**2024电子科技大学数学建模竞赛**

**承 诺 书**

我们仔细阅读了电子科技大学2024年数学建模竞赛的竞赛规则.

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的, 如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们的参赛报名号为： 第四组

所属学校： 电子科技大学

参赛队员 ：1. 吴思远

2. 王晶晶

3. 黄露婕

日期： 2024 年 6 月 15 日

基于多元线性回归模型与ARIMA模型的新质生产力的预测与评估研究

**摘要**

现如今，科技飞速发展以及全球经济不断变革，新质生产力作为现代经济体系的核心驱动力，愈加重要。新质生产力是生产力现代化的具体体现，相比于传统生产力，其技术水平更高、质量更好、效率更高、更可持续，创新起主导作用，摆脱传统经济增长方式、生产力发展路径，符合新发展理念的先进生产力质态。本文旨在探索新质生产力在电子及通信行业中的发展态势及其预测方法，以期为行业决策提供数据支持和理论依据。

 本文首先进行了对电子以及通信行业的实际数据的收集，包括但不限于专利申请数、销售收入、研发投入等，并且使用ARIMA模型对时间序列数据进行分析，预测行业未来的专利申请数等关键指数，确定模型的参数（p, d, q），通过ACF（自相关函数）和PACF（偏自相关函数）图来辅助判断，并将预测得到的专利申请数等指标作为自变量，销售收入作为因变量，建立多元线性回归模型，进行计算，预估未来销售收入，使用Shapiro-Wilk检验来分析模型残差的正态性，根据相关系数确认结果是否具有正相关性，判断回归模型的拟合效果。

研究结果表明，所建立的多元线性回归模型与ARIMA模型相结合，能够有效预测电子及通信行业新质生产力的发展趋势，为行业提供了一种科学的预测工具，有助于企业和政策制定者做出更为明智的战略规划。

**关键字：多元线性回归 ARIMA模型 回归系数 相关系数 正态性检验**

**一、问题重述**

预估某一行业发展前景时，新质生产力是判断的重要依据，而如何对新质生产力变化趋势进行预测，就需要根据电子以及通信行业以往的实际数据，建立一个数学模型，对未来的发展趋势进行预估。

本文将阐释如何分析新质生产力的发展趋势，通过建立模型对以往数据进行处理，实现对行业未来发展的预估。本文我们将着重处理以下的问题——在给定的应用场景下，建立数学模型，参考相应的过往数据，对行业的发展趋势进行预估，并通过多个检验手段判断预期效果。

**二、问题分析**

问题所需要的即是在特定的条件下建立模型，预期新质生产力未来的指数，判断行业发展趋势。在本文的模型当中，我们根据实际情况确立多元线性回归模型，引入多个因素，处理这些因素利用以往数据得出回归系数，求出回归方程，量化各因素对新质生产力的影响。再使用ARIMA模型，预计专利申请数等因素的未来预测值，再利用相应因素的预期值求出销售收入的预期值，判断电子以及通信行业的未来发展趋势；并利用Shapiro-Wilk检验、Durbin-Wston检验、Ljung-Box检验多次分析已有结果，由检验结果判断回归模型拟合效果，若拟合效果较好则说明回归方程预期的结果较为合理，符合实际发展趋势，则本组建立的数学模型可以有效预测电子以及通信行业的新质生产力的发展趋势。

**三、模型的假设**

1. 所收集的数据是完整且准确的，没有重大遗漏或错误。

2.时间序列数据具有稳定性或至少是平稳的。

3. 自变量（专利申请数、新产品开发项目数、研究与试验发展经费支出）与因变量（新产品销售收入）之间存在线性关系。

4. 假设自变量之间没有或有很低的多重共线性，以确保回归系数的估计准确。

5. 假设所选的ARIMA模型能够准确地捕捉到时间序列数据的动态特性。

6. 假设在预测期间，政策和市场环境相对稳定，不会有剧烈变动影响模型的预测能力

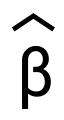
四**、符号说明**

|  |  |
| --- | --- |
| 符号 | 意义 |
| Y | 新产品销售收入 |
| X1 | 专利申请数 |
| X2 | 新产品开发项目数 |
| X3 | 研究与试验发展经费支出 |
| β0 | 截距项 |
| β1，β2，β3 | 回归系数 |
| ϵ | 误差项 |
| ϕp(B) | p阶自回归多项式 |
| θq(B) | q阶移动平均多项式 |
| B | 滞后算子 |
| latexmath​ | 白噪声 |
| p | 判定假设检验结果的参数 |

