随机分为数量相等的两部分，分别进行回归后利用残差平方和构造出类似 G-Q 检验中的 F 统计量，该过程重复进行多次，可得到多个 F 统计量，然后运用 K-S 检验，比较由样本观测值所形成的经验分布与对应自由度的 F 分布是否具有显著差异，从而判断模型是否具有异方差。该方法将 G-Q 检验与 K-S 检验结合，可以运用于多元线性回归，大大提高了检验的适用范围和检验效果，但计算量偏大，检验步骤较为复杂。

（3）其他的异方差检验方法

为了提高异方差检验的准确率，近年来也有很多学者提出了新的改进的异方

差检验方法。

彭作祥[27]（2003）等提出了基于极值指数估计量的的异方差检验方法，但其要求顺序统计量满足独立同分布，在实际问题中有一定的局限性。兰嘉庆[28]（2004）等基于游程检验提出了一种新的异方差检验方法，该方法不需要参数分布，且需要的样本量较小，但该检验过程中要对数据进行转化，会导致数据信息量的损失。张荷观[29]（2006）等提出利用分组的方法获得重复数据，由此对随机扰动项的方差进行估计和检验，这种检验方法可以适用于各种形式的异方差，且计算量较小，但在数据分组的过程中可能会掩盖某种类型的异方差从而使检验的精度降低。夏帆[30]（2012）等基于分布拟合提出了一种异方差检验方法，该方法不依赖随机扰动项的方差与解释变量的关联，通过对比样本数据分布与同方差条件下的 F 分布的差异对模型的异方差性进行检验，但该方法计算量较大，计算过程较为繁琐。唐裔[31]（2018）等利用样本主成分的方法对 Glejser 检验进行了改进，使用样本主成分重新组合观测值，既能检验出异方差的具体形式，又减小了计算量，提高了检验效率。谭馨[32]（2019）等运用主成分的思想，根据系数的显著性检验提出一种改进的 Park 检验，提高了 Park 检验的检验效率。刘锋[33]（2019）等基于经验似然方法，构造出渐近服从卡方分布的似然比统计量，并依此进行异方差检验。

1.2.2 协方差矩阵估计方法

（1）异方差一致性协方差矩阵估计方法

1980 年，White[20] 证明了异方差-一致性协方差矩阵（Heteroskedasticity-Consis- tent Covariance Matrix Estimator，简记为 HCCMEs）是 OLS 估计量的渐近协方差矩阵的一致估计量，无论在同方差模型还是异方差模型中，该方法均可以使用。在样本量足够大的条件下，即使在异方差模型下，使用 HCCMEs 也能得到一致的参数协方差矩阵估计量，进而进行参数的假设检验。

记 White 提出的协方差矩阵估计量为 HC0，当样本量足够大时，使用 HC0 可以做出准确的参数检验。然而在现实应用中，样本容量经常不够大，导致 HC0 存在较大误差，从而影响参数检验的准确性。为此，很多学者对 HC0 做了修正。Hinkley[34]提出了用自由度进行修正的估计量 HC1，Horn 和 Duncan[35]提出了用投

影矩阵的对角元素进行修正的估计量 HC2，MacKinnon 和 White[36]提出了用刀切法对参数的协方差矩阵进行估计，称为 HC3，且估计的效果优于 HC1 和 HC2。Davidson 和 MacKinnon[37]提出了便于计算且效果与刀切法近似的 HC3 估计量。Cribari-Neto 先后提出了 HC4[38]、HC5[39]和 HC4m[40]估计量。并论证了 HC4 估计量在在存在高杠杆点的情况下表现优于 HC3，HC5 估计量在存在高杠杆点的情况下表现优于 HC4，而 HC4m 估计量在不存在高杠杆点时表现最好。李顺勇[41]等提出了 HC5m 估计量，Nuzhat Aftab 和 Sohail Chand[42]提出了 HC6 估计量，他们都论证了其提出的估计量在特定情况下的估计效果优于之前的其他估计量。

（2）基于正交表的异方差估计方法

一些学者提出了利用正交表方法对线性回归模型中的参数协方差矩阵进行估计，并获得了更为准确的估计。

张晓琴[43]（2015）等基于正交表的优良性质，考虑到每个变量的影响，对因变量选择和容差的选取进行了修正，从而提出了 OR 估计。张晓琴[44]（2016）等针对现有的基于正交表的估计方法中，确定因变量和自变量之间的关系时未充分利用原始数据信息的问题，以及在应用正交表时，对于容差的确定存在的不合理之处，对 OR 估计进行了改进，得出了 OR1 估计，从而提高了估计的精度。冯军芳[45]（2021）等在正交表 OR 估计的基础上，增加了样本的相对误差作为容差的选取标准，使容差与样本误差相对应，从而提出了基于正交表的 OR2 估计，并与传统的两阶段最小二乘法相结合，解决了分组选择困难和可能产生的部分样本信息丢失的问题。郭雅静[46]（2022）等将非参数估计与正交表估计相结合，提出了 M-OR1 估计，其在异方差形式未知的情形下，首先采用正交列表对模型进行扩展，再利用非参数方法对各项的方差进行估计，从而获得了更为精确的结果。

（3）其他的参数协方差矩阵估计方法

为了提高异方差模型下参数协方差矩阵估计的准确率，近年来也有很多学者提出了其他的异方差模型下协方差矩阵的估计方法。

广义最小二乘法[47]（GLS）是一种常见而有效的方法，它的基本思想是通过对变量进行转换，使转换后模型的随机扰动项不存在自相关性和异方差性，进而利用 OLS 对回归参数进行估计。Box-Cox 变换法[48]通过对模型因变量做特定变换来获得恰当的参数，使变换后的因变量服从正态分布，从而满足经典线性回归模

型的各项假设，然后再利用 OLS 对模型进行估计。YAN[49]（2003）等运用贝叶斯估计方法，对被解释变量的均值和方差都服从正态分布时的情况下，提出了一种新的异方差模型中的参数估计方法。Leslie[50]（2007）等对此方法进行了进一步的拓展，并通过贝叶斯方法和混合先验狄氏过程得出了一个适用性更广的的异方差模型中的参数估计方法。张晓琴[51]（2020）等将 N-W 估计方法引入异方差模型参数估计中，然后利用 GLS 方法对未知参数进行估计，得到一种新的异方差估计方法，称为 Kernel Nadaraya Watson（KNW）估计。李顺勇[52]（2022）等针对 KNW 使用固定窗宽造成的估计误差偏大问题，与自适应 N-W 核回归估计的思想结合，在选取窗宽时引入了可变窗宽，提出了 AKNW 估计，提升了估计方法的普适性与准确性。张晓琴[53]（2021）等在异方差模型的估计中引入了局部多项式的非参数方法，提出了一种新的 LP 估计方法。

1.3 本文研究内容

本文首先介绍了一些已有的异方差检验和估计方法，指出了其优点和不足，然后从异方差的检验和估计两方面展开了研究，主要分为改进的 G-Q 异方差检验和改进的 HCCMEs 估计两部分研究内容。

在第二章中，介绍了传统的 G-Q 检验和其他学者对其的改进，以及 HCCMEs 和其他学者对其的改进。

在第三章中，提出了一种改进的 G-Q 异方差检验，并通过大量的数值模拟与其他改进的 G-Q 检验进行比较，论证了其良好的检验效果，然后通过实例分析论证了新检验方法的可行性。

在第四章中，提出了一种新的 HCCMEs 估计方法，并通过大量的数值模拟与其他改进的 HCCMEs 进行比较，论证了新估计方法准确性，然后通过实例分析论证了其可行性。

最后，总结本文内容，指出新的研究方向。

1.4 技术路线

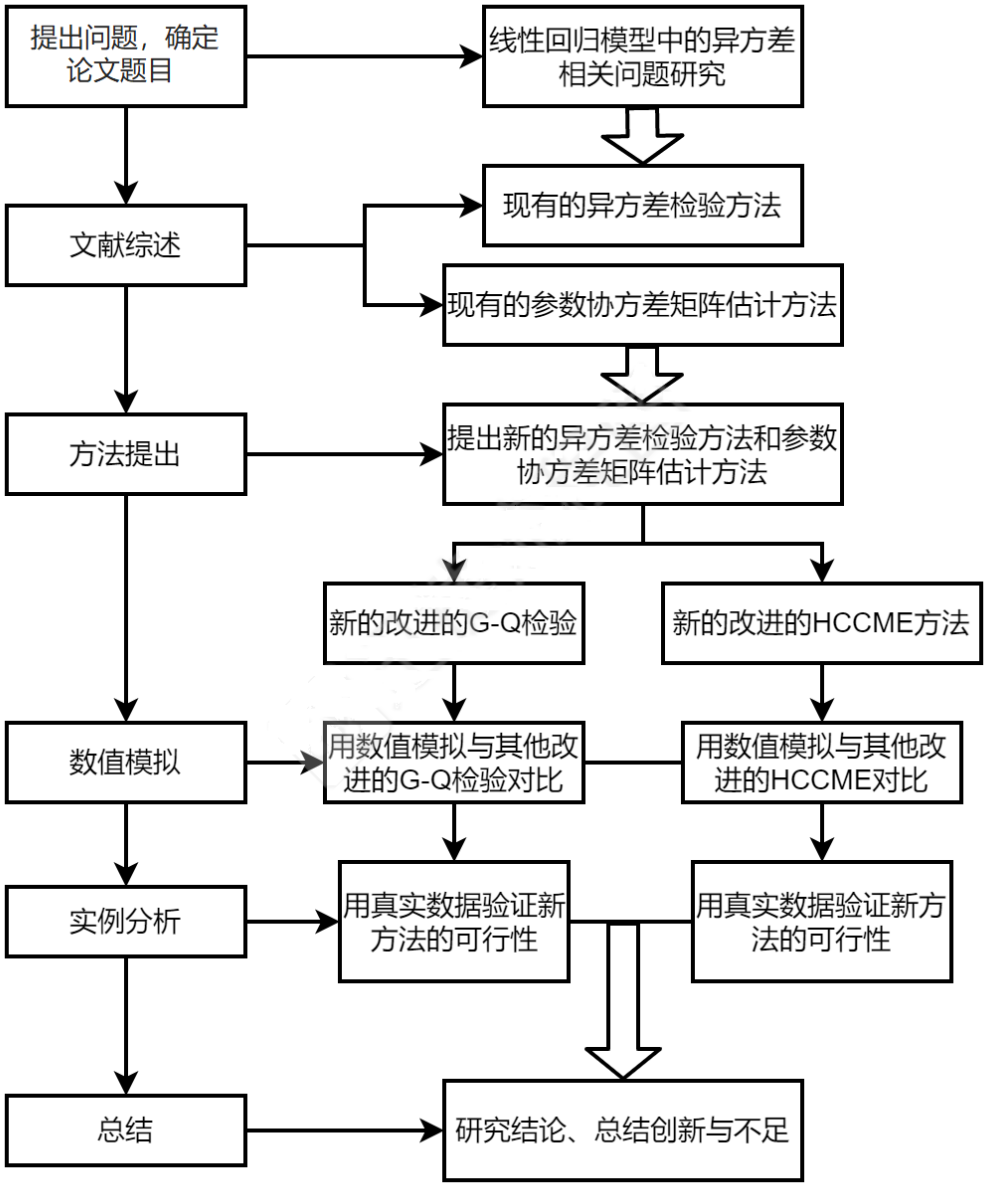

图 1.1 技术路线图

第 2 章 理论与方法概述

本章主要分为两部分：首先分别介绍了传统的 G-Q 检验、改进的 G-Q 检验、几种改进的 G-Q 检验的比较；然后介绍了异方差一致性协方差矩阵估计、改进的异方差一致性协方差矩阵估计，以及几种异方差一致性协方差矩阵估计的比较。

2.1 G-Q 检验

2.1.1 传统的 G-Q 检验

考虑一元回归模型：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \cdots, n \tag{2.1}$$

其中，y_i 为被解释变量，x_i 为解释变量，β_0 为截距项，β_1 为解释变量 x_i 的系数，ε_i 为随机扰动项，n 为样本容量。若 $\forall x_1, x_2, \cdots, x_n$，（2.1）中的每个 ε_i 的方差 σ_i^2 均相等，即 $Var(\varepsilon_i | x_i) = \sigma^2$，则称模型（2.1）具有同方差性，反之则称模型具有异方差性。对异方差性的检验，即考虑如下的假设检验问题：

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2 \quad Vs \quad H_1 : \sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2 \text{不全相等} \tag{2.2}$$

传统的 G-Q 检验用于检验一元回归模型是否存在随 x_i 递增的异方差，有着方便快捷、准确度高等优点。对一元回归模型（2.1）进行 G-Q 检验的步骤如下：

（i）将样本点按照解释变量 x_i 由小到大排列。

（ii）将序列中间的 $c(c \approx n/4)$ 个样本点删去，将序列两端的样本点各自作为一组子样本。

（iii）分别对两组子样本进行 OLS 回归，计算出各自的残差平方和，将 x_i 较小组的残差平方和记为 SSR_1，x_i 较大组的残差平方和记为 SSR_2。

（iv）构造如下在原假设成立下服从 F 分布的统计量：

$$F = \frac{SSR_2 \Big/ (\frac{n-c}{2} - 2)}{SSR_1 \Big/ (\frac{n-c}{2} - 2)} = \frac{SSR_2}{SSR_1} \sim F(\frac{n-c}{2} - 2, \frac{n-c}{2} - 2)$$

（v）给定显著性水平α，确定相应的临界值F_α。若$F > F_\alpha$，则拒绝原假设，认为模型存在异方差；若$F < F_\alpha$,则不拒绝原假设。

2.1.2 已有的改进的 G-Q 检验

近年来有很多学者对 G-Q 检验进行了改进，使其可以应用于多元线性回归模型。龚秀芳通过主成分分析法（PCA）将观测值按照第一主成分的顺序进行由小到大的顺序排列，然后对观测值进行 G-Q 检验。刘明等提出以被解释变量拟合值作为排序标准，将观测值按照由小到大的顺序排列，然后进行 G-Q 检验。郑红艳等将多元回归模型分解为多个一元模型，依次对其进行 G-Q 检验，若有一个回归模型存在异方差，则认为该多元模型存在异方差。

考虑多元回归模型：

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} \sim (\boldsymbol{0_n}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} = diag\left\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2\right\} \end{cases} \tag{2.3}$$

其中：

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \boldsymbol{0}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

龚秀芳提出首先计算$\boldsymbol{X}$的第一主成分$\boldsymbol{X}_f = (x_{1f}, x_{2f}, \cdots, x_{nf})^T$，然后按照$x_{1f}, x_{2f}, \cdots, x_{nf}$的大小对其对应的样本点进行排序，将所有样本点按照$x_{if}$（$i = 1, 2, \cdots, n$)由小到大的顺序排列，之后进行 G-Q 检验，将此方法记为 PCA-G-Q 检验。刘明提出首先对样本进行多元线性回归，然后按照回归的拟合值$\hat{Y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)^T$的大小对其对应的样本点进行排序，之后进行 G-Q 检验，将此方法记为 Yhat-G-Q 检验。之后的具体步骤为：

（i）对样本点进行排序后，得到对应的观察值数列为$(y_i'; x_{i1}', x_{i2}', \cdots, x_{ik}')$，$i = 1, 2, \cdots, n$.

（ii）删除数列$(y_i'; x_{i1}', x_{i2}', \cdots, x_{ik}')$中间的$c(c \approx n/4)$个数据，得到两列样本数列：观测值$x_{im}$较小的样本数列：$(y_i'; x_{i1}', x_{i2}', \cdots, x_{ik}')$，其中$i = 1, 2, \cdots, n_1$，观测值$x_{im}$较大

的样本数列：$(y_i';x_{i1}',x_{i2}',\cdots,x_{ik}')$，其中$i=n_1+c+1,n_1+c+2,\cdots,n$，　n_1与n_2表示两组样本数列的样本容量，其中：$n_1=n_2=\dfrac{n-c}{2}$，且$n_1+n_2+c=n$.

