中国林业产业与其影响因素分析

张庶杰 2018011415 计 85

刘程华 2018011687 计 91

周加 2018030041 生 85

摘要：林业是国民经济重要组成部分之一，其发展受到多重因素影响，具有随机性和动态性，与我国而言，地域异质性导致林业管理难度较大，合理规划林业生产需要进行严谨的分析。针对 1998-2017 年间全国 31 省林业生产总值及其解释变量数据，本文将研究和探讨我国林业经济发展中林总产值的产业结构优化及因素分析。首先，通过 K-Means 分类将 31 个省直辖市分为 5 类，针对各类具体情况将 PCA、lasso、box-cox 等方法应用到多元线性回归中，并对模型进行诊断和假设检验，针对各省份给出了具体的参考建议。其次，运用空间面板数据模型，我们首给出了林业产值的另几种模型，并综合残差分析、数据挖掘等方法给出了 31 省林业产值的固定效应空间自回归模型，最后从社会学角度分析了与林业产值最有关的几个社会学因素，并给出了简单的政策建议。

关键词：多元线性回归；主成分分析；K-Means 分类；空间面板数据模型

一．引言

1. 中国林业情况

林业是国民经济重要组成部分之一，指在维护生态平衡的前提下取得林产品和开发其他生产部门。在林业管理中，主要工作不仅为森林资源的利用，还需综合当地经济、人口、资源分配因素，实现森林资源的持续经营。于我国而言，地

域广阔导致的地理异质性使林业管理难度增加。重点林区森林资源遭受长期超强度采伐，2001 年六大林业重点工程确立，促进了林业合理规划和林业总产值的逐年增长。2006 年由于国际局势对外贸易受限，2008 年遭受地震和冰冻灾害影响和金融危机对外出口受阻，林业产值受挫。2014 年起，林业部门全面深化改革后，我国林业总产值持续起色。2017 年，我国实现了天然林商业性采伐全面禁止。在林业管理方面，需要持续的关注并提出新的优化方法。

林业的发展受到多重因素影响，具有随机性和动态性，这里我们引入当地房地产、规模以上企业数量、林业从业人数、林业国家投资额、林业企业数量、人均 GDP、人口、森林覆盖率和市林业站数为可能的解释变量进行分析。

2. 面板数据模型

面板数据融合了时间序列数据和截面数据，因此也被称为时间序列截面数据。面板数据可实现个体的横向和纵向对比，横剖面来看面板数据是多个截面数据按时间排布，纵剖面来看面板数据就是多个时间序列数据。相比经典的截面数据模型，面板数据可以降低由不可观测变量带来的最小二乘估计偏差，减少对参数估计的扰动。相比时间序列模型，面板数据含有更多的数据信息，能降低变量间的共线性，提高估计量的有效性。

本文研究对象是 1998-2017 年中国 31 省的林业发展状况，不同省之间的数据存在差异，在地域上也存在关联性，比如，我们推测空间上临近的省份的林业状况相关性更强。首先，由于差异性的存在，认为直接将全体数据归入一个线性回归是不合适的，因此在上个方法中，我们通过聚类将 31 省分成五类，依次做线性回归。在此处，我们将用面板数据模型来分析数据。

2.1. 静态与动态面板数据模型

静态模型和动态模型的区别在于后者的解释变量中含有响应变量的滞后项。从实际应用角度理解，因为很多经济关系是动态的，如时间上存在着记忆性，因此引入动态模型是必要的。简便起见，我们通常事先假定不同个体的自回归系数相等，且不随时间改变，除非系统遭到重大影响，该模型一般是成立的。

首先，我们给出静态面板数据模型：

$$y_{it}=\alpha_i+\lambda_t+x_{it}\beta+\varepsilon_{it}$$

其中，y_{it} 为 N×1 的因变量，x_{it} 为 N×k 的自变量，ε_{it} 为误差项，β 为待估参数。α_i 为个体效应，不随时间变化；λ_t 为时间效应，体现的是随时间改变的部分。

若加入响应变量的滞后项，则转化为动态模型：

$$y_{it}=\alpha_i+\lambda_t+x_{it}\beta+y_{i,t-l}\gamma+\varepsilon_{it}$$

其中，$y_{i,t-l}$是因变量在时刻 t 的滞后值，γ是自回归系数。

本文构建的模型中，空间是影响因变量的重要因素，比如地理位置可决定湿度、温度等，进而从生态层面影响林业发展，同时，相邻地域的经济发展情况和土地开发程度较为相近，从经济建设上间接影响了林业发展。因此，除了个体效应外，各省间林业发展的互相影响也应当被纳入考虑范围，故我们将一个空间自回归项引入模型。至于具体将自回归项放在模型的哪个位置，将在后文中阐述。