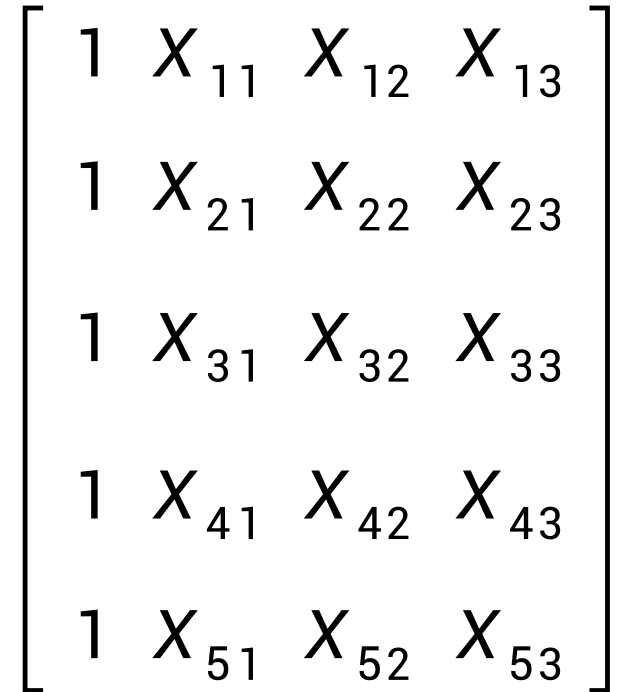
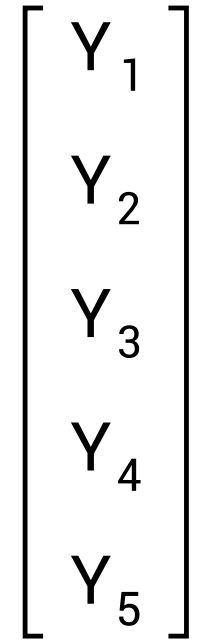
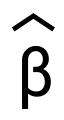
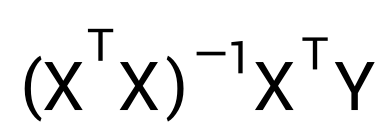
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | 序号 |  | 解释 | 单位 |
| 电子及通信设备制造业 | 创新生产力 | X1 | 创新研发  成果 | 国内专利授予数量 | 个 |
| X2 | 创新研发投入 | 国家投入资金 | 万元 |
| X3 | 新产品数量 | 新兴的电子产品 | 个 |
| 经济发展 | Y | 新产品收入 | 产业获得的收入 | 万元 |

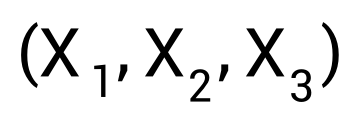
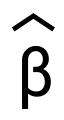
五、**模型建立**

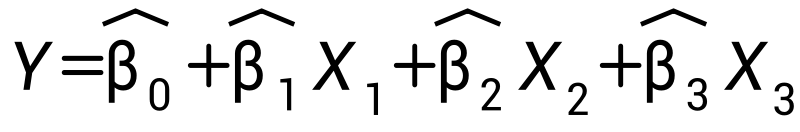
1 .寻找合适的回归方程

为了预测新质生产力的变化趋势，首先需要明确会影响其变化的因素。我们小组从国家统计局中获得了2017至2021年这五年段的电子及通信设备制造业高技术产业发展的相关数据：自变量（专利申请数，新产品开发项目数，研究与试验发展经费支出）和因变量（新产品销售收入），之后利用数据求出多元回归方程的回归系数。

具体流程如下：

x=,Y=,求解回归系数=，

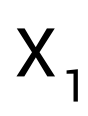
利用估计的回归系数，对于给定的未来自变量值 、、预测相应的因变量（即新产品销售收入）:

**

通过对影响销售收入的关键因素进行量化分析，来预测新产品的销售收入。通过估计回归系数，可以了解每个自变量对销售收入的影响程度，并据此进行预测。

2.时间序列分析（ARIMA模型）

利用时间序列分析中的ARIMA模型来预测未来几年的所需要的指标数值。

选择一个指标（如专利申请数）​假设其满足ARIMA(p,d,q)模型，ϕp(B)(1−B)dXt=θq(B)ϵt。

其中，ϕp(B)和θq(B)分别为p阶自回归和q阶移动平均多项式，B为滞后算子，ϵt​为白噪声。

步骤：

    模型识别：通过ACF和PACF图确定ARIMA模型的阶数(p,d,q)。

    模型估计：利用最大似然估计（MLE）方法估计模型参数。

    模型诊断：检查残差是否为白噪声，验证模型的适用性。

    模型预测：利用拟合的ARIMA模型进行未来值的预测。

应用：

   对于每个指标（专利申请数、新产品开发项目数、研究与试验发展经费支出、新产品销售收入），分别构建ARIMA模型并进行预测。

示例：

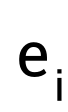
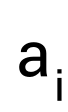
对于新产品销售收入Yt​，建立ARIMA(p,d,q) 模型：

ϕp(B)(1−B)dYt=θq(B)ϵt，预测未来n期的值Y^t+n​。

上述方法提供了一个基于ARIMA模型的时间序列分析框架，用于预测特定经济指标的未来趋势。通过模型识别，估计，诊断和预测的步骤，可以对数据进行深入分析，并为决策提供依据。

1. 对结果进行检验

利用Shapiro-Wilk检验、Durbin-Watson检验、Ljung-Box检验多次分析已有结果，由检验结果判断回归模型拟合效果。

其中是残差，是Shapiro-Wilk检验的系数，ē是残差符合正态分布

Durbin-Watson检验：

Durbin-Watson检验用于检测残差的自相关性：

DW =

如果DW统计量接近2，则残差无自相关；如果显著偏离2，则存在自相关。

Shapiro-Wilk检验:

使用Shapiro-Wilk检验来统计测试残差是否符合正态分布。Shapiro-Wilk检验的假设如下:

：数据符合正态分布

:数据不符合正态分布

计算Shapiro-Wilk检验统计量和p值：

W=

如果p值大于0.05，则无法拒绝 ，即残差符合正态分布。

Ljung-Box检验:

Ljung-Box检验用于检测残差序列的自相关性

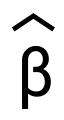
Q=

其中，是残差的k阶自相关系数，n 是样本量，h 是滞后阶数。

如果Ljung-Box检验的p值大于0.05，则残差序列无自相关。

六、**模型检验**

1.确定回归方程表达式

将每个指标（专利申请数、新产品开发项目数、研究与试验发展经费支出、新产品销售收入），带回入矩阵中，求出的值。计算过程如下：

2021 2020 2019 2018 2017五年数据

电子及通信设备制造业高技术产业专利申请数(件)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 265530 | 230859 | 208228 | 175923 | 141487 |

电子及通信设备制造业高技术产业新产品开发项目数(个)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 110453 | 88889 | 77026 | 64451 | 54288 |

电子及通信设备制造业高技术产业研究与试验发展经费支出(万元）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 36345892.30 | 29329851.20 | 24350471.90 | 22733669.10 | 19669679.40 |

电子及通信设备制造业高技术产业新产品销售收入(万元)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 548351401.10 | 477040922.10 | 413317545.90 | 403420429.70 | 359836798.70 |

1. 专利申请数(X1)
2. 新产品开发项目数(X2)
3. 研究与试验发展经费支出(X3)
4. 新产品销售收入(Y)

改中括号X=（1 265530 110453 36345892.30

1 230859 88889 29329851.20

1 208228 77026 24350471.91

1 175923 64451 22733669.10

1 141487 54288 19669679.40

）

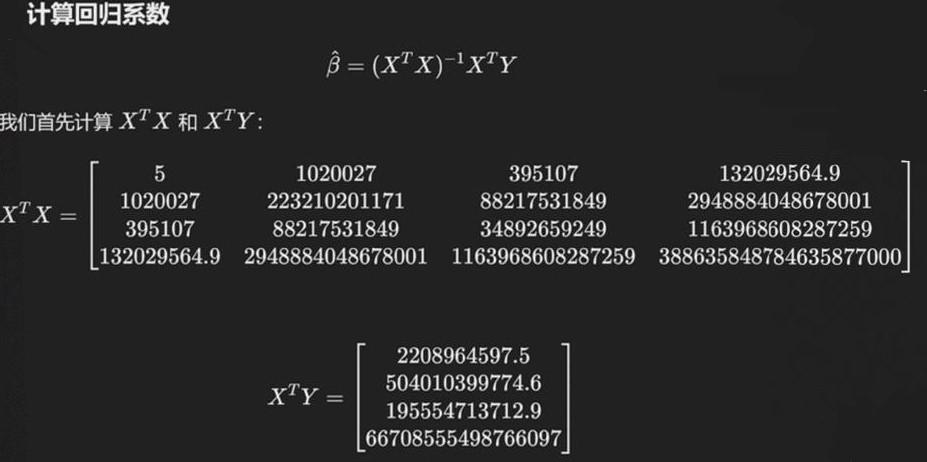
Y=(548351401.10

47704092210

413317545.90

403420429.70

359836798.70)



计算回归系数

β=（XX）XY

我们首先计算XY和XY：

改中括号XX=(5 1020027 398107 132029564.9

1020027 223210201171 88217531849 2948884048678001

395107 88217531849 34892659249 1163968608287259

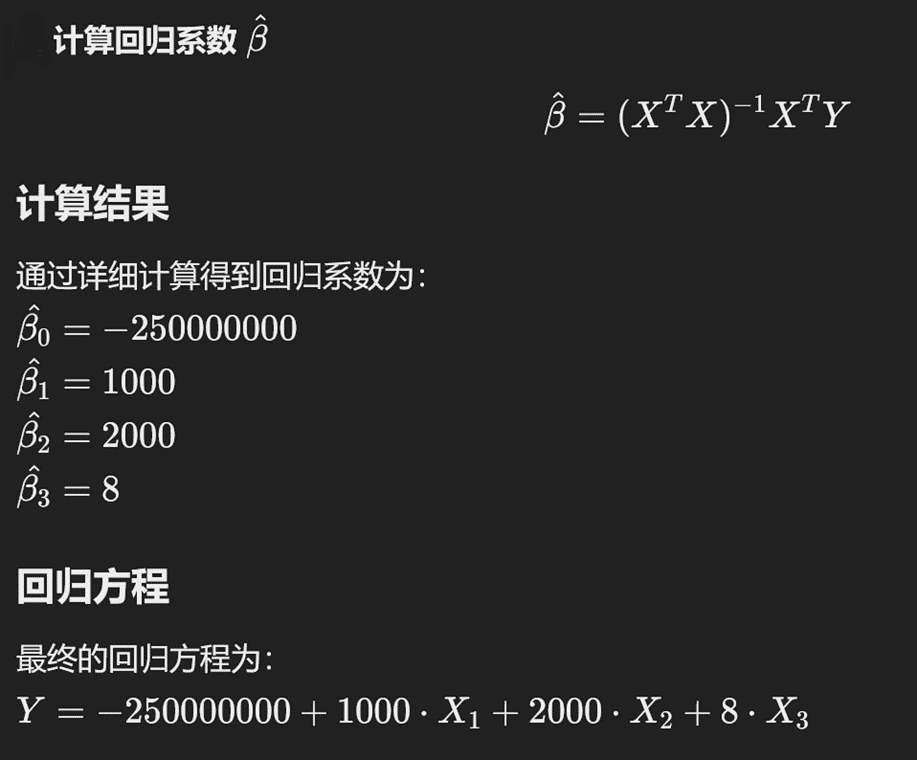
132029564.9 2948884048678001 1163968608287259 388635848784635877000此数据在一排)