（iii）由模型（2.3）假设这两部分样本数列的回归模型矩阵形式分别为：

$$\begin{cases}\boldsymbol{Y}_1=\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_1\sim N(\boldsymbol{0}_n,\boldsymbol{\Sigma}_1)\end{cases}\tag{2.4}$$

$$\begin{cases}\boldsymbol{Y}_2=\boldsymbol{X}_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}_2\\ \boldsymbol{\varepsilon}_2\sim N(\boldsymbol{0}_n,\boldsymbol{\Sigma}_2)\end{cases}\tag{2.5}$$

其中，$\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2$分别是$n_1,n_2$阶的对角矩阵，$\boldsymbol{Y}_1,\boldsymbol{\varepsilon}_1$与$\boldsymbol{Y}_2,\boldsymbol{\varepsilon}_2$分别是$n_1$维与$n_2$维列向量，$\boldsymbol{\beta}_1,\boldsymbol{\beta}_2$均是$k+1$维列向量，$\boldsymbol{X}_1,\boldsymbol{X}_2$分别是$n_1\times(k+1)$和$n_2\times(k+1)$的列满秩矩阵。

（iv）分别对模型（5）,（6）进行普通最小二乘回归，得出其各自残差平方和：

$$SSR_1=\sum_{i=1}^{n_1}(y_i'-\sum_{j=0}^{k}\hat{\beta}_{1j}x_{ij}')^2$$

$$SSR_2=\sum_{i=l+c+1}^{n}(y_i'-\sum_{j=0}^{k}\hat{\beta}_{2j}x_{ij}')^2$$

构造检验统计量F：

$$F=\frac{SSR_2/(n_2-k-1)}{SSR_1/(n_1-k-1)}=\frac{SSR_2}{SSR_1}$$

在（2.2）的原假设成立的情况下，F统计量服从自由度为(n_2-k-1,n_1-k-1)的F分布。

（v）给定显著性水平α，得到对应的检验临界值：

$$F_{\alpha/2}(n_2-k-1,n_1-k-1)，\ F_{1-\alpha/2}(n_2-k-1,n_1-k-1)，$$

若$F>F_{\alpha/2}(n_2-k-1,n_1-k-1)$或$F<F_{1-\alpha/2}(n_2-k-1,\ \ n_1-k-1)$则拒绝原假设，认为样本数据存在异方差；否则不拒绝原假设。

2.1.3 几种改进的 G-Q 检验的比较

龚秀芳提出的改进方法利用主成分对样本数据进行排序，解决了对多变量数据的排序问题。刘明提出的改进方法选取了被解释变量的拟合值来反映样本数据整体变动趋势，通过对被解释变量的拟合值进行排序来解决多元线性回归模型中

的数据排序问题。他们的改进方法都通过选取一个能代表整体的变量来进行排序，从而将 G-Q 检验推广到了多元线性回归中。郑红艳提出的改进方法是将多元回归模型分解为多个一元模型，依次对其进行 G-Q 检验。他们的方法都将 G-Q 检验进行了推广，扩展了使用范围，但是存在着适用性不广、准确性不高等问题。因此，本文提出了一种新的改进的 G-Q 检验方法，提高了检验的适用性和准确性。

2.2 异方差一致性协方差矩阵估计

2.2.1 异方差一致性协方差矩阵估计

考虑如下线性回归模型：

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

其中 $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)^T$ 为 $n \times 1$ 的被解释变量，$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$ 为 $n \times (k+1)$ 的解释变量，$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_k)^T$ 为 $(k+1) \times 1$ 的回归参数向量，$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ 为 $n \times 1$ 的随机扰动项，且 $\mathrm{cov}(\boldsymbol{\varepsilon}) = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2) \hat{=} \boldsymbol{\Omega}$.

使用最小二乘法得出 $\boldsymbol{\beta}$ 的估计量为 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$，则 $\hat{\boldsymbol{\beta}} \sim (\boldsymbol{\beta}, \boldsymbol{\Psi})$，且

$$\boldsymbol{\Psi} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \quad \boldsymbol{\Omega} = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$$

在同方差假定下，$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$，此时 $\boldsymbol{\Omega} = \sigma^2\boldsymbol{I}_n$，$\boldsymbol{I}_n$ 为 n 阶单位阵，则 $\hat{\boldsymbol{\beta}}$ 的协方差阵可简化为 $\boldsymbol{\Psi} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$，它的一致估计量为 $\hat{\boldsymbol{\Psi}} = \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$，其中 $\hat{\sigma}^2 = \boldsymbol{e}^T\boldsymbol{e}/(n-k-1)$，此处 $\boldsymbol{e} = (e_1, e_2, \cdots, e_n)^T = (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{Y}$ 为最小二乘残差向量。当存在异方差时，各个 σ_i^2 不再完全相等，为此，White 提出了 $\boldsymbol{\Psi}$ 的一致估计量 HC0：

$$HC0 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\Omega}}_0\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \quad \hat{\boldsymbol{\Omega}}_0 = diag(e_1^2, e_2^2, \cdots, e_n^2)$$

然而，通过蒙特卡罗方法可以验证，HC0 收敛到 $\boldsymbol{\Psi}$ 的速度并不快，正如 MacKinnon 和 White 在文献[36]中提到的，在样本量较小时，HC0 对 $\hat{\beta}_j$ $(j = 0,1,\cdots,k)$ 的方差估计值 $\hat{\sigma}_j^2$ 会低于 $\hat{\beta}_j$ 的真实方差。此时，考虑假设检验问题：

$$H_0:\beta_j=r \quad Vs \quad H_1:\beta_j\neq r$$

取显著性水平为α，检验的统计量$t_j=\dfrac{\hat{\beta}_j-\beta_j}{\hat{\sigma}_j}$会偏大，从而$t_j$落入拒绝域$W=\left\{\left|t_j\right|\geq t_{1-\alpha/2}(n-k-1)\right\}$的可能性变大。这会导致过度拒绝原假设，从而可能造成较为严重的误差。

2.2.2 已有的改进的异方差一致性协方差矩阵估计

为了解决误差方差估计偏小的问题，很多学者提出了 HC0 的改进估计量。HC1 估计量可以表示为：

$$HC1=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\Omega}}_1\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}，\text{其中}\ \hat{\boldsymbol{\Omega}}_1=n/(n-k-1)\hat{\boldsymbol{\Omega}}_0$$

其他 HCs 估计量可以表示为：

$$HCs=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\Omega}}_s\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1},s=2,3,4,5,4m,5m,6$$

其中：

$\hat{\boldsymbol{\Omega}}_s=diag\{\dfrac{e_1^2}{(1-h_1)^{\delta_1}},\dfrac{e_2^2}{(1-h_2)^{\delta_2}},\cdots,\dfrac{e_n^2}{(1-h_n)^{\delta_n}}\}$，

h_i为投影矩阵$\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ 的对角元素。

在 HC2 中，$\delta_i=1$；

在 HC3 中，$\delta_i=2$；

在 HC4 中，$\delta_i=\min\{4,h_i/\bar{h}\}$；

在 HC5 中，$\delta_i=\min\{h_i/\bar{h},\max\{4,0.7h_{\max}/\bar{h}\}\}$；

在 HC4m 中，$\delta_i=\min\{1,h_i/\bar{h}\}+\min\{1.5,h_i/\bar{h}\}$；

在 HC5m 中，$\delta_i=\min\{1,h_i/\bar{h}\}+\min\{h_i/\bar{h},\max\{4,0.7h_{\max}/\bar{h}\}\}$；

在 HC6 中，$\delta_i=\min\{h_i/\bar{h},\sqrt{h_{\max}/(2\bar{h})}\}$.

上述表达式中，$\bar{h}=\dfrac{1}{n}\sum_{i=1}^{n}h_i$，$h_{\max}=\max\{h_1,h_2,\cdots,h_n\}$，$i=1,2,\cdots,n$.

2.2.3 几种改进的异方差一致性协方差矩阵估计的比较

很多学者对 HC0 进行了改进，在不同方面与不同程度上提升其在小样本下参

数检验的准确性，但它们仍有各自的不足之处。例如 HC1 和 HC2 的偏差虽然比 HC0 小，但它仍然存在。文献[36]说明了 HC3 的估计效果优于 HC1 和 HC2，但当异方差程度较大时，HC3 也会产生较大的偏差。HC4 考虑了当数据存在高杠杆点时的情况，HC5 考虑了数据中最大的杠杆点，提高了异方差程度较大时的估计效果，但在异方差程度较小时它们表现较差。HC4m 在异方差程度较小时表现较好，但在异方差程度较大时表现不佳。HC5m 仅在部分情形下表现较好，在多数情况下存在高估参数方差的情况。为了对上述 HCCMEs 进行改进，本文提出了一种新的 HCCv 估计，可以适应不同的异方差情况，从而提高估计的准确性与适用性。

第 3 章 一种新的 G-Q 检验方法

第 2 章介绍的几种改进的 G-Q 检验方法都针对 G-Q 检验无法直接应用于多元模型的问题进行了改进，几种方法各有其优势，但也存在着一些不足，例如准确性不高、适用性不广等问题。为解决这些问题，本文借鉴 White 检验的思想，提出了一种新的改进的 G-Q 检验方法，通过大量的数值模拟实验论证了新方法的准确性，并通过实例分析论证了新方法的可行性。

3.1 基于变量选择的异方差 G-Q 检验

由第 2 章的内容可知虽然传统的 G-Q 检验优点众多，但其一般情况下仅能应用于一元回归模型，而无法应对多元情形。要在多元回归模型中使用 G-Q 检验，关键在于如何选择排序的标准。龚秀芳[21]提出了使用主成分分析法计算出样本的第一主成分来代表所有解释变量，使用第一主成分作为排序标准进行 G-Q 检验。但在实际应用中发现，使用该方法进行检验的准确度并不够高，尤其在第一主成分的贡献率较低的情况下。与龚秀芳[21]改进的 G-Q 检验的思想类似，本文考虑找出一个能代表所有解释变量对随机扰动项方差的影响的解释变量来进行排序。在变量选择的方法上，借鉴 White 检验的思想，通过对残差平方与所有解释变量进行 OLS 回归，找出参数 p 值最小的解释变量，该解释变量即为所有解释变量中对随机扰动项方差影响最大的解释变量，以该解释变量为排序标准进行 G-Q 检验。

考虑多元回归模型：

$$\begin{cases} \boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}\sim(\boldsymbol{0_n},\boldsymbol{\Sigma}),\ \boldsymbol{\Sigma}=diag\left\{\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2\right\} \end{cases} \tag{3.1}$$

其中：

$$\boldsymbol{Y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\quad \boldsymbol{X}=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix},\quad \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},\quad \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},\quad \boldsymbol{0}_n=\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

考虑如下的假设检验问题：

$$H_0:\sigma_1^2=\sigma_2^2=\cdots=\sigma_n^2=\sigma^2 \quad Vs \quad H_1:\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2\text{不全相等} \qquad (3.2)$$

为对假设（3.2）进行检验，建立随机扰动项的方差$\sigma_i^2(i=1,2,\cdots,n)$与解释变量$x_{i1},x_{i2},\cdots,x_{ik}\ (i=1,2,\cdots,n)$的回归模型，参考在 White 检验中，使用样本点的残差平方e_i^2代替随机扰动项的方差σ_i^2不影响检验统计量的分布[20]，所以本文使用残差平方e_i^2代替随机扰动项的方差σ_i^2与解释变量$x_{i1},x_{i2},\cdots,x_{ik}(i=1,2,\cdots,n)$做回归。其中$e_i^2=(y_i-\hat{y}_i)^2$，样本拟合值$\hat{y}_i$为：

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\cdots+\hat{\beta}_kx_{ik},i=1,2,\cdots,n$$

其中，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_k$为参数$\beta_0,\beta_1,\beta_2,\cdots,\beta_k$的普通最小二乘估计，且有$E(\hat{\beta}_j)=\beta_j$，$j=0,1,\cdots,k$。

本文利用样本残差平方$e_i^2(i=1,2,\cdots,n)$与解释变量$x_{i1},x_{i2},\cdots,x_{ik}(i=1,2,\cdots,n)$得到回归模型：

$$e_i^2=\alpha_0+\alpha_1x_{i1}+\cdots+\alpha_kx_{ik}+\nu_i,i=1,2,\cdots,n \qquad (3.3)$$

其中，$\alpha_0,\alpha_1,\cdots,\alpha_k$为回归模型的参数，$\nu_i(i=1,2,\cdots,n)$为随机扰动项。

对模型（3.3）中$\alpha_0,\alpha_1,\cdots,\alpha_k$进行$t$检验，找出检验$p$值最小的参数对应的解释变量，记为$x_{im}(1\le m\le k,i=1,2,\cdots,n)$。即在所有解释变量中，$x_{im}$对随机扰动项方差的影响最大，即最有可能引起随机扰动项产生异方差的解释变量是x_{im}。将所有样本点按照x_{im}由小到大的顺序排列，相当于样本点根据随机扰动项方差的影响因素进行排序，这与 G-Q 检验在一元回归模型中的思想一致，之后进行 G-Q 检验。将此方法称为基于变量选择的 G-Q 检验（简称为 M-G-Q 检验）。具体步骤为：

（i）按照最优解释变量x_{im}由小到大的顺序对样本点进行排序，得到对应的观察值数列为$(y_i';x_{i1}',x_{i2}',\cdots,x_{ik}'),i=1,2,\cdots,n$。

（ii）删除数列$(y_i';x_{i1}',x_{i2}',\cdots,x_{ik}')$中间的$c(c\approx n/4)$个数据，得到两列样本数列：

观测值x_{im}较小的样本数列：$(y_i';x_{i1}',x_{i2}',\cdots,x_{ik}')$，其中$i=1,2,\cdots,n_1$。

观测值x_{im}较大的样本数列：$(y_i';x_{i1}',x_{i2}',\cdots,x_{ik}')$，其中$i=n_1+c+1,n_1+c+2,\cdots,n$。用$n_1$与$n_2$表示两组样本数列的样本容量，其中：$n_1=n_2=\dfrac{n-c}{2}$，

且 $n_1+n_2+c=n$ 。

（iii）由模型（3.1）假设这两部分样本数列的回归模型矩阵形式分别为：

$$\begin{cases}\boldsymbol{Y}_1=\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_1\sim N(\boldsymbol{0}_n,\boldsymbol{\Sigma}_1)\end{cases}\tag{3.4}$$

$$\begin{cases}\boldsymbol{Y}_2=\boldsymbol{X}_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}_2\\ \boldsymbol{\varepsilon}_2\sim N(\boldsymbol{0}_n,\boldsymbol{\Sigma}_2)\end{cases}\tag{3.5}$$

其中， $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2$ 分别是 n_1,n_2 阶的对角矩阵， $\boldsymbol{Y}_1,\boldsymbol{\varepsilon}_1$ 与 $\boldsymbol{Y}_2,\boldsymbol{\varepsilon}_2$ 分别是 n_1 维与 n_2 维列向量， $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2$ 均是 $k+1$ 维列向量， $\boldsymbol{X}_1,\boldsymbol{X}_2$ 分别是 $n_1\times(k+1)$ 和 $n_2\times(k+1)$ 的列满秩矩阵。

（iv）分别对模型（3.4），（3.5）进行普通最小二乘回归，得出其各自的残差平方和：

$$SSR_1=\sum_{i=1}^{n_1}(y_i'-\sum_{j=0}^{k}\hat{\beta}_{1j}x_{ij}')^2$$

$$SSR_2=\sum_{i=l+c+1}^{n}(y_i'-\sum_{j=0}^{k}\hat{\beta}_{2j}x_{ij}')^2$$

构造检验统计量 F ：

$$F=\frac{SSR_2/(n_2-k-1)}{SSR_1/(n_1-k-1)}=\frac{SSR_2}{SSR_1}$$

在（3.2）的原假设成立的情况下， F 统计量服从自由度为 (n_2-k-1,n_1-k-1) 的 F 分布，证明如下：

令 $P_1=X_1(X_1^TX_1)^{-1}X_1^T$ ， I_{n_1} 是 n_1 阶单位矩阵，根据正交投影矩阵知识有 $M_1+P_1=I_{n_1}$ ， P_1 与 M_1 为互补投影，且均为对称幂等矩阵。

$$\begin{aligned}e_1&=Y_1-\hat{Y}_1=Y_1-X_1\hat{\tilde{\beta}}_1=Y_1-X_1(X_1^TX_1)^{-1}X_1^TY_1=(I_{n_1}-X_1(X_1^TX_1)^{-1}X_1^T)Y_1\\&=(I_{n_1}-X_1(X_1^TX_1)^{-1}X_1^T)(X_1\tilde{\beta}_1+\tilde{\varepsilon}_1)=X_1\tilde{\beta}_1-X_1\tilde{\beta}_1+M_1\tilde{\varepsilon}_1\end{aligned}$$

故 $rank(e_1)=rank(M_1)$ ，由对称幂等矩阵性质有：

$$\begin{aligned}rank(M_1)&=rank(I_{n_1}-P_1)=trace(I_{n_1}-P_1)=n_1-trace(P_1)\\&=n_1-trace(X_1(X_1^TX_1)^{-1}X_1^T)=n_1-trace((X_1^TX_1)^{-1}X_1^TX_1)\\&=n_1-rank((X_1^TX_1)^{-1}X_1^TX_1)=n_1-rank(X_1^TX_1)\\&=n_1-rank(X_1)=n_1-k-1\end{aligned}$$

在原假设成立条件下，有 $\tilde{\varepsilon}_1/\sigma_1\sim N(0_{n_1},I_{n_1})$ ，由正态向量二次型知识有：

$$\frac{SSR_1}{\sigma_1^2}=\frac{e_1'e_1}{\sigma^2}=\frac{1}{\sigma}\tilde{\varepsilon}_1'M_1\tilde{\varepsilon}_1\frac{1}{\sigma}\sim\chi^2(n_1-k-1)$$

由于$\tilde{\varepsilon}_1,\tilde{\varepsilon}_2$相互独立，故同理可得：

$$\frac{SSR_2}{\sigma_2^2}=\frac{e_2'e_2}{\sigma^2}=\frac{1}{\sigma}\tilde{\varepsilon}_2'M_2\tilde{\varepsilon}_2\frac{1}{\sigma}\sim\chi^2(n_2-k-1)$$

原假设成立条件下有$\sigma^2=\sigma_1^2=\sigma_2^2=\cdots=\sigma_n^2$，有

$$\frac{SSR_2/\sigma_2^2}{SSR_1/\sigma_1^2}=\frac{SSR_2}{SSR_1}=F$$

故可得

$$F=\frac{\frac{SSR_2/\sigma_1^2}{(n_2-k-1)}}{\frac{SSR_1/\sigma_2^2}{(n_1-k-1)}}=\frac{SSR_2/(n_2-k-1)}{SSR_1/(n_1-k-1)}=\frac{SSR_2}{SSR_1}\sim F(n_2-k-1,n_1-k-1)$$

（v）给定显著性水平α，得到对应的检验临界值：

$$F_{\alpha/2}(n_2-k-1,n_1-k-1)，\ F_{1-\alpha/2}(n_2-k-1,n_1-k-1)，$$

若$F>F_{\alpha/2}(n_2-k-1,n_1-k-1)$或$F<F_{1-\alpha/2}(n_2-k-1,\ n_1-k-1)$则拒绝原假设，认为样本数据存在异方差；否则不拒绝原假设。

本文所提出的 M-G-Q 检验通过类似 White 检验的t检验从多个解释变量中挑选出对随机扰动项方差影响最大的解释变量作为 G-Q 检验的排序标准，进而进行 G-Q 异方差检验。与龚秀芳改进的 G-Q 检验相比，新方法选择的排序标准本身就是解释变量，能够更好地反映对随机扰动项方差的影响。在 White 检验中，需要存在某个解释变量、二次项或交叉项参数的 p 值低于给定的显著性水平才能拒绝原假设，认为存在异方差。而在新方法中，不必拘泥于给定的显著性水平，只需找出参数 p 值最小的解释变量，然后再以该解释变量作为排序标准进行 G-Q 检验，提高了检验的灵敏度，使检验结果更加准确。本文的F检验使用了双侧检验，这样可以同时检验随解释变量递增和随解释变量递减的异方差，提高了检验的适用性与准确性。

3.2 数值模拟

本小节从数值模拟分析的角度对龚秀芳改进的 G-Q 检验（记为 PCA-G-Q 检

验）、刘明改进的 G-Q 检验（记为 Yhat-G-Q 检验）、White 检验和本文提出的 M-G-Q 检验进行比较。本节的数值模拟分析通过 Python 实现。

3.2.1 生成数据

使用如下线性回归模型：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i, i = 1,2,\cdots,n$$

其中样本容量 n 为 50，100 或 200，$\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$，解释变量 $x_{i1}, x_{i2}, x_{i3}, x_{i4}$ 之间的相关系数 r 为 0，0.1，0.3，0.5，0.7 或 0.9，且产生自正态分布 $N(0,1)$，ε_i 产生自正态分布 $N(0,\sigma_i^2)$，为了详细对比上述各方法的检验效果，使数值模拟实验更具说服力，本文模拟了多种不同异方差，σ_i^2 的取值有以下六种情况：

（a）$\sigma_i^2 = e^{x_{i1}}$；

（b）$\sigma_i^2 = e^{-x_{i1}}$；

（c）$\sigma_i^2 = e^{x_{i1}-x_{i2}}$；

（d）$\sigma_i^2 = e^{x_{i1}+x_{i2}+x_{i3}+x_{i4}}$；

（e）$\sigma_i^2 = \left|\ln({x_{i1}}^2)\right|$；

（f）$\sigma_i^2 = \left|\ln((x_{i1}+x_{i2})^2)\right|$.