2.2. 混合模型、变截距模型和变系数模型

以上，我们从静动态角度区分了两种面板数据模型。而从截距项和解释变量系数的性质来看，面板数据模型又可分为混合模型、变截距模型和变系数模型，其中，变截距模型又可分为固定效应模型和随机效应模型。

混合模型假定，对同一截面中的所有个体而言，截距项和解释变量的系数都是一样的，即同一截面上不存在个体影响。同时，对于任一截面，这些系数也是相同的。

而变截距模型，顾名思义就是对于不同的截面个体，截距项不同。但截距项不同不等价于模型截距项不同。若截距项与解释变量有关，则被称为固定效应模型，模型截距项不一致。若截距项与解释变量无关，则被称为随机效应模型。在随机效应中，直观来看截距项是一致的，个体间差异被归入随机干扰项之中，因此被称为随机效应模型。变系数模型则认为对于截面不同个体，截距项和解释变量系数都不相同。

2.3. 固定效应模型、随机效应模型和 Hausman 检验

如 1.2.所说，变截距模型比混合模型更具实际意义，而选择其中的固定效应模型还是随机效应模型，可以根据研究背景直观判断，或者根据 Hausman 检验结果判断。

比如若研究对象是具体的若干个体，研究目的也仅局限于对这些个体的分析，则应使用固定效应模型。若研究对象是从总体中随机抽取的某些个体，而研究目的针对整个总体，则随机效应模型比较合适。在本文中，我们认为使用固定效应模型比较合理。

Hausman 检验的原理在于，在固定的模型下，固定效应估计和随机效应估计的结果存在差异，通过检测该差异是否存在，可判断使用哪一种模型比较合适。不论研究背景是固定还是随机的，基于固定效应模型的估计都能收敛，因此固定效应模型是普适的模型。在随机抽取的背景下，随机效应模型的效率更好，但在固定的背景下，随机效应模型的估计有偏差。Hausman 检验基于对该偏差的检测，判断是否能够使用随机效应模型。

3. 空间面板数据模型

如 1.1.所说，除了个体效应外，各省间林业发展的空间相关性也应当被纳入考虑范围，故我们将一个空间自回归项引入模型，最终构建一个空间面板数据模型。

3.1. 空间滞后模型与空间误差模型

根据空间自回归项的位置差别，可将空间面板数据模型分为空间滞后模型和空间误差模型。空间滞后模型主要刻画空间相依性，而空间误差模型主要刻画空间异质性。

空间滞后模型中的空间自回归项为空间滞后被解释变量，参考动态面板数据模型，空间滞后模型可表示为，

$$Y = \lambda\left(I_T \otimes W_N\right)Y + \left(1_T \otimes I_N\right)\mu + X\beta + \varepsilon$$

其中，λ为空间自回归系数，是我们需要预测的变量。μ表示的是个体效应，

随机效应模型中，$\mu = [\mu_1,...\mu_{TN}], \mu_i \sim N(0,\sigma_\mu^2)(i.i.d)$)；而固定效应中，$\mu_i$都是各自独立、固定的常数。

$\epsilon = [\epsilon_1,.....\epsilon_{NT}]$为正态残差项，$\epsilon_i$服从$(i.i.d)N(0,\sigma_\epsilon^2)$的正态分布。

涉及的参数为ρ,λ，以及方差比参数$\phi = \sigma_\mu^2/\sigma_\epsilon^2$

W_N为空间权重矩阵，$\otimes$表示矩阵的克罗内克积运算。

空间误差模型中的空间自回归项为空间滞后误差项，模型可表示为，

$$Y = (1_T \otimes I_N)\mu + X\beta + u$$

此处的u做为误差项包含了空间相关性，可将其表示为：

$$u = \rho(I_T \otimes W_N)u + \epsilon$$

代入上式后可知，空间误差模型和空间滞后模型的差别仅在于空间自回归项的位置。

在本文中，我们的研究对象是中国 31 省的林业发展状况，由于缺乏专业知识，我们无法在空间误差模型和空间滞后模型中选出更为合适的模型。直观上看，省份之间既存在空间相依性，又存在空间异质性，这两点都是我们研究的对象，因此在设计模型时，我们同时引入空间误差模型和空间滞后模型。