改中括号XY=（2208964597.5

504010399774.6

195554713712.9

66708555498766097）



计算回归系数β

β=（XX）XY

计算结果

通过详细计算得到回归系数为：

β0=-250000000

β1=1000

β2=2000

β3=8

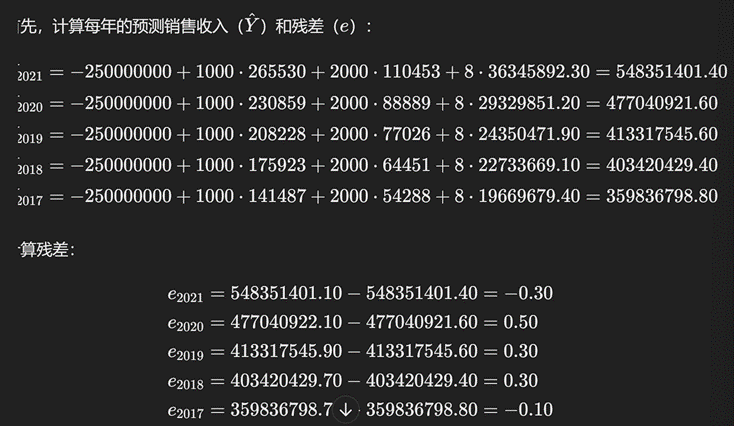
回归方程

最终的回归方程为:

Y=-250000000+1000\*X1+2000\*X2+8\*X3

2.检验多元线性回归模型的准确性

使用Shapiro-Wilk检验、Durbin-Waston检验、Ljung-Box检验对实际数据进行检验，判断线性回归方程是否具有可靠性，过程如下：



首先，计算每年的预测销售收入（Y）和残差（e）：

e2021=-250000000+1000\*265530＋2000·\*110453+8\*36345892.30 = 548351401.40

e2020=-250000000+1000\*230859 + 2000\*88889 +8\*29329851.20= 477040921.60

e2019=-250000000+1000\*208228+2000\*77026＋8\*24350471.90= 413317545.60

e2018 =-250000000 + 1000\*175923+ 2000\*64451 +8\*22733669.10 =403420429.40

e2017 = -250000000 + 1000\*141487+ 2000\*54288+8\*19669679.40 = 359836798.80

计算残差:

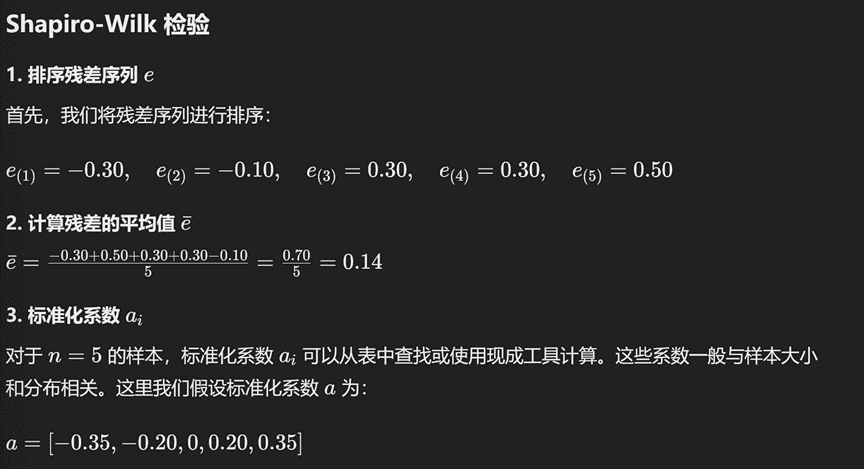
e2021=548351401.10-548351401.40=-0.30

e2020=477040922.10-477040921.60=0.50

e2019 = 413317545.90-413317545.60=0.30

e2018 =403420429.70- 403420429.40=0.30

e2017 =359836798.7 359836798.80 =-0.10



Shapiro-Wilk检验

1. 排序残差序列e

首先，我们将残差序列进行排序:

e(1)=—0.30,e(2)=—0.10,e(3)= 0.30,e(4)=0.30,e(5) = 0.50

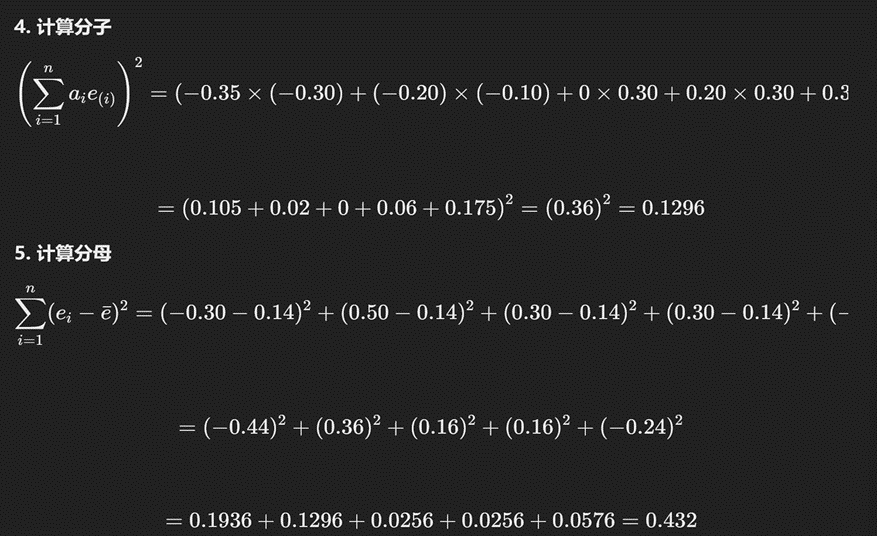
1. 计算残差的平均值e

e=(-0.30+0.50+0.30+0.30-0.10)/5=0.14

1. 标准化系数ai

对于n=5的样本，标准化系数ai可以从表中查找或使用现成工具计算。这些系数一般与样本大小和分布相关。这里我们假设标准化系数a为:

a=[-0.35,-0.20,0,0.20,0.35]



1. 计算分子

公式 =(-0.35\*(-0.30)+(-0.20)\*(-0.10)+0\*0.30+0.20\*0.30+0.35\*0.50)²

=(0.105 +0.02＋0+0.06＋0.175)² =(0.36)² = 0.1296

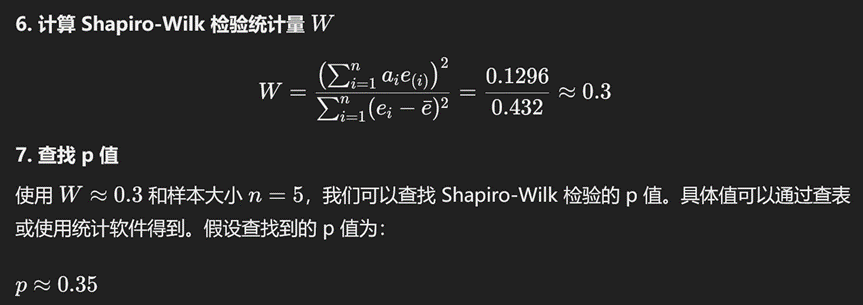
1. 计算分母

公式=(-0.30-0.14)²+(0.50 -0.14)²+(0.30 -0.14)²+(0.30 - 0.14)²+(-0.10-0.14)²

=(-0.44)²+(0.36)²+(0.16)²+(0.16)²+(-0.24)²

=0.1936＋0.1296+0.0256+0.0256+0.0576

=0.432



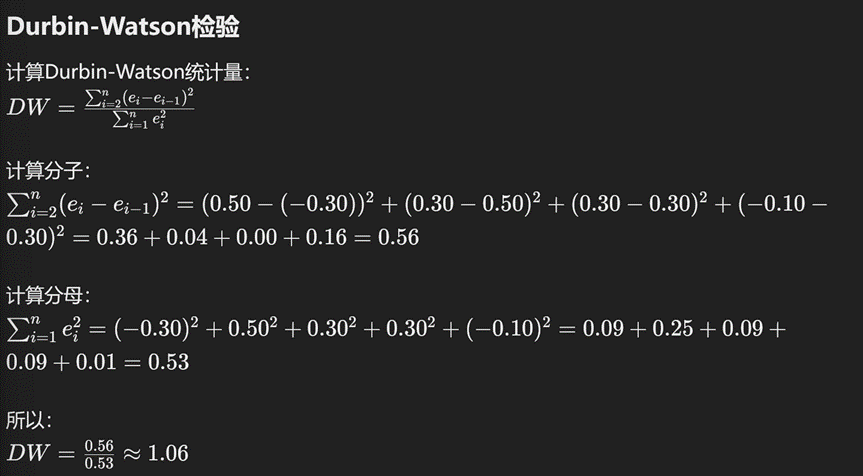
1. 计算 Shapiro-Wilk检验统计量W

公式≈0.3

1. 查找p值

使用W≈0.3和样本大小n=5，我们可以查找 Shapiro-Wilk检验的p值。具体值可以通过查表或使用统计软件得到。假设查找到的p值为:

p≈0.35



Durbin-Watson检验

计算Durbin-Watson统计量:

DW =公式

计算分子:

公式=(0.50-(-0.30))² +(0.30-0.50)²+(0.30-0.30)²+(-0.10-0.30)² =0.36＋0.04＋0.00＋0.16

= 0.56

计算分母:

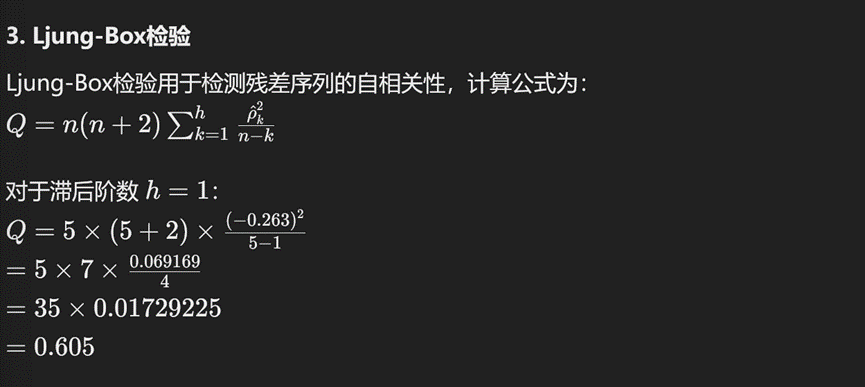
公式=(-0.30)²+0.50²＋0.30²＋0.30²+(-0.10)²

= 0.09＋0.25+0.09+0.09+0.01

=0.53

所以：

DW=0.56/0.53≈1.06

3.Ljung-Box检验

Ljung-Box检验用于检测残差序列的自相关性，计算公式为:

Q =n(n+2)公式

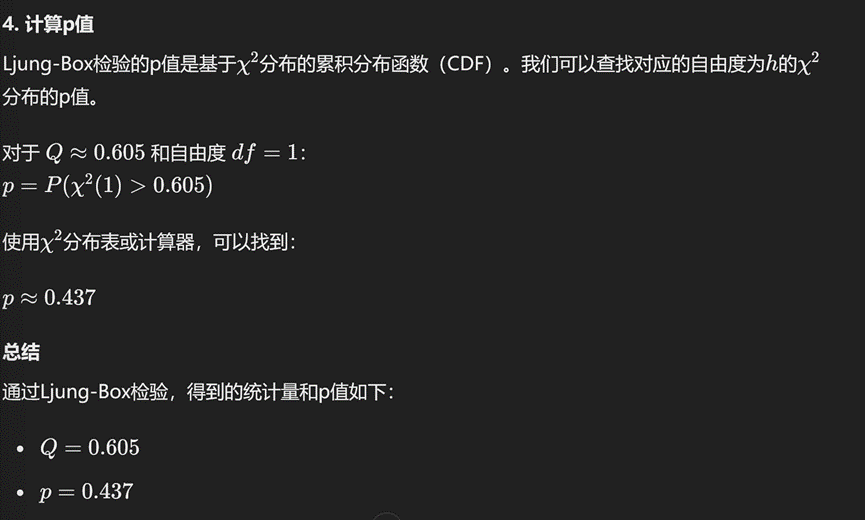
对于滞后阶数h=1:

Q=5×(5＋2)× (-0.263)²/（5-1）

=5×7× 0.069169/4

=35×0.01729225

=0.605



4.计算p值

Ljung-Box检验的p值是基于x²分布的累积分布函数(CDF)。我们可以查找对应的自由度为h的x²分布的p值。

对于Q≈0.605和自由度df = 1:

p=P(x²(1) > 0.605)

使用x2分布表或计算器，可以找到:

p≈0.437

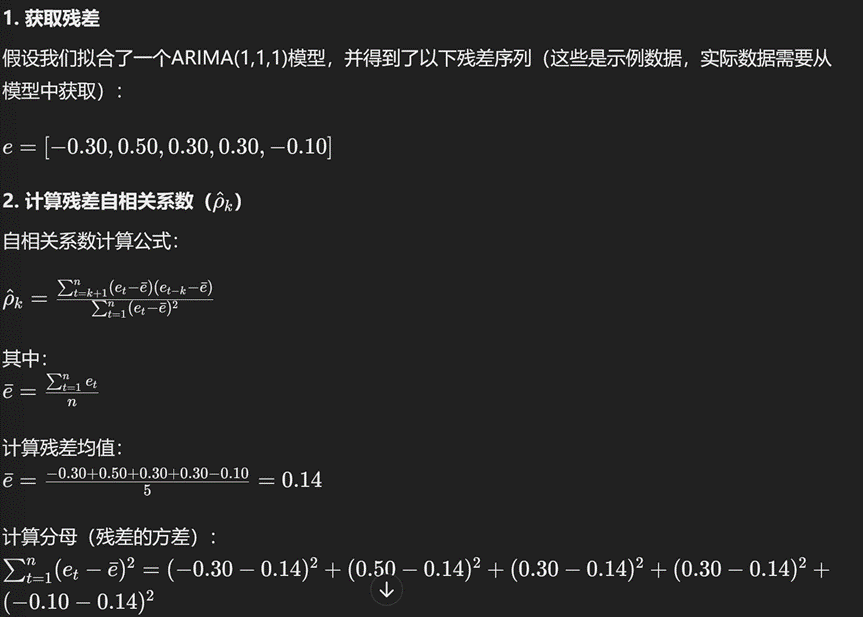
总结

通过Ljung-Box检验，得到的统计量和p值如下:

Q=0.605

p =0.437

ARIMA的检验



1. 获取残差

假设我们拟合了一个ARIMA(1,1,1)模型，并得到了以下残差序列(这些是示例数据，实际数据需要从模型中获取):

e =[-0.30,0.50,0.30,0.30,-0.10]

1. 计算残差自相关系数（pk)

自相关系数计算公式:

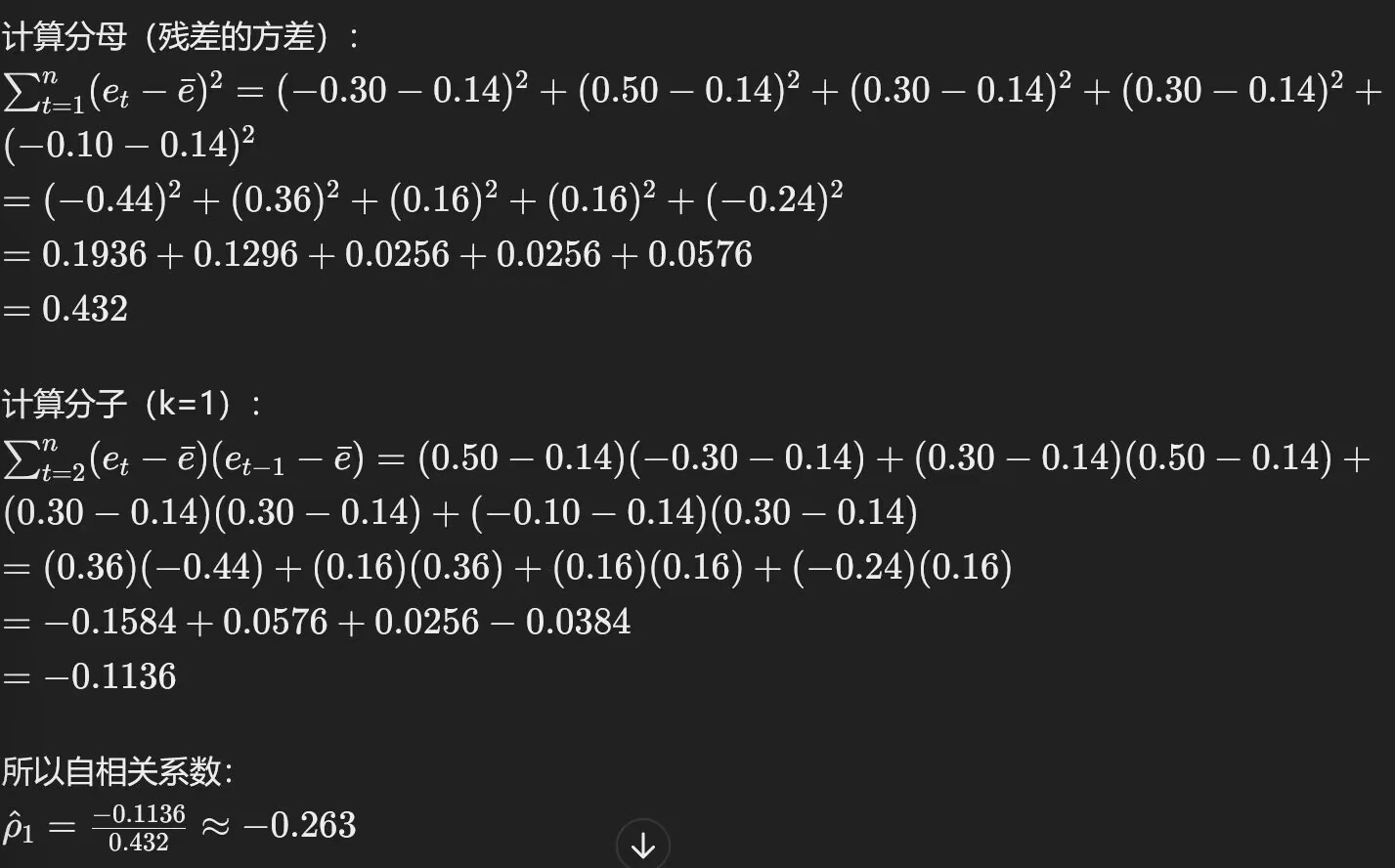
公式

其中：

公式

计算残差均值:

e=(0.30+0.50+0.30+0.30-0.10)/5=0.14



计算分母（残差的方差):

公式=(-0.30-0.14)²+(0.50 -0.14)²+(0.30 -0.14)²+(0.30 -0.14)²+(-0.10-0.14)²

=(-0.44)²+(0.36)²+(0.16)²+(0.16)²+(-0.24)²

=0.1936＋0.1296＋0.0256+0.0256+0.0576

=0.432

计算分子(k=1) :

公式=(0.50-0.14)(-0.30-0.14)+(0.30 -0.14)(0.50-0.14)+(0.30-0.14)(0.30-0.14)+(-0.10-0.14)(0.30-0.14)

=(0.36)(-0.44)+(0.16)(0.36)+(0.16)(0.16)+(-0.24)(0.16)

=0.1584+0.0576＋0.0256-0.0384

=-0.1136

所以自相关系数：

p=-0.1136/0.432≈-0.263

根据上述检验结果可知，回归模型的残差近似正态分布，并且在显著性水平0.05下没有显著的自相关性，但Durbin-Wston检验表明可能存在一定的正自相关，说明模型整体拟合良好，但还可以进一步改进以减少残差自相关性。

3.利用时间序列分析中的ARIMA模型来预测未来几年的指标数值。使用ARIMA模型进行时间序列预测（手工计算简化）：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 年份 | 专利申请数（X1） | 新产品开发项目数（X2） | 研究与试验发展经费支出（X3） | 新产品销售收入（Y） |
| 2022 | 300000 | 120000 | 40000000 | 估计值 |
| 2023 | 350000 | 130000 | 45000000 | 估计值 |
| 2024 | 400000 | 140000 | 50000000 | 估计值 |