其中，情况（a）模拟的是 σ_i^2 随某个解释变量递增的情况，情况（b）模拟的是 σ_i^2 随某个解释变量递减的情况，情况（c）模拟的是 σ_i^2 同时受两个解释变量的影响且方向相反的情况，情况（d）模拟的是 σ_i^2 同时受四个解释变量同方向的影响的情况。情况（e）和（f）模拟了两种较为复杂的异方差情况。

3.2.2 异方差检验

模拟实验中原假设为模型不存在异方差，给定显著性水平 $\alpha = 0.05$，对每种不同的异方差情况在不同的样本容量情况下生成的数据分别进行 PCA-G-Q 检验、Yhat-G-Q 检验检验、White 检验及 M-G-Q 检验。每种情况重复 10000 次实验，统计各方法拒绝原假设的次数，以各方法对异方差的拒绝频率（拒绝次数与实验次数的比率）作为评价指标来对比各方法在不同条件下的检验效果，结果如表 3.1-表 3-4 所示。

1. $r = 0$ 时，

表 3.1 $r=0$ 时异方差检验结果（10000 次实验）

σ_i^2	n	PCA-G-Q 检验拒绝次数	Yhat-G-Q 检验拒绝次数	White 检验拒绝次数	**M-G-Q 检验拒绝次数**
$e^{x_{i1}}$	50	4592	4265	1057	**9017**
	100	6196	6997	7314	**9987**
	200	7362	9082	9949	**10000**
$e^{-x_{i1}}$	50	4530	4231	984	**9035**
	100	6296	7086	7294	**9989**
	200	7303	9111	9952	**10000**
$e^{x_{i1}-x_{i2}}$	50	6068	3130	2217	**9292**
	100	7369	4001	8928	**9986**
	200	8200	4465	9993	**10000**
$e^{x_{i1}+x_{i2}+x_{i3}+x_{i4}}$	50	7079	9892	3519	**9360**
	100	8130	9991	9542	**9949**
	200	8657	9999	9997	**9999**
$\left\|\ln(x_{i1}^{\ 2})\right\|$	50	1652	1545	35	**3288**
	100	1966	1916	295	**4019**
	200	2101	2088	590	**4451**
$\left\|\ln((x_{i1}+x_{i2})^2)\right\|$	50	1398	1359	45	**3118**
	100	1801	1626	388	**3929**
	200	1865	1925	468	**4315**

表 3.1 以及图 3-1 到图 3-6 展示了各解释变量相互独立时在六种异方差情况下不同方法随样本容量变化的拒绝频率变化情况，通过表 3.1 以及图 3-1 到图 3-6，可在各解释变量相互独立的情况下得出如下结论：

（1）在六种不同的异方差情况下，M-G-Q 检验的异方差检出率均高于 PCA-G-Q 检验，例如，在情况（a）中，样本容量为 50 的情况下，相比之前性能最好的 PCA-G-Q 方法 45.9%的拒绝频率，本文提出的 M-G-Q 检验的拒绝频率提升到了 90.2%，可以说明通过 M-G-Q 检验选择出的最优解释变量比 PCA-G-Q 检验的第一主成分更能反映随机扰动项的方差情况。

（2）Yhat-G-Q 检验在情况（a）和（b）中表现尚可，说明用样本的拟合值 $\hat{y}_i$ 进行排序具有一定的合理性，但远不如用 M-G-Q 检验中选择出的最优解释变量效果好。在情况（d）中 Yhat-G-Q 检验表现最佳，因为在情况（d）中，四个解释变

量协同影响随机扰动项的方差，此时用样本的拟合值 $\hat{y}_i$ 进行排序效果很好，而本文提出的 M-G-Q 检验仅在样本容量为 50 时的拒绝频率比 Yhat-G-Q 检验略低，在样本容量为 100 或 200 时检验效果与 Yhat-G-Q 检验不相上下。在情况（c）中，使用 Yhat-G-Q 检验的效果较差，因为此时两个解释变量对随机扰动项方差的影响可能会相互抵消，用 $\hat{y}_i$ 排序并不合理，但使用 M-G-Q 检验依然可以很好地识别出异方差。在情况（e）和（f）中，本文提出的 M-G-Q 检验效果均比 Yhat-G-Q 检验要好，例如，在情况（f）中，样本容量为 200 的情况下，相比之前性能最好的 Yhat-G-Q 方法 19.3%的拒绝频率，本文提出的 M-G-Q 方法的拒绝频率提升到了 43.2%。因此 M-G-Q 检验比 Yhat-G-Q 检验更具有合理性与泛用性。

（3）White 检验是多元线性回归中最常用的异方差检验方法之一，但它的缺陷也显而易见，即在解释变量较多时，自由度损失严重，所以 White 检验要求的样本容量较大，这一点也在模拟结果中得以体现，当 $n=50$ 时，White 检验的效果是极差的。M-G-Q 检验要求的样本容量远小于 White 检验，在情况（a）到（d）中，样本容量为 200 时，White 检验效果很好，拒绝频率达到 99.5%以上，此时本文提出的 M-G-Q 检验的拒绝频率与 White 检验十分接近，在情况（e）和（f）中，M-G-Q 检验的表现均优于 White 检验。

（4）在相对复杂且其他方法较难识别的异方差情况（e）和（f）中，M-G-Q 的检验效果仍远比其他几种方法要好。因此，M-G-Q 在异方差检验中灵敏性更高。

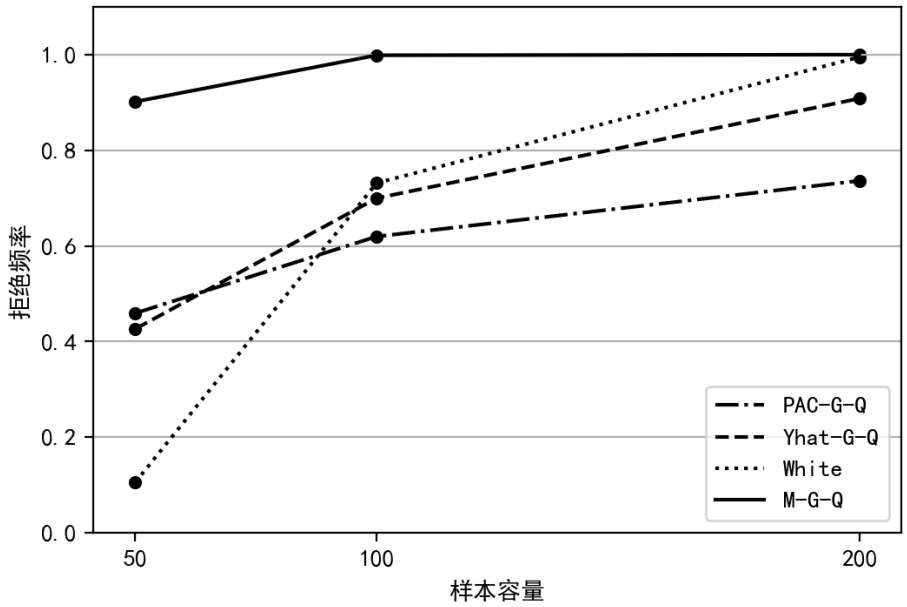

图 3.1 $r=0,\ \sigma_i^2=e^{x_{i1}}$ **时各方法的拒绝频率（拒绝次数与实验次数的比率）**

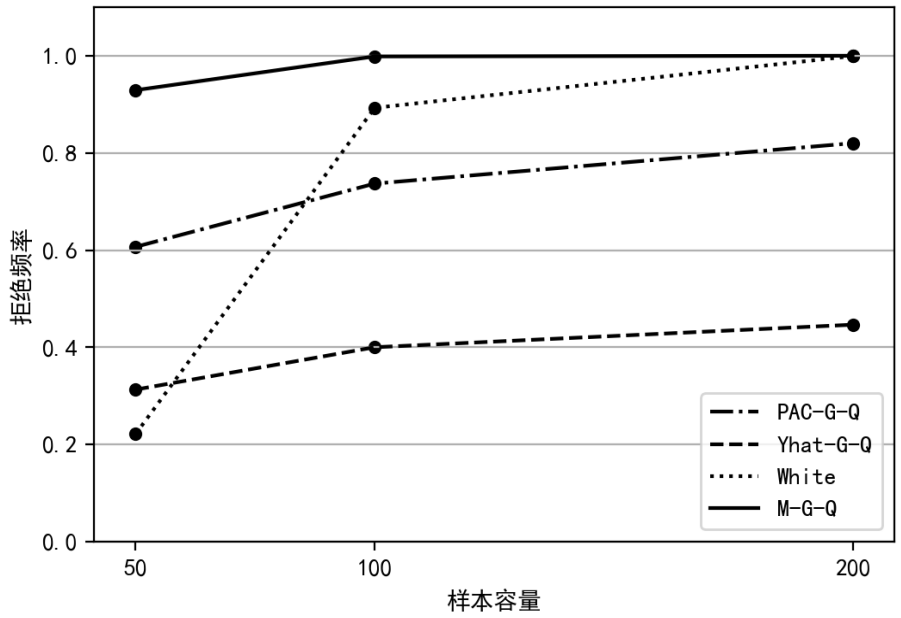

图 3.2 $r=0,\ \sigma_i^2=e^{-x_{i1}}$ **时各方法的拒绝频率（拒绝次数与实验次数的比率）**

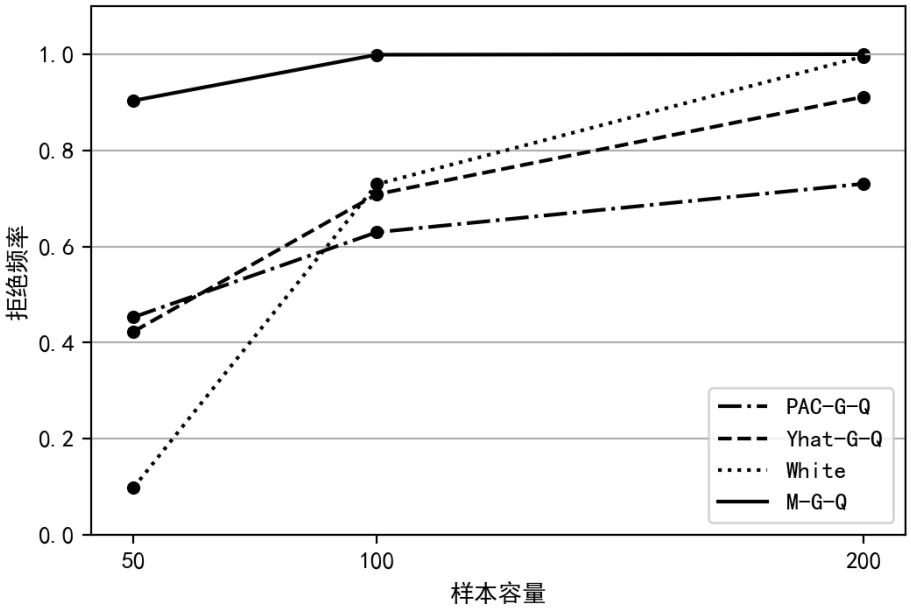

图 3.3 $r=0,\ \sigma_i^2=e^{x_{i1}-x_{i2}}$ **时各方法的拒绝频率（拒绝次数与实验次数的比率）**

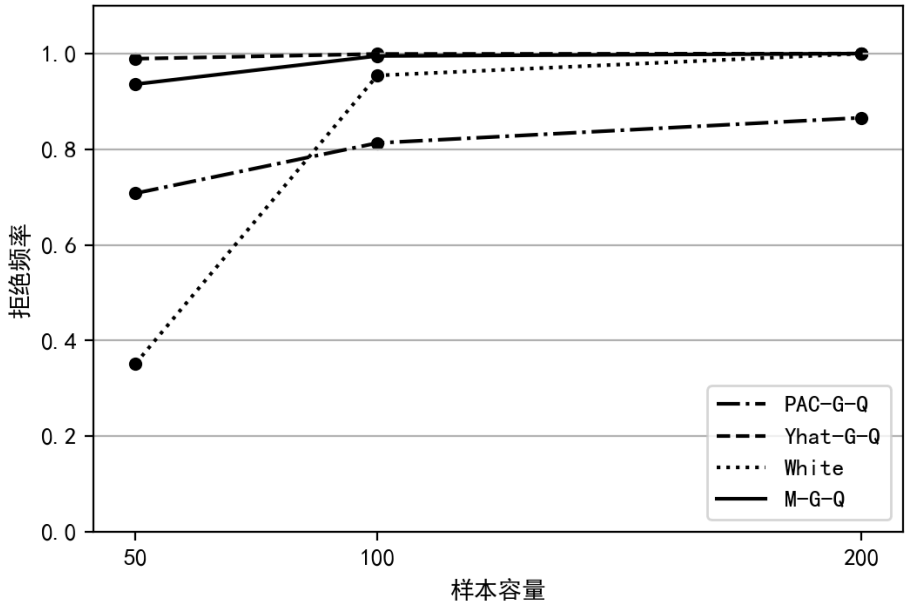

图 3.4 $r=0,\ \sigma_i^2=e^{x_{i1}+x_{i2}+x_{i3}+x_{i4}}$ **时各方法的拒绝频率（拒绝次数与实验次数的比率）**

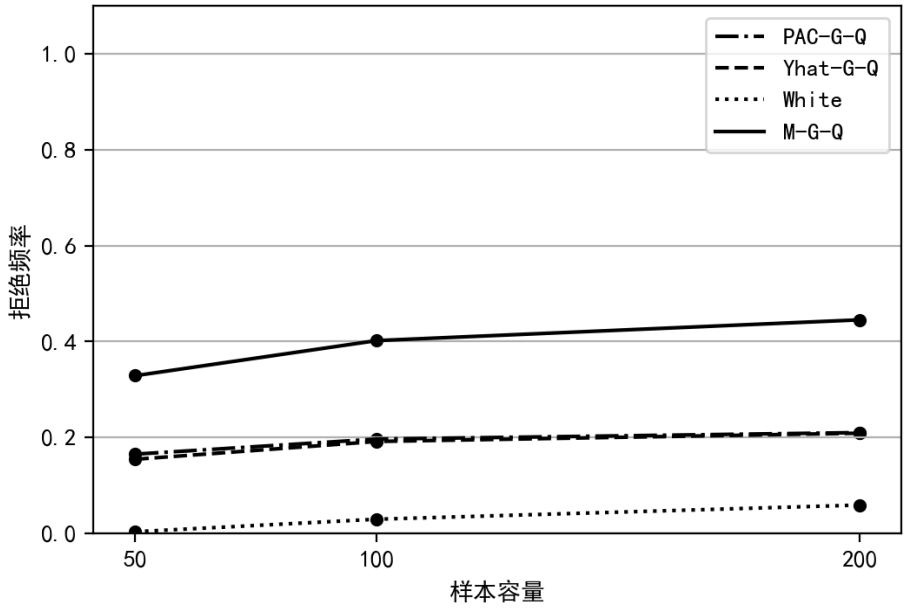

图 3.5 $r=0,\ \sigma_i^2=\left|\ln(x_{i1}{}^2)\right|$ **时各方法的拒绝频率（拒绝次数与实验次数的比率）**

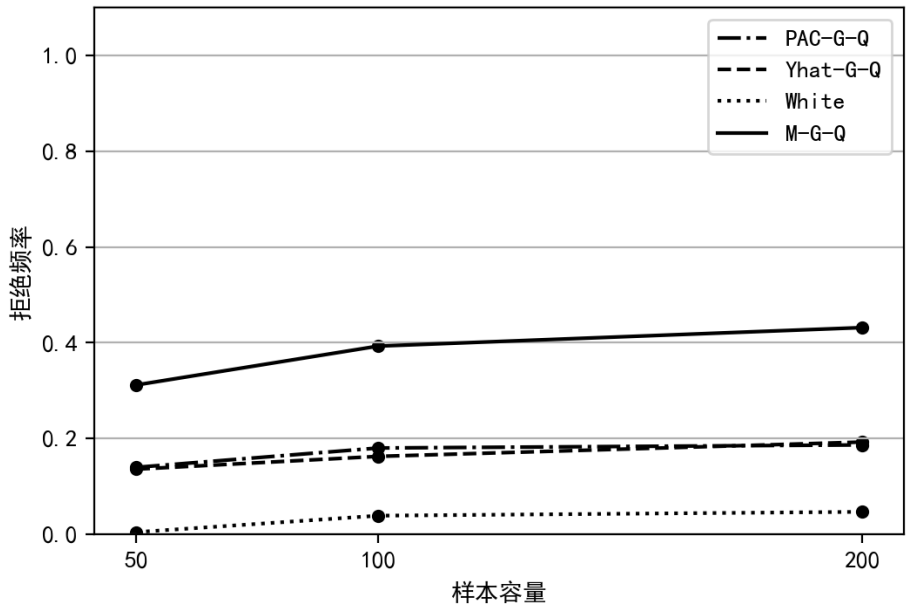

图 3.6 $r=0, \sigma_i^2=\left|\ln((x_{i1}+x_{i2})^2)\right|$ **时各方法的拒绝频率（拒绝次数与实验次数的比率）**

在实际应用中，各解释变量并不一定是相互独立的情况，因此，本文对各解释变量间的相关系数取不同的数值的情况进行了实验，并选取了几组具有代表性的结果进行分析。

2. $r=0.1$时，

如表 3.2 所示，通过数值模拟实验，在各个解释变量间的相关系数$r=0.1$的情况下可以看出：

（1）在情况（*a*）、（*b*）和（*d*）中，PCA-G-Q 检验和 Yhat-G-Q 检验的检验效果均有小幅度增加，例如在情况（*a*）中，样本容量为 50 的情况下，相比各解释变量相互独立情况，PCA-G-Q 方法的拒绝频率从 45.9%提升到了 49.3%，White 检验在这三种情况下的拒绝频率基本保持不变，本文提出的 M-G-Q 检验的拒绝频率也基本保持不变，且在各个样本容量大小的条件下检验效果均优于其他检验方法。

（2）在情况（*c*）中，四种检验方法的拒绝频率均有所下降，PCA-G-Q 检验的下降幅度最大，例如在样本容量为 50 的情况下，PCA-G-Q 检验的拒绝频率从 60.7%降低到了 45.6%，Yhat-G-Q 检验的拒绝频率下降幅度次之，例如在样本容量为 100 的情况下，Yhat-G-Q 检验的拒绝频率从 40.0%降低到了 35.7%，而 White 检验和本文提出的 M-G-Q 检验下降幅度很小，且 M-G-Q 检验在情况（*c*）中的表现均优于其他几种方法。

（3）在情况（e）中，四种检验方法的拒绝频率与各解释变量相互独立的情况

较为接近，在情况（f）中，M-G-Q 检验的拒绝频率较 的情况略有降低，但仍远远高于其他几种检验方法。

表 3.2 $r=0.1$时异方差检验结果（10000 次实验）

σ_i^2	n	PCA-G-Q 检验拒绝次数	Yhat-G-Q 检验拒绝次数	White 检验拒绝次数	**M-G-Q 检验拒绝次数**
$e^{x_{i1}}$	50	4931	5020	1065	**9015**
	100	7052	7919	7239	**9986**
	200	8801	9606	9945	**10000**
$e^{-x_{i1}}$	50	4847	4989	1027	**9037**
	100	7113	7948	7281	**9990**
	200	8829	9580	9952	**10000**
$e^{x_{i1}-x_{i2}}$	50	4560	2865	2083	**8978**
	100	5269	3571	8760	**9964**
	200	5688	4107	9992	**10000**
$e^{x_{i1}+x_{i2}+x_{i3}+x_{i4}}$	50	9313	9864	3980	**9717**
	100	9869	9978	9556	**9983**
	200	9984	9997	9999	**10000**
$\left\|\ln(x_{i1}^2)\right\|$	50	1604	1632	37	**3280**
	100	2008	2011	292	**4057**
	200	2113	2162	561	**4467**
$\left\|\ln((x_{i1}+x_{i2})^2)\right\|$	50	1348	1348	50	**2999**
	100	1598	1656	418	**3776**
	200	1702	1722	554	**4172**

3. $r=0.3$时，

如表 3.3 所示，通过数值模拟实验，在各个解释变量间的相关系数$r=0.3$的情况下可以发现：

（1）随着各解释变量间相关系数的增加，在情况（a）、（b）和（d）中，PCA-G-Q 检验和 Yhat-G-Q 检验的检验效果变得更好，例如在情况（a）中，样本容量为 50 的情况下，相比$r=0.1$情况，PCA-G-Q 方法的拒绝频率从 49.3%提升到了 62.3%，在情况（d）中，PCA-G-Q 检验获得了相当准确的检验结果。White 检验在这三种情况下的拒绝频率基本保持不变，本文提出的 M-G-Q 检验的拒绝频率在这三种情况下也基本保持不变，在这几种情况下均保持较为准确的检验结果。

（2）在情况（c）中，四种检验方法的拒绝频率较 的情况均有所下降，PCA-G-Q

检验的下降幅度最大，例如在样本容量为 50 的情况下，PCA-G-Q 检验的拒绝频率从 45.6%降低到了 25.0%，Yhat-G-Q 检验的拒绝频率从 28.7%降至 22.2%，而 White 检验和本文提出的 M-G-Q 检验下降幅度较小，同时可以看出，White 检验受样本量影响很大，在样本容量 的情况下，White 检验与 M-G-Q 检验的拒绝频率均接近 100%，而在 时，White 检验的拒绝频率下降至 15.7%，相较于 White 检验，M-G-Q 检验受样本容量的影响较小，在样本容量较小的情况下的表现优于 White 检验。

表 3.3 $r=0.3$ 时异方差检验结果（10000 次实验）

σ_i^2	n	PCA-G-Q 检验拒绝次数	Yhat-G-Q 检验拒绝次数	White 检验拒绝次数	**M-G-Q 检验拒绝次数**
$e^{x_{i1}}$	50	6233	6236	1106	**8951**
	100	8956	9092	7279	**9988**
	200	9919	9954	9961	**10000**
$e^{-x_{i1}}$	50	6118	6150	1026	**8902**
	100	8959	9075	7322	**9982**
	200	9910	9939	9957	**10000**
$e^{x_{i1}-x_{i2}}$	50	2500	2215	1567	**7711**
	100	3258	2980	8184	**9721**
	200	3604	3348	9985	**9998**
$e^{x_{i1}+x_{i2}+x_{i3}+x_{i4}}$	50	9995	9794	4384	**9945**
	100	10000	9943	9656	**9998**
	200	10000	9981	9997	**10000**
$\left\|\ln(x_{i1}^2)\right\|$	50	1663	1664	22	**3126**
	100	1947	1998	286	**3844**
	200	2158	2169	551	**4357**
$\left\|\ln((x_{i1}+x_{i2})^2)\right\|$	50	1248	1278	53	**2721**
	100	1610	1614	491	**3559**
	200	1711	1679	693	**3872**

（3）在情况（e）中，四种检验方法的拒绝频率与 $r=0.1$ 的情况类似，在情况（f）中，M-G-Q 检验的拒绝频率较 $r=0.1$ 的情况略有降低，但仍高于其他几种检验方法。

4. $r=0.5$ 时，

如表 3.4 所示，通过观察在各个解释变量间的相关系数 的情况下的数值模拟

实验结果，可以得出：

表 3.4 $r=0.5$ 时异方差检验结果（10000 次实验）

σ_i^2	n	PCA-G-Q 检验拒绝次数	Yhat-G-Q 检验拒绝次数	White 检验拒绝次数	**M-G-Q 检验拒绝次数**
$e^{x_{i1}}$	50	7324	7313	1090	**8877**
	100	9675	9673	7254	**9969**
	200	9996	9998	9951	**9999**
$e^{-x_{i1}}$	50	7349	7308	1080	**8860**
	100	9679	9682	7295	**9968**
	200	9992	9995	9943	**10000**
$e^{x_{i1}-x_{i2}}$	50	1743	1679	989	**5678**
	100	2221	2175	7300	**8528**
	200	2590	2540	9947	**9814**
$e^{x_{i1}+x_{i2}+x_{i3}+x_{i4}}$	50	10000	9671	4605	**9990**
	100	10000	9865	9646	**10000**
	200	10000	9961	9996	**10000**
$\left\|\ln(x_{i1}^2)\right\|$	50	1685	1727	25	**3049**
	100	2035	2056	308	**3623**
	200	2340	2344	550	**4067**
$\left\|\ln((x_{i1}+x_{i2})^2)\right\|$	50	1131	1118	60	**2394**
	100	1380	1401	615	**3009**
	200	1475	1476	913	**3306**

（1）随着各解释变量间相关系数的继续增加，在情况（a）、（b）和（d）中，PCA-G-Q 检验和 Yhat-G-Q 检验的拒绝频率进一步提高，在样本容量 $n=100$ 和 $n=200$ 的情况下与 M-G-Q 检验十分接近。White 检验和 M-G-Q 检验的拒绝频率基本保持不变，M-G-Q 检验在这几种情况下仍保持较为准确的检验结果。

（2）在情况（c）中，四种检验方法的拒绝频率较 $r=0.3$ 的情况均有所下降，在样本容量为 50 的情况下，M-G-Q 检验的拒绝频率下滑较为明显，从 77.1%降低到了 56.8%。 White 检验的结果保持稳定，在样本容量 $n=200$ 时的拒绝频率略微超过了 M-G-Q 检验，除此之外 M-G-Q 检验的拒绝频率均高于其他检验方法。

（3）在情况（e）中，PCA-G-Q 检验和 Yhat-G-Q 检验的拒绝频率较 $r=0.3$ 的情况略有提升，例如在样本容量为 $n=200$ 的情况下，相比 $r=0.3$ 情况，PCA-G-Q

检验的拒绝频率从 21.6%提升到了 23.4%， Yhat-G-Q 检验的拒绝频率从 21.7%提升到了 23.4% ，而 M-G-Q 检验的拒绝频率略有下降，但仍高于其他的方法。

表 3.5 $r=0.7$ 时异方差检验结果（10000 次实验）

σ_i^2	n	PCA-G-Q 检验拒绝次数	Yhat-G-Q 检验拒绝次数	White 检验拒绝次数	**M-G-Q 检验拒绝次数**
$e^{x_{i1}}$	50	8175	8154	1002	**8708**
	100	9900	9890	7323	**9942**
	200	10000	10000	9943	**10000**
$e^{-x_{i1}}$	50	8155	8107	1052	**8773**
	100	9893	9889	7273	**9953**
	200	10000	10000	9961	**9999**
$e^{x_{i1}-x_{i2}}$	50	1136	1130	521	**3097**
	100	1506	1521	5463	**5409**
	200	1709	1700	9677	**7803**
$e^{x_{i1}+x_{i2}+x_{i3}+x_{i4}}$	50	10000	9566	4764	**9997**
	100	10000	9793	9627	**10000**
	200	10000	9900	9991	**10000**
$\left\|\ln(x_{i1}^2)\right\|$	50	1826	1812	31	**2636**
	100	2315	2304	306	**3423**
	200	2518	2504	541	**3696**
$\left\|\ln((x_{i1}+x_{i2})^2)\right\|$	50	1152	1147	67	**1885**
	100	1351	1349	691	**2453**
	200	1460	1465	1156	**2768**

（4）在情况（f）中，White 检验的拒绝频率较 $r=0.3$ 的情况略有提高，但仍处于一个很低的水平，另外三种方法的拒绝频率相较于 $r=0.3$ 的情况略有降低，M-G-Q 检验仍是其中表现最好的。

5. $r=0.7$ 时，

如表 3.5 所示，通过 $r=0.7$ 时数值模拟实验结果，可以看出：随着解释变量间相关系数增加，各检验方法拒绝频率的变化趋势与上文的分析基本保持一致，在情况（c）中，样本容量 $n=100$ 和 $n=200$ 的情况下，White 检验的拒绝频率超过了 M-G-Q 检验，但 M-G-Q 检验在多数情况下仍是效果最佳的。

6. $r=0.9$ 时，

表 3.6 $r=0.9$ 时异方差检验结果（10000 次实验）

σ_i^2	n	PCA-G-Q 检验拒绝次数	Yhat-G-Q 检验拒绝次数	White 检验拒绝次数	**M-G-Q 检验拒绝次数**
$e^{x_{i1}}$	50	9868	8795	1070	**8837**
	100	9985	9985	7327	**9967**
	200	10000	10000	9942	**10000**
$e^{-x_{i1}}$	50	8778	8771	1037	**8805**
	100	9978	9981	7290	**9972**
	200	10000	10000	9960	**10000**
$e^{x_{i1}-x_{i2}}$	50	734	728	120	**1055**
	100	788	783	1949	**1412**
	200	815	838	5924	**2017**
$e^{x_{i1}+x_{i2}+x_{i3}+x_{i4}}$	50	10000	9534	4858	**10000**
	100	10000	9750	9626	**10000**
	200	10000	9867	9991	**10000**
$\left\|\ln(x_{i1}^{\ 2})\right\|$	50	1987	1963	38	**2377**
	100	2504	2488	333	**2977**
	200	2772	2797	568	**3228**
$\left\|\ln((x_{i1}+x_{i2})^2)\right\|$	50	1050	1065	72	**1440**
	100	1164	1167	742	**1711**
	200	1256	1286	1358	**1860**

如表 3.6 所示，通过对比 $r=0.9$ 时与之前的数值模拟实验结果，可以看出：在各解释变量间相关性很强时，在情况（a）、（b）和（d）中，PCA-G-Q 检验和 Yhat-G-Q 检验的检验效果很好，在情况（a）中，样本容量 $n=50$ 的情况下，PCA-G-Q 检验的拒绝频率达到了 98.7%，超过了 M-G-Q-检验的 88.4%，在 $n=100$ 和 $n=200$ 时的拒绝频率与 M-G-Q 检验十分接近。在这三种情况中 M-G-Q 检验保持着较为准确的检验结果。在情况（c）中 M-G-Q 检验的拒绝频率继续降低，在 $n=100$ 和 $n=200$ 时的检验效果差于 White 检验，但仍优于 PCA-G-Q 检验和 Yhat-G-Q 检验。在情况（e）和（f）中，M-G-Q 的检验效果虽持续下降，但始终优于其他三种方法。

纵观各个相关系数下的数值模拟结果，可以得出：PCA-G-Q 检验和 Yhat-G-Q 检验在某些情况下表现较好，例如在解释变量间高相关性条件下的情况（a）、（b）和（d）中，而在有些情况下的缺陷较为明显，例如解释变量间相关性较低的情况。

White 检验受解释变量间的相关系数影响很小，而受样本容量的影响很大，当提高样本容量时，White 检验的检验效果均获得了较大的提升，但 White 检验在样本容量较小的情况下检验效果较差。本文提出的 M-G-Q 检验方法的效果虽受相关性的影响而有所变化，但在绝大多数情况下的检验效果均优于其他几种方法，因此 M-G-Q 检验是一种效果优良的改进的 G-Q 检验方法。

3.3 实例分析

本节将通过实例来验证本章提出的新方法的可行性。从各地区的统计局官网上获取 2021 年 31 个省份（不含港澳台）的人均生产总值(y)，人均消费支出(x_1)，人均第三产业增加值(x_2)，人均对外进出口总值(x_3)，具体数值如表 3.7 所示，该数据由该项经济指标总值除以当地总人口得到。

模型设定如下：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad i = 1, 2, \cdots, 31.$$

使用 31 省份的数据估计出的模型为：

$$\hat{y}_i = 4.145259 - 0.6877 x_{i1} + 1.6837 x_{i2} - 0.6742 x_{i3}, \quad i = 1, 2, \cdots, 31.$$

分别使用 PCA-G-Q 检验、Yhat-G-Q 检验、White 检验以及 M-G-Q 检验进行异方差检验，并比较检验结果。

1. PCA-G-Q 检验

使用 PCA-G-Q 检验对数据进行异方差检验，按照第二章中的 PCA-G-Q 检验方法，样本容量$n = 31$，首先计算$c = 7$，$n_1 = n_2 = \dfrac{n-c}{2} = 12$，然后按第一主成分对样本点进行排序后，计算得出$SSR_1 = 19.4439$，$SSR_2 = 0.1259$，$F$ 统计量的自由度为$n-3-1=8$，可以求得其检验统计量：

$$F_{PCA-G-Q} = \frac{SSR_2/(n-3-1)}{SSR_1/(n-3-1)} = 0.0065 < F_{0.025}(8,8) = 0.2256$$