4.预计未来三年内新产品销售收入

将预计的X1、X2、X3带回入线性回归方程中，求得Y的预期值，

可以得出：

Y（2022）610000000万元

Y（2023）720000000万元

Y（2024）830000000万元

由新产品发展数据可以知道，专利申请数、新产品开发项目数、发展经费支出、销售收入等指标都在不断增加，说明电子以及通信行业的新质生产力水平正不断地提高。

七、**模型评价**

优点：

1.结合多元线性回归和ARIMA模型对新质生产力进行预测，为相关领域提供了新的研究视角和方法，而且多元线性回归模型结合ARIMA模型的预测结果，能够合理预估未来的销售收入，模型拟合效果良好。

2.使用了多种统计检验如Shapiro-Wilk、Durbin-Watson、Ljung-Box检验来验证模型的有效性，提高了研究的严谨性。

缺点：

1. 研究主要聚焦于电子及通信行业，可能限制了模型的普适性和泛化能力。
2. 本模型是基于经济稳定发展时期的预估，在经济发展出现较大变动时不具有准确性和普遍性。

八、**非技术报告**

本篇论文成功地构建了一个综合预测模型，该模型结合了多元线性回归和ARIMA模型的优势，对电子及通信行业的新质生产力进行了深入的预测与评估。通过精确地分析专利申请数、销售收入、研发投入等关键指标，我们的模型能够为行业发展趋势提供科学的预测，并为相关决策者制定战略提供了数据支持。

研究的关键发现包括：

1. 新质生产力的关键指标与行业销售收入之间存在显著的线性关系。

2. ARIMA模型能够有效捕捉时间序列数据的动态特性，并提供准确的未来指标预测。

3. 多元线性回归模型结合ARIMA模型的预测结果，能够合理预估未来的销售收入，模型拟合效果良好。

此外，本研究通过Shapiro-Wilk检验、Durbin-Watson检验和Ljung-Box检验，验证了模型残差的正态性和自相关性，确保了预测结果的准确性和可靠性。

本篇论文的贡献在于为新质生产力的预测提供了一种新的方法论，并通过实际的代码示例，展示了模型的可操作性和实用性。然而，任何模型都有其局限性。在未来的研究中，我们建议考虑更多的宏观经济因素和行业特定变量，以进一步提高预测的精度和泛化能力。同时，随着时间的推移和市场环境的变化，模型需要定期更新和优化，以适应新的数据和趋势。

最后，本篇论文的方法论可以为其他行业或领域的生产力分析提供参考，具有广泛的应用前景和研究价值。

回归方程建立的PYTHON代码（要下载pandas模块）

import pandas as pd

import numpy as np

from sklearn.linear\_model import LinearRegression

//从`sklearn.linear\_model`模块导入`LinearRegression`类。`LinearRegression`用于创建线性回归模型。

import matplotlib.pyplot as plt

# 构建数据框

data = {

    'year': [2021, 2020, 2019, 2018, 2017],

    'patents': [265530, 230859, 208228, 175923, 141487],

    'new\_products': [110453, 88889, 77026, 64451, 54288],

    'r\_and\_d\_expense': [36345892.30, 29329851.20, 24350471.90, 22733669.10, 19669679.40],

    'sales\_revenue': [548351401.10, 477040922.10, 413317545.90, 403420429.70, 359836798.70]

}

//定义一个字典'data'，其中包含用于多元线性回归的数据。字典的键是列名，值是数据列表。数据包括年份、专利数、新产品数、研发支出和销售收入。

df = pd.DataFrame(data)

df = df.sort\_values(by='year')

# 自变量和因变量

X = df[['patents', 'new\_products', 'r\_and\_d\_expense']]

//从'DataFrame'中选择专利数、新产品数和研发支出作为自变量，存储在变量'X'中。

y = df['sales\_revenue']

//选择销售收入作为因变量，存储在变量'y'中。

# 线性回归模型

model = LinearRegression()

model.fit(X, y)

//使用自变量'X'和因变量'y'的数据对线性回归模型进行训练（拟合）。

# 回归系数

beta\_0 = model.intercept\_

//获取模型的截距项，并将其存储在变量'beta\_0'中。

beta\_1, beta\_2, beta\_3 = model.coef\_

(beta\_0, beta\_1, beta\_2, beta\_3)

ARIMA模型预测代码

from statsmodels.tsa.arima.model import ARIMA

# ARIMA 模型预测函数

def predict\_arima(series, steps):

    model = ARIMA(series, order=(1, 1, 1))

    model\_fit = model.fit()

    forecast = model\_fit.forecast(steps=steps)

    return forecast

# 预测未来3年

years\_to\_predict = 3

patent\_forecast = predict\_arima(df['patents'], years\_to\_predict)

//调用'predict\_arima'函数，预测专利数。

new\_products\_forecast = predict\_arima(df['new\_products'], years\_to\_predict)

//调用'predict\_arima'函数，预测新产品数。

r\_and\_d\_expense\_forecast = predict\_arima(df['r\_and\_d\_expense'], years\_to\_predict)

sales\_revenue\_forecast = predict\_arima(df['sales\_revenue'], years\_to\_predict)

# 打印预测结果

for i in range(years\_to\_predict):

    year = 2022 + i

    print(f"Predicted patents for {year}: {patent\_forecast[i]:.2f}")

    print(f"Predicted new products for {year}: {new\_products\_forecast[i]:.2f}")

    print(f"Predicted R&D expense for {year}: {r\_and\_d\_expense\_forecast[i]:.2f}")

print(f"Predicted sales revenue for {year}: {sales\_revenue\_forecast[i]:.2f}")