因此拒绝原假设，认为该模型存在异方差。

2. Yhat-G-Q 检验

使用 Yhat-G-Q 检验对数据进行异方差检验，按照第二章中的 Yhat-G-Q 检验方

法，样本容量$n=31$，首先计算$c=7$，$n_1=n_2=\frac{n-c}{2}=12$，然后按被解释变量的估计量对样本点进行排序后，计算得出$SSR_1=0.1947$，$SSR_2=19.4439$，F统计量的自由度为$n-3-1=8$，可以求得其检验统计量：

$$F_{Yhat-G-Q}=\frac{SSR_2/(n-3-1)}{SSR_1/(n-3-1)}=99.8667>F_{0.975}(8,8)=4.4333$$

因此拒绝原假设，认为该模型存在异方差。

表 3.7 各地区指标值（一）

单位：万元

地区	人均生产总值（y）	人均消费支出(x_1)	人均第三产业增加值(x_2)	人均对外进出口总值(x_3)
北京	16.49	3.89	13.83	6.94
天津	10.16	2.85	6.54	4.39
河北	4.85	1.8	2.51	2.71
山西	5.06	1.57	2.59	2.52
内蒙古	7.22	1.98	3.52	3.15
辽宁	5.9	2.07	3.15	3.27
吉林	5.11	1.73	2.67	3.34
黑龙江	4.3	1.71	2.13	2.49
上海	15.56	4.25	11.38	7.22
江苏	12.12	2.62	6.37	4.34
浙江	10.01	3.13	5.58	5.24
安徽	6.34	2.27	3.25	2.81
福建	10.57	2.51	5.02	3.72
江西	5.69	1.8	2.74	2.8
山东	7.2	2.09	3.86	3.29
河南	5.53	1.61	2.69	2.48
湖北	7.52	1.92	3.86	2.79
湖南	6.29	2.1	3.25	2.94
广东	8.79	4.1	4.96	2.85
广西	4.42	2.46	2.29	1.64
海南	5.49	2.79	3.31	2.79
重庆	7.8	3.08	4.12	2.17
四川	5.81	2.65	3.04	2.65
贵州	4.62	2.18	2.35	1.49
云南	5.19	2.33	2.68	2.33
西藏	5.22	2.17	2.61	1.32
陕西	6.62	2.62	3.18	1.74
甘肃	35.93	2.03	19.79	1.62
青海	5.07	2.4	2.58	1.83
宁夏	5.44	2.57	2.74	2.57
新疆	5.34	2.38	2.74	1.65

3. White 检验

使用 White 检验对数据进行异方差检验，计算求得其检验统计量为 $\chi^2 = 16.9144 < \chi^2_{0.95}(9) = 16.9190$，故不拒绝原假设，认为该模型不存在异方差。

4. M-G-Q 检验

使用 M-G-Q 检验对数据进行异方差检验，按照本章提出的的 M-G-Q 检验方法，样本容量 $n = 31$，首先计算 $c = 7$，$n_1 = n_2 = \frac{n-c}{2} = 12$，用残差平方对各个解释变量进行如式 3.3 中的 OLS 回归，并对回归参数进行 t 检验，得出三个解释变量系数的检验统计量分别为：$t_1 = 0.3997$，$t_2 = 1.3857$，$t_3 = 2.2053$，由 t 检验的性质可得最显著的解释变量是 x_3，按 x_3 对样本点进行排序后，计算得出 $SSR_1 = 0.2994$，$SSR_2 = 11.6636$，F 统计量的自由度为 $n-3-1=8$，可以求得其检验统计量：

$$F_{M-G-Q} = \frac{SSR_2/(n-3-1)}{SSR_1/(n-3-1)} = 38.9547 > F_{0.975}(8,8) = 4.4333$$

因此拒绝原假设，认为该模型存在异方差。

通过对比 M-G-Q 检验与其他几种检验的结果，可以看出，M-G-Q 检验可以顺利实施，且比 White 检验的结果更灵敏，因此 M-G-Q 检验具有可行性。

3.4 本章小结

G-Q 检验是一元线性模型中常用的异方差检验方法，本文提出的 M-G-Q 检验是 G-Q 检验借鉴了 White 检验的思想，将 G-Q 检验推广至多元线性模型，并且与前人提出的几种改进的 G-Q 检验进行了详细的对比，论证了 M-G-Q 检验的灵敏度优于其他几种改进的 G-Q 检验。

第 4 章 一种新的 HCCMEs 估计方法

第 2 章介绍的几种改进的回归参数协方差矩阵估计方法都在不同程度上提高了估计的准确性，几种方法各有其优势，但也存在着一些不足，例如准确性不高、适用性较窄等问题。为解决这些问题，本章提出了一种新的改进估计方法 HCCv，并通过大量的模拟实验与前人提出的多种改进方法进行了比较分析，论证了在多种异方差情况下 HCCv 的估计效果比其他改进的效果更好，同时通过实例分析验证了新方法的可行性。

4.1 异方差一致性协方差矩阵的 HCCv 估计

在第 2 章中介绍了异方差一致性协方差矩阵估计（HCCMEs），并介绍了其他学者在 HC0 基础上做出的改进：HC1，HC2，HC3，HC4，HC5，HC4m，HC5m 和 HC6。这些改进都在一定程度上提高了 HC0 的估计效果，但仍存在各自的不足，本文基于以上研究，提出了一种准确率更高的 HCCMEs 估计。

MacKinnon 和 White 在文献中提到了通过 OLS 计算得到的残差比真实的随机扰动项偏小，Cribari-Neto 在其文章中阐述了 OLS 残差比真实扰动项偏小的程度与 $1-h_i$ 有关[40]，且其提出的 HC4、HC5 和 HC4m 估计量均利用 $(1-h_i)^{\delta_i}$ 在一定程度上对 HC0 进行了修正。但这三种估计量的修正因子 δ_i 取值方式较为固定，使得它们在不同的条件下表现各有优劣，不能适应不同的异方差情况，因此本文考虑添加一个能衡量异方差程度的参数，来提高修正因子对不同异方差情况的适应能力。

在回归模型设定正确的条件下，残差可以看作随机扰动项的观测值，而残差平方可以看作随机扰动项方差的观测值。我们所需要的衡量异方差程度大小的参数，正是衡量随机扰动项方差离散程度的参数。若采用残差平方的标准差来衡量随机扰动项方差的离散程度，则会受到变量值平均水平的影响，所以我们考虑使用残差平方的变异系数。变异系数是标准差与平均值的比值，是一个无量纲量。一般来说，变量值平均水平高，其离散程度的测度值越大，反之越小。用变异系数来衡量样本的离散程度，可以消除平均值和量纲对离散程度测度的影响。因此，用变异系数来衡量异方差程度是可行的。

HC4 中的δ_i可表示为$\delta_i=\min\{k_1,k_2h_i/\bar{h}\}$，$k_1=4,k_2=1$。我们基于此形式，在确定修正因子$\delta_i$的取值时添加了 OLS 残差平方的变异系数作为异方差程度的测度，设k_1,k_2的取值与c_v有关。

首先确定δ_i的上限值k_1，其作用是防止在某些点处δ_i取值过大，HC4m 中δ_i的上限值为2.5，HC5 中δ_i的上限值为$\max\{4,0.7h_{\max}/\bar{h}\}$。结合 HC4m 在异方差程度较小时估计效果较好而 HC4 和 HC5 在异方差程度较大时估计效果较好的特点可以发现，对于异方差程度较大的情形需要较大的上限值，对于异方差程度较小的情形需要较小的上限值。k_1太小时，δ_i对$\hat{\boldsymbol{\Omega}}_0$的放大程度不够，不足以使$\hat{\boldsymbol{\Psi}}$接近真实的$\boldsymbol{\Psi}$；当$k_1$太大时，对$\delta_i$的限制作用不明显，从而可能存在某些点处$h_i/\bar{h}$太大使$\delta_i$过大的情况，导致估计误差较大。因此我们选取$k_1$的准则是要使其尽可能的小，但其对应的$\delta_i$对$\hat{\boldsymbol{\Omega}}_0$的放大效果必须足以使$\hat{\boldsymbol{\Psi}}$接近真实的$\boldsymbol{\Psi}$。为了形式简明同时又能满足上述要求，我们设$k_1$为$c_v$的线性函数：$k_1=dc_v,d>0$。在随机扰动项同方差条件下残差平方的变异系数最小，最容易出现k_1偏小的情况。因此我们只要在同方差条件下使k_1作为δ_i对$\hat{\boldsymbol{\Omega}}_0$的放大程度满足条件即可。

令$\delta_i=k_1=dc_v$，d分别取$0.1,0.2,\cdots,3.0$，在本章第二节模型同方差情形下用计算机模拟 10000 次，分别得出各个d值下拒绝原假设的频率p（对照检验拒绝频率为 0.0495），如图 4.1 所示。可以得出，当$d=1.6$时k_1可以满足条件，即$k_1=1.6c_v$。

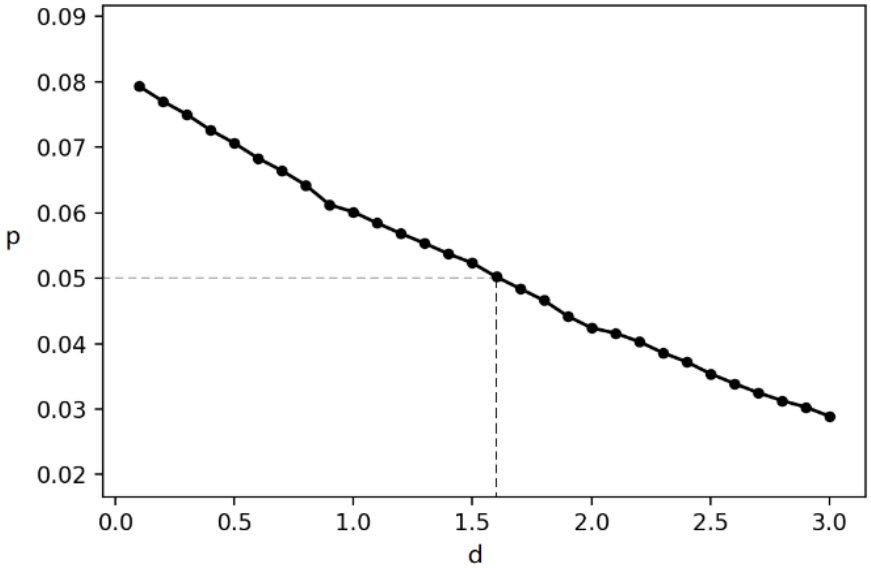

图 4.1　不同d取值下拒绝频率p的取值

然后确定 $h_i/\bar{h}$ 的系数 k_2。在 $k_1=1.6c_v$ 的条件下，设定不同的异方差形式使 c_v 取不同的值，在各个 c_v 不同的取值下寻找最佳的 k_2 值，然后将 k_2 与 c_v 拟合出一条曲线：$k_2=(4^{2.6-c_v}+0.5)$。如图 4.2 所示。

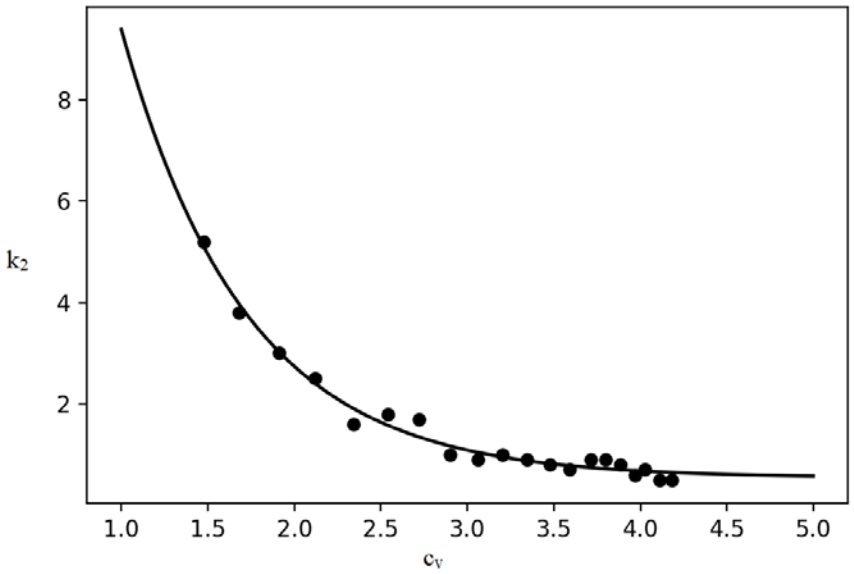

图 4.2　不同 c_v 值下最佳的 k_2 取值曲线

通过上述步骤，我们得到一个新的修正因子：$\delta_i=\min\{(4^{2.6-c_v}+0.5)h_i/\bar{h},1.6c_v\}$，其中 $c_v=\frac{s_e}{\bar{x}_{e^2}}$，$s_{e^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i^2-\frac{1}{n}\sum_{i=1}^{n}e_i^2)^2}$，$\bar{x}_{e^2}=\frac{1}{n}\sum_{i=1}^{n}e_i^2$，将其代入 HCs 的表达式便得到了新的估计量 HCCv.

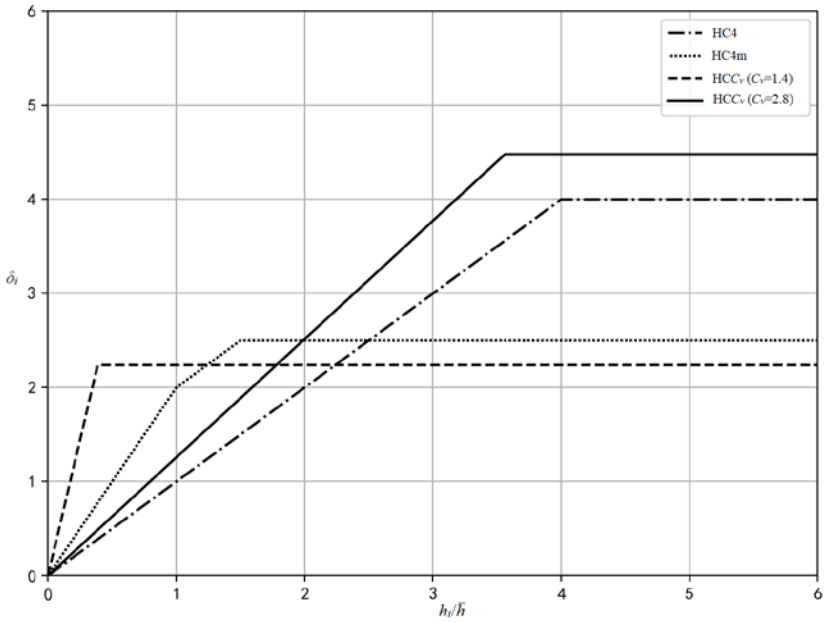

图 4.3　不同估计量的 δ_i 取值对比

HCCv 通过 OLS 残差平方的变异系数来衡量数据的异方差程度，故其修正因子δ_i的取值比其他的 HCs 估计量更为灵活。例如，在c_v=1.4 时，δ_i在低杠杆点（$h_i/\overline{h}$较小）处取值比 HC4m 大，在高杠杆点（$h_i/\overline{h}$较大）处取值比 HC4m 小；c_v=2.8 时，δ_i在低杠杆点处取值比 HC4m 小，在高杠杆点处取值比 HC4m 大，如图 4.3 所示。HCCv 的修正因子取值随异方差程度的不同而变化，从而可以适应各种不同的异方差情况。通过大量的数值模拟实验可以验证，HCCv 在多种不同的异方差情况下的估计效果优于上述其他估计量。

4.2 数值模拟

本节的模拟模拟分析通过 Python 实现，使用如下线性回归模型：

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\varepsilon_i,i=1,2,\cdots n$$

其中，n 为 25，50 或 100，$\beta_1=0$，$\beta_0=\beta_2=\beta_3=1$，解释变量$x_{i1},x_{i2},x_{i3}$互相独立且产生自正态分布$N(0,1)$，$\varepsilon_i$产生自正态分布$N(0,\sigma_i^2)$，为了详细对比上述各估计量的估计效果，本文模拟多种不同的异方差情况，σ_i^2的取值有以下几种类型：

（a）$\sigma_i^2=|x_{i1}|^{c_1}$，$c_1=0,1,2,4,6$；

（b）$\sigma_i^2=e^{|c_2x_{i1}|}$，$c_2=1,2,3$；

（c）$\sigma_i^2=\ln(|x_{i1}|+1)$，$\sigma_i^2=\lg(|x_{i1}|+1)$.

然后使用本章中提到的不同的参数协方差矩阵估计量对 OLS 估计量$\hat{\beta}_1$进行原假设为$\hat{\beta}_1=0$，备择假设为$\hat{\beta}_1\neq0$的t检验，同时使用真实方差（$\boldsymbol{\Psi}$）的检验结果作为参照，以对比不同的方差估计量的检验效果。重复实验 10000 次并记录显著性水平$\alpha=0.05$时的拒绝次数。结果如表 4-1 到表 4-3 所示。

（a）$\sigma_i^2=|x_{i1}|^{c_1}$的情况，$c_1=0,1,2,4,6$

表 4-1 展示了异方差与x_1成幂次关系时用各个方差估计量进行假设检验在 10000 次模拟中的拒绝次数，通过表 4-1 可以得出如下结论：

从检验结果来看，参照组的结果准确度十分良好，检验的拒绝频率始终很接近显著性水平 0.05，可见使用真实方差对参数的显著性进行 t 检验是十分可靠的，因此可以通过对比其他估计量与参照组的检验拒绝次数的接近程度来判断该估计

量的检验精确度。

表 4.1 异方差情况（*a*）下用各个方差估计量进行假设检验在 10000 次模拟中的拒绝次数

σ_i^2	n	OLS	HC0	HC3	HC4	HC5	HC4m	HC5m	HC6	HCCv	Ψ
1	25	614	1157	546	608	608	482	364	802	**518**	493
	50	575	859	538	567	567	497	427	669	**511**	503
	100	538	651	513	526	526	500	474	575	**498**	503
$\lvert x_{i1}\rvert$	25	1691	1461	710	789	789	611	523	1042	**624**	518
	50	1716	997	631	654	654	580	513	778	**559**	526
	100	1678	744	576	583	583	551	449	636	**525**	493
x_{i1}^2	25	2584	1707	794	843	843	673	548	1164	**646**	536
	50	2527	1011	614	612	612	551	513	759	**482**	488
	100	2566	773	573	562	562	539	520	644	**494**	504
x_{i1}^4	25	3716	1938	824	881	881	676	524	1238	**572**	533
	50	3691	1182	660	640	640	578	454	848	**523**	510
	100	3833	808	596	551	551	549	458	661	**549**	507
x_{i1}^6	25	4435	1935	691	739	739	556	411	1132	**487**	484
	50	4470	1095	585	548	548	494	395	773	**503**	519
	100	4566	729	475	443	443	420	366	550	**507**	502

在同方差条件下，使用基于 OLS 的协方差矩阵估计量的检验效果与参照组十分接近。当存在与解释变量成幂次关系的异方差时，基于 OLS 的协方差矩阵估计量的检验精确度很低。在 $\sigma_i^2 = x_{i1}^2$， $n=50$ 的情况下拒绝的次数接近参照组的 5 倍，在 $\sigma_i^2 = x_{i1}^6$， $n=100$ 的情况下拒绝的次数甚至达到参照组的 9 倍，此时，检验的结果已经严重偏离了真实情况。

在同方差条件下，基于 HC0 的检验精确度比 OLS 要低，在在 $\sigma_i^2=1$，$n=50$ 的情况下 HC0 的拒绝次数比 OLS 高出了 49.39%，在 $\sigma_i^2=1$， $n=100$ 的情况下 HC0 的拒绝次数比 OLS 高出了 21%，这也论证了在同方差条件下 HC0 的收敛速度比 OLS 要慢。在异方差情况下，基于 HC0 的检验精确度比 OLS 有所提高，并且在每种情况中均随着样本容量的增大，检验效果均越来越好，例如在 $\sigma_i^2 = \lvert x_{i1}\rvert$， $n=100$ 情况下基于 HC0 的检验拒绝次数为 744，仅为 $\sigma_i^2 = \lvert x_{i1}\rvert$， $n=25$ 时的一半，更接近参照组。但当样本容量较小时，HC0 的误差依然很大，在 $\sigma_i^2 = x_{i1}^6$， $n=25$ 的情况

下拒绝次数接近参照组的 4 倍，因此对其进行改进是很有必要的。

HC3 的表现较为稳定，其在同方差条件下与参照组的表现很接近。在表 4-1 所列的情况中，基于 HC3 的检验拒绝次数最高达到 824，最低达到 475，其精确度已经比 HC0 提高了很多，但其仍然不够精确，尤其异方差程度较大的情况下，例如，在 $\sigma_i^2 = x_{i1}^6$， $n = 25$ 的情况下拒绝次数接近参照组的 1.4 倍。

在表 4-1 中所列的情况中，HC4 和 HC5 表现完全相同，因为其都在低杠杆点处采用了较小的修正因子而在高杠杆点处采用了较大的修正因子，其表达式也较为接近。在同方差条件下，其表现与参照组接近。当异方差程度较低时，低杠杆点较多，其估计效果比 HC3 略差，例如在 $\sigma_i^2 = |x_{i1}|$， $n = 50$ 的情况下，基于 HC4 和 HC5 的检验拒绝次数均比 HC3 高出了 3.65%，当异方差程度增大时，高杠杆点也随之增多，其估计效果比 HC3 略好，例如在 $\sigma_i^2 = x_{i1}^6$， $n = 50$ 的情况下，其拒绝次数为 548，比 HC3 的 585 更接近参照组。

HC4m 在低杠杆点的修正因子较大而在高杠杆点的修正因子较小，故在异方差程度较低时估计效果较好，在同方差情况下，其检验效果与参照组十分接近。在 $\sigma_i^2 = |x_{i1}|$ 的情况下，基于 HC4m 的检验效果均比 HC3 更精确。当异方差程度增大时，其估计效果却逐渐变差，例如 $\sigma_i^2 = x_{i1}^6$， $n = 100$ 的情况下，基于 HC4m 的检验拒绝次数低至 420，比 HC3 的 475 离参照组更远。

基于 HC5m 的检验在多数情况下的拒绝次数比参照组偏小，存在高估参数方差的现象。而 HC6 与之相反，在多数情况下的拒绝次数比参照组偏大，存在低估参数方差的现象。基于 HC5m 和 HC6 的检验效果较之 HC0 均有所提高，但在多数情况下其精确度不如 HC4m。

HCCv 因为采用了更为灵活的修正因子，其在表 4-1 所列的大多数情况中表现均优于其他估计量，在同方差情况下的检验结果也十分精确，在异方差条件下，除了在 $\sigma_i^2 = |x_{i1}|$， $n = 25$ 和 $\sigma_i^2 = x_{i1}^2$， n=25 两种情况下的检验拒绝次数比参照组分别高出了 20.46%和 20.52%，其余情况均与参照组十分接近。故 HCCv 在异方差与解释变量成幂次关系时是一种精确度更高，泛用性更好的估计量。

（b） $\sigma_i^2 = e^{|c_2 x_{i1}|}$ 的情况， $c_2 = 1, 2, 3$

表 4-2 展示了异方差与 x_1 成指数关系时用各个方差估计量进行假设检验在

10000 次模拟中的拒绝次数，通过表 4-2 可以得出如下结论：

从检验结果来看，参照组的结果准确度依然十分良好，检验的拒绝频率始终很接近显著性水平 0.05，因此可以通过对比其他估计量与参照组的检验拒绝次数的接近程度来判断该估计量的检验精确度。

表 4.2 异方差情况（*b*）下用各个方差估计量进行假设检验在 10000 次模拟中的拒绝次数

σ_i^2	n	OLS	HC0	HC3	HC4	HC5	HC4m	HC5m	HC6	HCCv	Ψ
$e^{\lvert x_{i1}\rvert}$	25	1813	1614	757	824	824	651	481	1093	**708**	529
	50	1778	1004	629	631	631	572	476	772	**562**	516
	100	1847	755	582	578	578	551	469	653	**527**	506
$e^{\lvert 2x_{i1}\rvert}$	25	3099	1788	718	786	786	587	475	1135	**574**	514
	50	3265	1097	641	631	631	570	458	797	**505**	510
	100	3471	789	558	532	530	517	432	635	**498**	508
$e^{\lvert 3x_{i1}\rvert}$	25	4043	1930	692	759	759	572	453	1107	**505**	494
	50	4382	1071	576	564	564	495	404	732	**477**	516
	100	4524	705	479	440	440	435	337	550	**477**	500

当存在与解释变量成指数关系的异方差时，基于 OLS 的协方差矩阵估计量的检验精确度很低。在 $\sigma_i^2 = e^{\lvert 2x_{i1}\rvert}$，$n=50$ 的情况下检验的拒绝次数接近参照组的 6 倍，在 $\sigma_i^2 = e^{\lvert 3x_{i1}\rvert}$，$n=100$ 的情况下检验的拒绝次数达到了参照组的 9 倍，此时的检验的结果已经与真实情况偏差巨大。

基于 HC0 的检验精确度比 OLS 有明显提高，尤其在异方差程度较高的 $\sigma_i^2 = e^{\lvert 2x_{i1}\rvert}$ 和 $\sigma_i^2 = e^{\lvert 3x_{i1}\rvert}$ 情况下。并且在每种情况中均随着样本容量的增大，检验效果均越来越好，例如在 $\sigma_i^2 = e^{\lvert 3x_{i1}\rvert}$，$n=50$ 情况下，基于 HC0 的检验拒绝次数约是参照组的 2 倍，而在 $\sigma_i^2 = e^{\lvert 3x_{i1}\rvert}$，$n=100$ 情况下基于 HC0 的检验拒绝次数为 705，只比参照组高出了 41%，但 HC0 的误差依然较大，在 $\sigma_i^2 = e^{\lvert 2x_{i1}\rvert}$，$n=25$ 的情况下拒绝次数接近参照组的 3.5 倍，因此需要对其进行改进。

HC3 在存在与解释变量成指数关系的异方差的情况下的表现总体较好，比起 HC0 有显著提升。在表 4-2 所列的各种情况中，基于 HC3 的检验拒绝次数最高达到均比 HC0 更接近参照组，因此 HC3 的精确度比 HC0 高。但其在部分情况下仍然不够精确，依然有改进的空间。

基于 HC4 和 HC5 的检验拒绝次数几乎相同，且与 HC3 的表现较为接近，差距最大的情况是在 $\sigma_i^2 = e^{|3x_{i1}|}$， $n = 25$ 的情况下，基于 HC4 和 HC5 的检验拒绝次数均比 HC3 高出了 9.68%，比参照组高出了 53.64%，可见在存在与解释变量成指数关系的异方差时，HC4，HC5 与 HC3 的估计效果较为接近，均比 HC0 有较大提升。

HC4m 在表中所列的情况中表现较好，仅仅在异方差程度较高的情况中表现较差。在 $\sigma_i^2 = e^{|3x_{i1}|}$， $n = 100$ 的情况下，基于 HC4m 的检验拒绝次数比 HC3，HC4，HC5 略低，在其余情况均比它们更接近参照组。可见 HC4m 在异方差程度较低的情况下是一种较为准确的估计量，但其在异方差程度较高时准确度有所降低。

基于 HC5m 的检验仍然在多数情况下的拒绝次数比参照组偏小，存在高估参数方差的现象。HC6 与之相反，在多数情况下的拒绝次数比参照组偏大，存在低估参数方差的现象。基于 HC5m 和 HC6 的检验效果较之 HC0 均有所提高，但在多数情况下其精确度不如 HC4m。

HCCv 在存在与解释变量成指数关系的异方差情况下的表现依然十分优秀，其在表 4-2 所列的大多数情况中表现均优于其他估计量，仅在 $\sigma_i^2 = e^{|x_{i1}|}$， $n = 25$ 情况下的检验拒绝次数比参照组高出了 33.84%，其余情况均与参照组十分接近。故 HCCv 在异方差与解释变量成指数关系时的精确度和泛用性依然很高。

（c） $\sigma_i^2 = \ln(|x_{i1}|+1)$ 和 $\sigma_i^2 = \lg(|x_{i1}|+1)$ 的情况

表 4.3 异方差情况（c）下用各个方差估计量进行假设检验在 10000 次模拟中的拒绝次数

σ_i^2	n	OLS	HC0	HC3	HC4	HC5	HC4m	HC5m	HC6	HCCv	Ψ
$\ln(\|x_{i1}\|+1)$	25	1406	1374	653	721	721	564	478	943	**594**	493
	50	1454	958	613	641	641	573	470	753	**554**	513
	100	1416	773	596	602	602	557	516	653	**531**	504
$\lg(\|x_{i1}\|+1)$	25	1464	1455	715	792	792	625	466	1044	**647**	502
	50	1393	932	596	611	611	554	477	745	**537**	501
	100	1395	740	578	586	586	558	506	642	**544**	515

表 4-3 展示了异方差与 x_1 成对数关系时用各个方差估计量进行假设检验在 10000 次模拟中的拒绝次数，通过表 4-3 可以得出如下结论：

从检验结果来看，参照组的检验拒绝频率仍然始终很接近显著性水平 0.05，因此可以通过对比其他估计量与参照组的检验拒绝次数的接近程度来判断该估计

量的检验精确度。

当存在与解释变量成对数关系的异方差时，基于 OLS 的协方差矩阵估计量的检验精确度很低。在 $\sigma_i^2=\ln(|x_{i1}|+1)$ 和 $\sigma_i^2=\lg(|x_{i1}|+1)$ 的情况中，检验的拒绝次数均接近参照组的 2.8 倍，此时的检验的结果与真实情况偏差很大。

在样本容量较小的情况下，HC0 的精确度与 OLS 相当，在样本容量较大的情况下，基于 HC0 的检验精确度比 OLS 有明显提高。并且在每种情况中均随着样本容量的增大，检验效果均越来越好，在 $\sigma_i^2=\ln(|x_{i1}|+1)$ 和 $\sigma_i^2=\lg(|x_{i1}|+1)$ 的情况中，样本容量 $n=100$ 情况下的检验拒绝次数均比 $n=25$ 情况下低了近一半，更接近参照组。但 HC0 的误差依然较大，在 $\sigma_i^2=\lg(|x_{i1}|+1)$，$n=25$ 的情况下拒绝次数接近参照组的 2.9 倍，因此需要对其进行改进。

HC3 在存在与解释变量成对数关系的异方差的情况下的表现总体较好，比起 HC0 有显著提升。在表 4-3 所列的各种情况中，基于 HC3 的检验拒绝次数均比参照组偏大 20%到 40%，仍然不够准确，可以进行进一步的改进。

基于 HC4 和 HC5 的检验拒绝次数完全相同，且均比参照组偏大，比起 HC3 离参照组的距离更远，差距最大的情况在 $\sigma_i^2=\lg(|x_{i1}|+1)$，$n=25$ 的情况下，基于 HC4 和 HC5 的检验拒绝次数均比参照组高出了 57.77%，可见在存在与解释变量成对数关系的异方差时，HC3，HC4，HC5 对参数的方差估计都偏小，导致检验拒绝次数偏高。

HC4m 在表中所列的情况中表现均比 HC3 好，在各个情况下，基于 HC4m 的检验拒绝次数比 HC3，HC4，HC5 低，更接近参照组，但比起参照组仍然偏大，差距最大的情况是在 $\sigma_i^2=\lg(|x_{i1}|+1)$，$n=25$ 的情况下，检验的拒绝次数比参照组高出 24.5%，因此 HC4m 不失为一种较好的估计，但仍有改进空间。

基于 HC5m 的检验在多数情况下的拒绝次数与参照组十分接近，与参照组差距最大的情况是在 $\sigma_i^2=\ln(|x_{i1}|+1)$，$n=50$ 的情况下，检验的拒绝次数比参照组低 8.38%，说明了 HC5m 在存在与解释变量成对数关系的异方差情况下表现较好。

基于 HC6 的检验拒绝次数比参照组偏大，其精确度不如 HC4m。

HCCv 在存在与解释变量成对数关系的异方差情况下的表现仍然较好，其与参照组差距最大的情况是在 $\sigma_i^2=\lg(|x_{i1}|+1)$，$n=25$ 的情况下，检验的拒绝次数比参

照组高出了 28.88%，在表 4-3 所列的大多数情况中表现均优于除 HC5m 之外的其他估计量，而且其在多数情况下与 HC5m 和参照组十分接近。故 HCCv 在异方差与解释变量成对数关系时的精确度仍然很高，充分说明了其具有较高的稳定性和良好的泛用性。

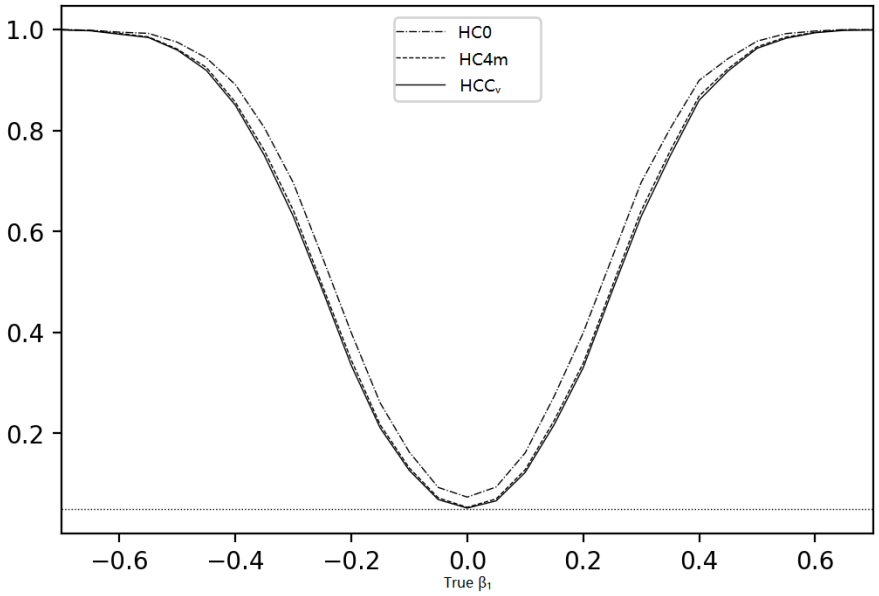

图 4.4　在 $\sigma_i^2=\left|x_{i1}\right|$ 异方差情况下不同估计量的势

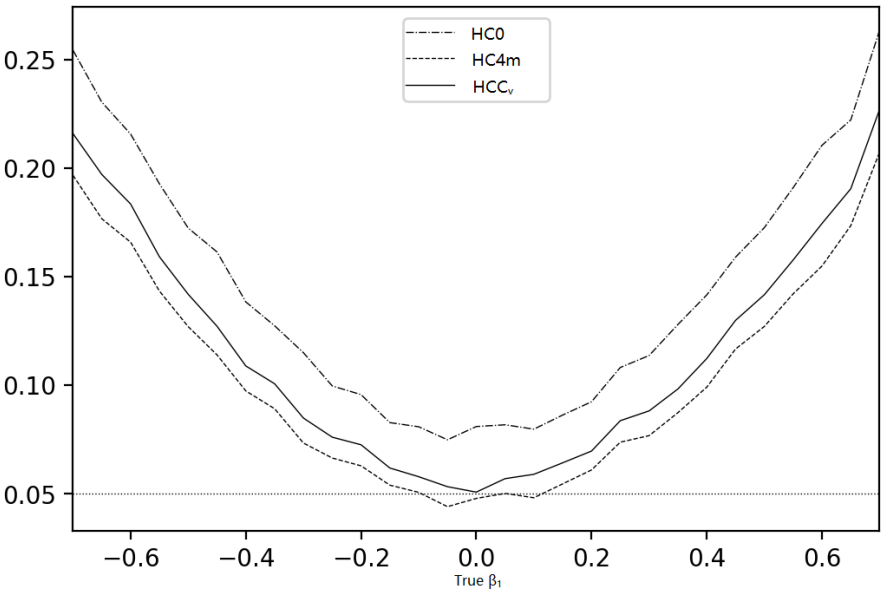

图 4.5　在 $\sigma_i^2=\left|x_{i1}\right|^6$ 异方差情况下不同估计量的势

本文还比较了在不同的 β_1 真实值情况下 HC0，HC4m 和 HCCv 的势。在 $n=100$，显著性水平 $\alpha=0.05$ 的条件下，不断改变 β_1 的真实值，并在不同的 β_1 真

实值情况下重复 10000 次实验，得到在该情况下的对原假设 β_1=0 的拒绝频率。当 β_1=0 时，检验的势越接近显著性水平，检验的就越好。当 $\beta_1 \neq 0$ 时，检验的势越大，犯第二类错误的概率也就越小。

图 4.4 和图 4.5 显示了当异方差程度较小和较大时，使用 HC0，HC4m 和 HCCv 进行检验的势。可以看出，当异方差程度较低时，使用 HCCv 和 HC4m 的检验效果非常接近，而且它们在 β_1=0 的情况下的势都比 HC0 更接近显著性水平 α，因而有更好的效果，但在 $\beta_1 \neq 0$ 的情况下的表现均略差于 HC0。当异方差程度较高时，使用 HCCv 的检验在 β_1=0 的势最接近显著性水平 α，所以在该情况下使用 HCCv 的检验效果最好，且在 $\beta_1 \neq 0$ 的情况下使用 HCCv 估计量进行检验的势高于 HC4m，此时使用 HCCv 的检验效果也优于 HC4m。

4.3 实例分析

本节将通过实例来验证本章提出的 HCCv 估计方法的可行性。仍然采用 3.3 中的例子，从各地区的统计局官网上获取 2021 年 31 个省份（不含港澳台）的人均生产总值(y)，人均消费支出(x_1)，人均第三产业增加值(x_2)，人均对外进出口总值(x_3)，该数据由该项经济指标总值除以当地总人口得到。

表 4.4 各个方差估计量进行 *t* 检验的结果

方差估计量	t	拒绝情况
OLS	-1.8002	不拒绝原假设
HC0	-2.6627	拒绝原假设
HC1	-2.5306	拒绝原假设
HC2	-1.7471	不拒绝原假设
HC3	-0.6332	不拒绝原假设
HC3	-0.0412	不拒绝原假设
HC5	-0.0090	不拒绝原假设
HC4m	-0.3284	不拒绝原假设
HC5m	-0.0022	不拒绝原假设
HC6	-0.7115	不拒绝原假设
HCCv	-0.05678	不拒绝原假设

模型设定如下：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad i = 1, 2, \cdots, 31.$$

使用 31 省份的数据估计出的模型为：

$$\hat{y}_i = 4.145259 - 0.6877 x_{i1} + 1.6837 x_{i2} - 0.6742 x_{i3}, \quad i = 1, 2, \cdots, 31.$$

然后使用本章中提到的不同的方差估计量对 OLS 估计量 $\hat{\beta}_1$ 进行原假设为 β_1=0，备择假设为 $\beta_1 \neq 0$ 的 t 检验，并比较检验结果，各方法得到的 t 统计量如下：

通过对比 HCCv 估计量与其他几种估计量的 t 检验结果，可以看出，HCCv 估计的表现符合预期，比除 HC5m 之外的其他估计量的 p 值都要大，更不易产生过度拒绝原假设的情况，因此 HCCv 估计具有可行性。

4.4 本章小结

HCCMEs 是一种使用广泛的参数协方差矩阵估计方法。然而在现实应用中，样本容量经常不够大，导致 White 最初提出的 HC0 存在较大误差，从而影响参数检验的准确性，很多学者对其做出了改进，提高了其在某些情况中的估计准确性。本文提出的 HCCv 估计将样本残差平方的变异系数融入到 HCCMEs 的表达式中，大大提高了估计的准确性与适用范围，从而可以提高参数检验的精度，并且与前人提出的几种改进的 HCCMEs 估计进行了详细的对比，论证了 HCCv 的准确性在多数情形下优于其他几种改进的 HCCMEs 估计。

第 5 章 总结与展望

5.1 总结

线性回归模型是统计学中使用最为广泛的模型，然而异方差的存在对线性回归模型中参数的显著性检验、结果的预测等问题的准确度造成了一定的影响，本文主要研究线性回归模型中的异方差相关问题，从异方差的检验和异方差模型下的参数协方差矩阵估计两条线路进行研究。首先通过大量阅读文献，研究已有的异方差检验方法和参数协方差矩阵估计方法；接着，通过对已有的方法做出有效的改进，得到新的异方差检验方法和参数协方差矩阵估计方法；然后设计数值模拟实验，使用 Python 实现了新方法和一些文献中已有的方法并进行对比，以论证新方法的有效性；最后，使用实际数据验证了新方法的可行性。

G-Q 检验是一种常用的异方差检验方法，但其仅适用于一元情形，很多学者对其进行了改进以适应多元情况，但仍存在准确率较低的缺陷，本文针对这些改进的不足，提出一种新的 M-G-Q 检验，通过借鉴 White 检验的思想，对残差平方与所有解释变量进行 OLS 回归，找出参数 p 值最小的解释变量，进而进行 G-Q 检验，提高了检验的准确率。在实际应用中，可以更精准地识别出模型中的异方差现象，从而做出进一步的处理，提高模型的准确性。

为了解决模型存在异方差情况下的参数协方差矩阵估计问题，White 提出了异方差一致性协方差矩阵估计（记 White 提出的估计量为 HC0），即使模型存在异方差，也可以据此得出参数方差的一致估计。然而，HC0 的收敛的速度并不快，对样本容量的要求较高，在实际问题中经常遇到样本容量较小的情形。因此，很多学者对 White 提出的 HC0 进行了改进，提出了很多改进的估计量，这些估计量在 HC0 中增加了系数和修正因子，提升了 HC0 的性能，但也存在着准确率不够高、适用范围不够广等问题。本文提出了新的 HCCv 估计，通过在修正因子中添加 OLS 残差平方的变异系数作为异方差程度的测度，并通过数值拟合的方法确定修正因子的表达式，来提升估计量在不同异方差情况的适应性，从而获得更准确的参数协方差矩阵估计。

5.2 展望

本文基于已有的异方差检验、参数协方差矩阵估计相关研究，进一步改进了异方差的检验方法和参数协方差矩阵的估计方法，提高了异方差检验和参数协方差矩阵估计的准确率，属于方法上的创新，并为解决线性回归模型中的异方差问题提供了新的思路。本文设计了大量的数值模拟实验，从多角度对不同的异方差检验和参数协方差矩阵估计方法进行了探究，对其他学者的进一步研究具有一定的参考价值。

但本文也存在一些不足之处：一是对新方法的评价指标单一，二是缺少理论证明。在未来的研究中，可以从多角度设计异方差的检验和参数协方差矩阵估计的评价指标，进一步提高数值模拟实验的说服力；可以从理论角度对新方法进行证明，从而提升结论的严谨性。

附　录

附录一 M-G-Q 检验数值模拟 Python 代码

```
import numpy as np
import math
from sklearn import datasets,linear_model
import statsmodels
from scipy.stats import f,chi2

c,d1,d2,d3,d4 = 0,0,0,0,0
r = 0
num_s = []
b0,b1,b2,b3,b4 = 1,1,1,1,1
#b0,b1,b2,b3,b4 = 10,10,10,10,10

test_number = 1000
n = 50
n1 = n*3//8
fr = n1-5

def findbiggest(arr):
    biggest = arr[0] #存储最小的值
    biggest_index = 0 #存储最小元素的索引
    for i in range(1,len(arr)):
        if arr[i] > biggest:
            biggest = arr[i]
            biggest_index = i
    return biggest_index
#  数据中心化(求平均值以及各个特征的样本数据与平均值之差)
def zeroMean(dataMat):
```

```
    # 按列求平均，即各个特征的平均
    meanVal = np.mean(dataMat, axis=0)
    newData = dataMat - meanVal
    return newData, meanVal

# 计算协方差的函数
def xcov(data):
    meanVal = np.mean(data, axis=0)
    newData = data - meanVal
    L = np.array(np.zeros((newData.shape[1], newData.shape[1])))
    for i in range(len(newData)):
        rowData = newData[i, :]
        rowData = rowData[:, np.newaxis]
        L = L + np.dot(rowData, rowData.T)
    return L / (len(newData) - 1)
# GQ
def gq(xy):
    xy_s = xy[np.lexsort(xy.T)]
    #print(xy_s)
    x_1 = xy_s[:n1,:4]
    y_1 = xy_s[:n1,4]
    x_2 = xy_s[-n1:,:4]
    y_2 = xy_s[-n1:,4]
    x_1_1 = x_1[:,0]
    x_1_2 = x_1[:,1]
    x_1_3 = x_1[:,2]
    x_1_4 = x_1[:,3]
    x_2_1 = x_2[:,0]
    x_2_2 = x_2[:,1]
    x_2_3 = x_2[:,2]
    x_2_4 = x_2[:,3]

    regr1 = linear_model.LinearRegression()
```

```
    regr1.fit(x_1,y_1)
    b1_hat_1 = regr1.coef_[0]
    b2_hat_1 = regr1.coef_[1]
    b3_hat_1 = regr1.coef_[2]
    b4_hat_1 = regr1.coef_[3]
    y_hat_1 = b1_hat_1*x_1_1 + b2_hat_1*x_1_2 + b3_hat_1*x_1_3 + b4_hat_1*x_1_4 + regr1.intercept_
    ssr1 = np.sum((y_1-y_hat_1)**2)

    regr2 = linear_model.LinearRegression()
    regr2.fit(x_2,y_2)
    b1_hat_2 = regr2.coef_[0]
    b2_hat_2 = regr2.coef_[1]
    b3_hat_2 = regr2.coef_[2]
    b4_hat_2 = regr2.coef_[3]
    y_hat_2 = b1_hat_2*x_2_1 + b2_hat_2*x_2_2 + b3_hat_2*x_2_3 + b4_hat_2*x_2_4 + regr2.intercept_
    ssr2 = np.sum((y_2-y_hat_2)**2)

    F0 = ssr2/ssr1
    return F0
def cm(F,fr):
    if F>=f.ppf(0.975,dfn=fr,dfd=fr):
        m = 1
    elif F<=f.ppf(0.025,dfn=fr,dfd=fr):
        m = -1
    else:
        m = 0
    return m
for i in range(test_number):

    #生成数据
    #mean = [2,2,2,2]
```

```
mean = [0,0,0,0]
cov = np.array([[ 1, r, r ,r],
                [ r, 1, r ,r],
                [ r, r, 1 ,r],
                [ r, r, r ,1]])
x = np.random.multivariate_normal(mean, cov,n)
#x = np.abs(np.random.multivariate_normal(mean, cov,n))
x0 = np.ones(n)
x1 = x[:,0]
x2 = x[:,1]
x3 = x[:,2]
x4 = x[:,3]

a = 0
v = []
u = []
while (a<n):
    #va = 1 #方差
    va = np.abs(np.log((x1[a]+x2[a])**2))
    #va = np.abs(np.log((x1[a])**2))
    #va = (x1[a])**2
    #va = math.pow(np.e,x1[a])
    #va = math.pow(np.e,-x1[a])
    #va = math.pow(np.e,x1[a]-x2[a])
    #va = math.pow(np.e,x1[a]+x2[a]+x3[a]+x4[a])
    #va = math.pow(np.e,0.5*x1[a]+0.5*x2[a]+0.5*x3[a]+0.5*x4[a])
    #va = 0.5*math.pow(np.e,x1[a])+0.2*math.pow(np.e,x2[a])
    #va =
0.4*math.pow(np.e,x1[a])+0.3*math.pow(np.e,x2[a])+0.2*math.pow(np.e,x3[a])+0.1*
math.pow(np.e,x4[a])
    #va = (x1[a])**2+(x2[a])**2
    #va = (x1[a])**2+(x2[a])**2+(x3[a])**2
    ua = np.random.normal(0, np.sqrt(va), 1) #回归扰动项
```

```
        v = np.r_[v,va]    #真实方差
        u = np.r_[u,ua]    #扰动项
        a = a+1

    y = b0*x0 + b1*x1 + b2*x2 + b3*x3 + b4*x4 + u

    #  改进的 GQ 检验
    regr0 = linear_model.LinearRegression()
    regr0.fit(x,y)
    b1_hat = regr0.coef_[0]
    b2_hat = regr0.coef_[1]
    b3_hat = regr0.coef_[2]
    b4_hat = regr0.coef_[3]
    y_hat = b1_hat*x1 + b2_hat*x2 + b3_hat*x3+    b4_hat*x4 + regr0.intercept_
    e = y-y_hat
    e2 = e**2

    #找系数最显著的 x
    xr = x
    regr0_1 = linear_model.LinearRegression()
    regr0_1.fit(xr,e2)
    gamma1_hat = regr0_1.coef_[0]
    gamma2_hat = regr0_1.coef_[1]
    gamma3_hat = regr0_1.coef_[2]
    gamma4_hat = regr0_1.coef_[3]
    gamma_hat = np.array([])
    e2_hat  =  gamma1_hat*xr[:,0]  +   gamma2_hat*xr[:,1]  +  gamma3_hat*xr[:,2]  +
gamma4_hat*xr[:,3] + regr0_1.intercept_
    ef = e2 - e2_hat

    xe = np.c_[x0,xr]
    xe_t = np.transpose(xe)
    q = xe_t@xe
```

```
q_i = np.linalg.inv(q)
#OLS
w0 = np.sum(ef**2)/(n-5)
vm0 = w0*q_i #b_hat 的协方差矩阵
t1 = gamma1_hat/np.sqrt(vm0[1,1])
t2 = gamma2_hat/np.sqrt(vm0[2,2])
t3 = gamma3_hat/np.sqrt(vm0[3,3])
t4 = gamma4_hat/np.sqrt(vm0[4,4])
t = np.array([t1,t2,t3,t4])
num = findbiggest(np.abs(t))
#print(np.abs(t))
num_s.append(num)
x_best = x[:,num]
#print(x_best)

#PCA
#data = np.c_[xr,y]
data = x
newData, meanVal = zeroMean(data)
covMat = np.cov(newData, rowvar=0)
# 利用 Numpy 求特征值和特征向量
eigValue, eigVectors = np.linalg.eig(np.mat(covMat))
# 对特征值从小到大排序，返回索引
eigValueIndex = np.argsort(eigValue)
# 确定降维后的维度
dim = 1
# 取最大的 dim 个特征值的下标
n_eigValueIndexs = eigValueIndex[-1:-(dim + 1):-1]
# 取最大的特征值对应的特征向量
n_eigVectors = eigVectors[:, n_eigValueIndexs]
# 得到降维后低维特征空间的数据
lowDataMat = newData @ n_eigVectors
```

```
    # 利用低维数据重构数据(将降维的步骤反向执行)
    reconMat = (lowDataMat @ n_eigVectors.T) + meanVal

    xy1 = np.c_[x,y,x_best]
    xy2 = np.c_[x,y,np.ravel(lowDataMat)]
    xy3 = np.c_[x,y,y_hat]
    xy4_1 = np.c_[x,y,x1]
    xy4_2 = np.c_[x,y,x2]
    xy4_3 = np.c_[x,y,x3]
    xy4_4 = np.c_[x,y,x4]

    F1 = gq(xy1)
    F2 = gq(xy2)
    F3 = gq(xy3)
    F4_1 = gq(xy4_1)
    F4_2 = gq(xy4_2)
    F4_3 = gq(xy4_3)
    F4_4 = gq(xy4_4)

    m1 = cm(F1,fr)
    m2 = cm(F2,fr)
    m3 = cm(F3,fr)
    if m1!=0:
        c = c+1
    if m2!=0:
        d1 = d1+1
    if m3!=0:
        d2 = d2+1
    #white 检验
    x_white                                                              =
np.c_[x,x1**2,x2**2,x3**2,x4**2,x1*x2,x1*x3,x1*x4,x2*x3,x2*x4,x3*x4]
    regr0_2 = linear_model.LinearRegression()
    regr0_2.fit(x_white,e2)
```

```
        r2 = regr0_2.score(x_white,e2)
        #print(r2)
        if (n-14-1)*r2>chi2.ppf(0.95,df=14):
            d4 = d4+1
# print('r=',r)
print('模拟次数',test_number)
print(' ')
print('PCA 方法检验出异方差次数',d1)
print('刘明改进检验出异方差次数',d2)
print('White 检验出异方差次数',d4)
print('M-G-Q 检验出异方差次数',c)

##不同 x 被选中的次数
num0 = str(num_s).count("0")
print(num0)
num0 = str(num_s).count("1")
print(num0)
num0 = str(num_s).count("2")
print(num0)
num0 = str(num_s).count("3")
print(num0)
```

附录二 HCCv 估计数值模拟 Python 代码

```
import numpy as np
import math
from sklearn import datasets,linear_model

c,c00,c0,c1,c2,c3,c4,c5,c4m,c5m,c6,c7 = 0,0,0,0,0,0,0,0,0,0,0,0
cx1 = 0
gamma_sum,gamma2_sum,cv1_sum,cv2_sum,ratio_sum = 0,0,0,0,0
b1 = 0
b0,b2,b3 = 1,1,1
```

```
n = 50
critical_value = 1.96
test_number = 10000
for i in range(test_number):
    #b1t = 2*np.random.random(1)
    x0 = np.ones(n)
    x1 = np.random.normal(0, 1, n)
    x2 = np.random.normal(0, 1, n)
    x3 = np.random.normal(0, 1, n)

    #构造异方差
    a = 0
    u = []
    v = []
    while (a<n):

        va = 1
        #va = np.abs(x1[a])
        #va = (x1[a])**2
        #va = (x1[a])**4
        #va = (x1[a])**6
        #va = (x1[a])**8 #x
        #va = math.pow(np.e,1*x1[a])
        #va = math.pow(np.e,2*x1[a])
        #va = math.pow(np.e,3*x1[a])
        #va = math.pow(np.e,np.abs(x1[a]))
        #va = math.pow(np.e,np.abs(2*x1[a]))
        #va = math.pow(np.e,np.abs(3*x1[a]))
        #va = math.pow(np.e,np.abs(4*x1[a]))
        #va = math.pow(np.e,np.abs(5*x1[a]))
        #va = np.log(np.abs(x1[a])+1)
        #va = np.log10(np.abs(x1[a])+1)
```

```
        #va = 1/(np.abs(x1[a])+0.1)
        #va = (x1[a])**2+(x2[a])**2
        #va = (x1[a])**2+(x2[a])**2+(x3[a])**2
        ua = np.random.normal(0, np.sqrt(va), 1)
        v = np.r_[v,va]   #真实方差
        u = np.r_[u,ua]   #扰动项
        a = a+1
xr = np.c_[x1,x2,x3]
y = b0+b1*x1+b2*x2+b3*x3+u
#回归
regr = linear_model.LinearRegression()
regr.fit(xr,y)
b0_hat = regr.intercept_
b1_hat = regr.coef_[0]
b2_hat = regr.coef_[1]
b3_hat = regr.coef_[2]
y_hat = b0_hat+b1_hat*x1+b2_hat*x2+b3_hat*x3
e = y-y_hat
x = np.c_[x0,x1,x2,x3]
#计算参数
x_t = np.transpose(x)
q = x_t@x
q_i = np.linalg.inv(q)
hm = x@q_i@x_t #投影矩阵
im = np.identity(n)
hd = np.diag(hm)
md = np.diag(im-hm)
hd_sum = np.sum(hd)
e2 = e**2
ea = np.abs(e)
e2_sum = np.sum(e**2)
e_std = np.std(e**2,ddof=0)
```

```
e_std2 = np.std(e**2,ddof=1)
cv1 = e_std/np.mean(e**2)
cv2 = e_std2/np.mean(e**2)
cv1_sum = cv1_sum+cv1
cv2_sum = cv2_sum+cv2
ratio = max(e2)/np.mean(e**2)
ratio_sum = ratio_sum+ratio

e2_sort = np.sort(e2)
#e2_s1 = e2_sort[:n//2]
e2_s2 = e2_sort[-(n//2):]
gamma = np.sum(e2_s2)/np.sum(e2)
#print(gamma)
e2_s4 = e2_sort[-(n//4):]
gamma2 = np.sum(e2_s4)/np.sum(e2)
gamma_sum = gamma_sum + gamma
gamma2_sum = gamma2_sum + gamma2


##检验 OLS
w00 = e2_sum/(n-4)
vm00 = w00*q_i #b_hat 的协方差矩阵
sigma00 = math.sqrt(vm00[1,1])
t00 = (b1_hat-b1)/sigma00
if (t00>critical_value)or(t00<-critical_value):
      c00 = c00+1

##检验  HC0
w0 = np.diag((e**2))
vm0 = q_i@x_t@w0@x@q_i #b_hat 的协方差矩阵
sigma0 = math.sqrt(vm0[1,1])
t0 = (b1_hat-b1)/sigma0
if (t0>critical_value)or(t0<-critical_value):
```

```
        c0 = c0+1
##检验  HC1
w1 = np.diag((e**2)*n/(n-3))
vm1 = q_i@x_t@w1@x@q_i
sigma1 = math.sqrt(vm1[1,1])
t1 = (b1_hat-b1)/sigma1
if (t1>critical_value)or(t1<-critical_value):
        c1 = c1+1

##检验  HC2
w2 = np.diag((e**2)/md)
vm2 = q_i@x_t@w2@x@q_i
sigma2 = math.sqrt(vm2[1,1])
t2 = (b1_hat-b1)/sigma2
if (t2>critical_value)or(t2<-critical_value):
        c2 = c2+1

##检验  HC3
w3 = np.diag((e**2)/(md**2))
vm3 = q_i@x_t@w3@x@q_i
sigma3 = math.sqrt(vm3[1,1])
t3 = (b1_hat-b1)/sigma3
if (t3>critical_value)or(t3<-critical_value):
        c3 = c3+1

##检验  HC4
a = 0
md4 = []
while (a<n):
        delta = min(n*hd[a]/hd_sum,4)
        md4 = np.r_[md4,md[a]**delta]
        a = a+1
        #print(delta)
```

```
w4 = np.diag((e**2)/(md4))
vm4 = q_i@x_t@w4@x@q_i
sigma4 = math.sqrt(vm4[1,1])
t4 = (b1_hat-b1)/sigma4
if (t4>critical_value)or(t4<-critical_value):
    c4 = c4+1

##检验 HC5
a = 0
hmax = max(hd)
md5 = []
while (a<n):
    delta = min(n*hd[a]/hd_sum,max(4,0.7*n*hmax/hd_sum))
    md5 = np.r_[md5,md[a]**delta]
    a = a+1
    #print(delta)
w5 = np.diag((e**2)/(md5))
vm5 = q_i@x_t@w5@x@q_i
sigma5 = math.sqrt(vm5[1,1])
t5 = (b1_hat-b1)/sigma5
if (t5>critical_value)or(t5<-critical_value):
    c5 = c5+1

##检验 HC4m
a = 0
md4m = []
while (a<n):
    delta = min(1,n*hd[a]/hd_sum)+min(1.5,n*hd[a]/hd_sum)
    md4m = np.r_[md4m,md[a]**delta]
    a = a+1
    #print(delta)
w4m = np.diag((e**2)/(md4m))
vm4m = q_i@x_t@w4m@x@q_i
```

```
    sigma4m = math.sqrt(vm4m[1,1])
    t4m = (b1_hat-b1)/sigma4m
    if (t4m>critical_value)or(t4m<-critical_value):
        c4m = c4m+1
    ##检验 HC5m
    a = 0
    md5m = []
    hmax = max(hd)
    while (a<n):
        delta =
min(1,n*hd[a]/hd_sum)+min(n*hd[a]/hd_sum,max(4,0.7*n*hmax/hd_sum))
        md5m = np.r_[md5m,md[a]**delta]
        a = a+1
        #print(delta)
    w5m = np.diag((e**2)/(md5m))
    vm5m = q_i@x_t@w5m@x@q_i
    sigma5m = math.sqrt(vm5m[1,1])
    t5m = (b1_hat-b1)/sigma5m
    if (t5m>critical_value)or(t5m<-critical_value):
        c5m = c5m+1

    ##检验 HC6
    a = 0
    md6 = []
    hmax = max(hd)
    while (a<n):
        delta = min(n*hd[a]/hd_sum,np.sqrt(n*hmax/(2*hd_sum)))
        md6 = np.r_[md6,md[a]**delta]
        a = a+1
        #print(delta)
    w6 = np.diag((e**2)/(md6))
    vm6 = q_i@x_t@w6@x@q_i
    sigma6 = math.sqrt(vm6[1,1])
```

```
    t6 = (b1_hat-b1)/sigma6
    if (t6>critical_value)or(t6<-critical_value):
        c6 = c6+1


    ##检验 HCCv

    a = 0
    mdx1 = []
    while (a<n):

        delta = min((4**(2.6-cv2)+0.5)*(n*hd[a]/hd_sum),1.6*cv2)

        mdx1 = np.r_[mdx1,md[a]**delta]
        a = a+1
        #print(delta)
    wx1 = np.diag((e**2)/(mdx1))
    vmx1 = q_i@x_t@wx1@x@q_i
    sigmax1 = math.sqrt(vmx1[1,1])
    tx1 = (b1_hat-b1)/sigmax1
    if (tx1>critical_value)or(tx1<-critical_value):
        cx1 = cx1+1



    ##参照组检验
    w = np.diag(v)
    vm = q_i@x_t@w@x@q_i
    sigma =    math.sqrt(vm[1,1])
    t = (b1_hat-b1)/sigma
    if (t>critical_value)or(t<-critical_value):
        c = c+1
print('OLS 估计量拒绝次数',c00)
```

```
print('HC0 估计量拒绝次数',c0)
print('HC1 估计量拒绝次数',c1)
print('HC2 估计量拒绝次数',c2)
print('HC3 估计量拒绝次数',c3)
print('HC4 估计量拒绝次数',c4)
print('HC5 估计量拒绝次数',c5)
print('HC4m 估计量拒绝次数',c4m)
print('HC5m 估计量拒绝次数',c5m)
print('HC6 估计量拒绝次数',c6)
print('HCCv 的估计量拒绝次数',cx1)
print(' ')
print('对照拒绝次数          ',c)
```

参考文献

[1] 钱莹，方秀男．多元线性回归模型及实例应用[J]．中国科技信息，2022，669 (04): 73-74.

[2] 冷建飞，高旭，朱嘉平．多元线性回归统计预测模型的应用[J]．统计与决策，2016, 451(07): 82-85.

[3] 任丹．基于多元线性回归模型的电影票房预测系统设计与实现[D]．中山大学, 2015.

[4] 蔡成旺，潘芷莹，蔡孝庆，等．基于线性回归的焦炭质量预测模型研究[J]．当代化工研究, 2023(01): 1-4.

[5] 钱莹，方秀男．多元线性回归模型及实例应用[J]．中国科技信息，2022(04): 73-74.

[6] 徐礼文，廖丹．大样本线性回归模型的子抽样及变量选择[J]．统计与决策，2022, 38 (02): 5-9.

[7] 廖文辉，黄颖强，何志锋，等．线性回归模型中参数估计稳健性比较及应用[J]．数理统计与管理, 2021, 40(05): 822-832.

[8] 龚艳冰，巢妍．基于正态云线性回归模型的企业员工绩效评价[J]．统计与决策, 2020, 36(18): 167-170.

[9] 张童巍．若干广义自回归条件异方差模型的统计推断[D].吉林大学, 2021.

[10] 阎颖，曲建民．计量经济学在经济应用中的异方差问题[J]．工业技术经济，2005(02): 111.

[11] Knaub J R. When Would Heteroscedasticity in Regression Occur?[J]. Pakistan Journal of Statistics, 2021, 37(4): 315-367.

[12] Schlaak T, Rieth M, Podstawski M. Monetary policy, external instruments, and heteroskedasticity[J]. Quantitative Economics, 2023, 14(1): 161-200.

[13] 徐睿．截面数据线性回归模型中异方差问题的研究[D]．延边大学, 2020.

[14] 严威．异方差和多重共线性的影响——模拟蒙特卡洛[J]．北方经贸，2014 (08): 26-27.

[15] 林天水，陈佩树．一元线性回归中异方差的处理[J]．统计与决策, 2015(21): 86-88.

[16] Park R E. Estimation With Heteroscedastic Error Terms[J]. Econometrica,

1966, 34(4)4: 888.

[17] Glejser H. A New Test for Heteroskedasticity[J]. Journal of American Statistical Association, 1969, 64(325): 316-323.

[18] Breusch T S, Pagan A R. A Simple Test Heteroscedasticity and Random Coefficient Variation[J]. Econometrica, 1979, 47(5): 1287-1294.

[19] Goldfield S M, Quandt R E. Some tests for Heteroskedasticity[J]. Publications of the American Statistical Association, 1965, 60(310): 539-547.

[20] White H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity[J]. Econometrica, 1980, 48(4): 817-838.

[21] 龚秀芳. 戈德菲尔德-匡特检验的推广[J]. 数理统计与管理, 2005(01): 98-102.

[22] 金蛟. 基于统计深度函数的 G-Q 检验[J]. 数理统计与管理, 2008(01): 100-105.

[23] 郑红艳, 夏乐天. 一种多变量线性回归模型的异方差检验方法[J]. 统计与决策, 2010(05): 152-154.

[24] 李俊领. 检验异方差的新方法-修正的 G-Q 检验[J]. 金融经济, 2012(24): 78-80.

[25] 刘明, 黄恒君. 异方差 G-Q 检验方法的改进[J]. 统计与信息论坛, 2018, 33(02): 3-9.

[26] 张晓琴, 牛建永, 李顺勇. 基于 G-Q 的 K-S 异方差检验方法[J]. 山西大学学报(自然科学版), 2019, 42(01): 95-104.

[27] 彭作祥, 庞皓. 计量经济模型中扰动项异方差性的一种检验方法[J]. 西南师范大学学报(自然科学版), 2003(01): 58-60.

[28] 兰嘉庆, 余宛泠. 异方差的游程检验[J]. 中山大学学报: 自然科学版, 2004 (S1): 9-11.

[29] 张荷观. 基于分组的异方差检验和两阶段估计[J]. 数量经济技术经济研究, 2006(01): 129-137.

[30] 夏帆, 倪青山. 基于分布拟合的异方差检验[J]. 数量经济技术经济研究, 2012, 29(08): 114-123.

[31] 唐裔, 冯长焕. 多元线性回归模型异方差检验研究[J]. 廊坊师范学院学报：自然科学版, 2018, 18(01): 8-11.

[32] 谭馨, 邓光明. 异方差帕克检验方法的改进[J]. 统计与信息论坛, 2019, 34

(06): 10-16.

[33]刘锋，胡悦，康新梅. 响应变量缺失下部分线性模型的异方差检验[J]. 重庆理工大学学报：自然科学, 2019, 33(02): 180-187.

[34]Hinkley D V. Jackknifing in Unbalanced Situations[J]. Technometrics, 1977, 19(3): 285-292.

[35]Susan D, Horn, et al. Estimating Heteroscedastic Variances in Linear Models[J]. Journal of the American Statistical Association, 1975, 70(350): 380-385.

[36]Mackinnon J G, White H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties[J]. Journal of Econometrics, 1985, 29(3): 305-325.

[37]Davidson R, Mackinnon J G. Estimation and Inference in Econometrics[M]. New York (NY): Oxford University Press; 1993.

[38]Cribari-Neto F. Asymptotic inference under heteroskedasticity of unknown form[J]. Computational Statistics & Data Analysis, 2004, 45(2): 215-233.

[39]Cribari-Neto F, Souza T C, Vasconcellos K L P. Inference Under Heteroskedasticity and Leveraged Data[J]. Communications in Statistics, 2007, 36(10): 1877-1888.

[40]Cribari-Neto F, Silva W B D. A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model[J]. AStA Advances in Statistical Analysis, 2011, 95(2): 129-146.

[41]Li S, Zhang N, Zhang X et al. A new heteroskedasticity-consistent covariance matrix estimator and inference under heteroskedasticity[J]. Journal of Statistical Computation and Simulation, 2017, 87(1): 198-210.

[42]Aftab N, Chand S. A simulation-based evidence on the improved performance of a new modified leverage adjusted heteroskedastic consistent covariance matrix estimator in the linear regression model[J]. Kuwait Journal of Science, 2018, 45(3): 29-38.

[43]Zhang X, Hao H, Liang J. A new nonparametric estimation method of the variance in a heteroscedastic model[J]. Hacettepe Journal of Mathematics and Statistics, 2015, 44(1): 239-245.

[44]张晓琴，王佳鸣. 基于正交表的异方差估计方法改进[J]. 数理统计与管理, 2016, 35(02): 225-231.

[45]冯军芳，张晓琴. 一种基于正交表的异方差估计方法[J]. 统计与决策, 2021,

37(08): 63-67.

[46]郭雅静. 基于非参数方法的异方差估计研究[D]. 山西大学, 2019.

[47]郎静波. 广义最小二乘法在建立回归模型中的应用[J]. 浙江统计, 2001(05): 19-20.

[48]Loh J. Asymptotic theory for Box–Cox transformations in linear models[J]. Statistics & Probability Letters, 2001, 54(4): 337-343.

[49]Yau P, Kohn R. Estimation and variable selection in nonparametric heteroscedastic regression[J]. Statistics & Computing, 2003, 13(3): 191-208.

[50]Leslie D S, Kohn R, Nott D J. A general approach to heteroscedastic linear regression[J]. Statistics & Computing, 2007, 17(2): 131-146.

[51]张晓琴, 郭雅静, 李顺勇. 一种基于 N-W 估计的异方差估计方法[J]. 统计学报, 2020, 1(02): 31-38.

[52]李顺勇, 李佳欣. 基于自适应 N-W 估计的异方差估计方法[J/OL]. 山西大学学报(自然科学版): 1-12[2023-02-22].

[53]张晓琴, 郭雅静, 米子川. 基于局部多项式方法的异方差估计[J]. 数理统计与管理, 2021, 40(06): 1019-1030.