R 语言的 splm 包提供了我们所需的这一功能，其支持的空间面板数据分析的模型为：

$$Y = \lambda(I_T \otimes W_N)Y + (1_T \otimes I_N)\mu + X\beta + u$$

$$u = \rho(I_T \otimes W_N)u + \epsilon$$

其中的 splm()函数提供了模型中参数λ,ρ,ϕ,β的最大似然估计量（MLE）,我们将以此作为我们最终选定模型的参数来源，同时，在分析过程中，我们将根据不同参数的 $p - value$ 排除不必要的、解释能力差的变量，而保留必要的、解释能力强的变量，将一般化模型转化为一个合适的模型。当然，此前讨论过的固定效应模型和随机效应模型，我们将分别进行讨论，并通过比对其解释能力选出其

中较为合适的模型。

3.2. 空间权重矩阵

为了定义 31 省的空间距离，我们引入一个二元对称空间权重矩阵，参考聚类分析中用到的距离矩阵，该空间权重矩阵可凭欧氏距离或自定义的广义距离来刻画个体的临近关系。空间矩阵不唯一，也不存在最优情况，实际上不可能用一个矩阵充分表示现实中的空间结构，相比之下，简洁且便于计算的空间矩阵在空间面板数据模型中更为常用。空间相关性随距离的增加而减少，以这一点为原则，我们可以用地图距离倒数为元素构建矩阵，也可以构造二进制空间权重矩阵，即构成元素均为 0,1，构成定性矩阵，其特点是简洁，但是信息损失量较大。

我们使用常用的空间权重矩阵之一——距离矩阵作为W_N。数据来源为 https://download.csdn.net/download/megantan/10771264。矩阵每一行都要归一化，且在使用距离倒数矩阵时，可通过乘一个合适的常数使矩阵元素不至于过小，本次试验中我们使用了常数 $c = 1000$。我们也可以使用经济距离矩阵作为空间权重矩阵，经济距离矩阵需要联系地区人均 GDP 进行构造，会造成一定程度上的解释变量损失，且经济距离矩阵并不能很好地处理本文的数据。

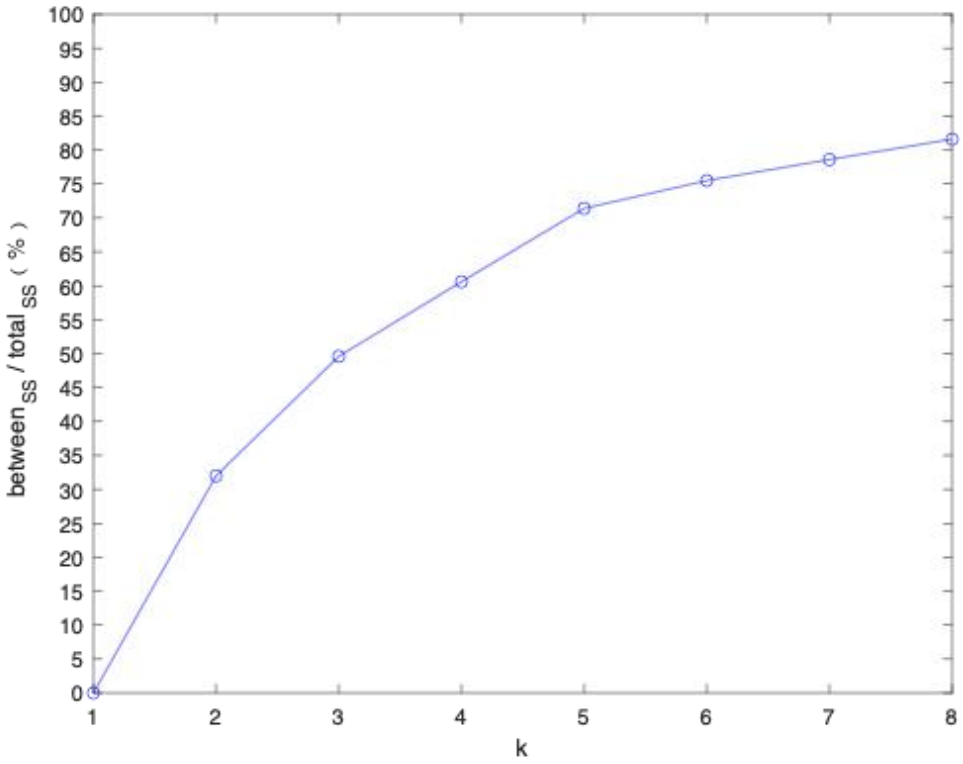
二．多元线性回归模型

1. 地区分类

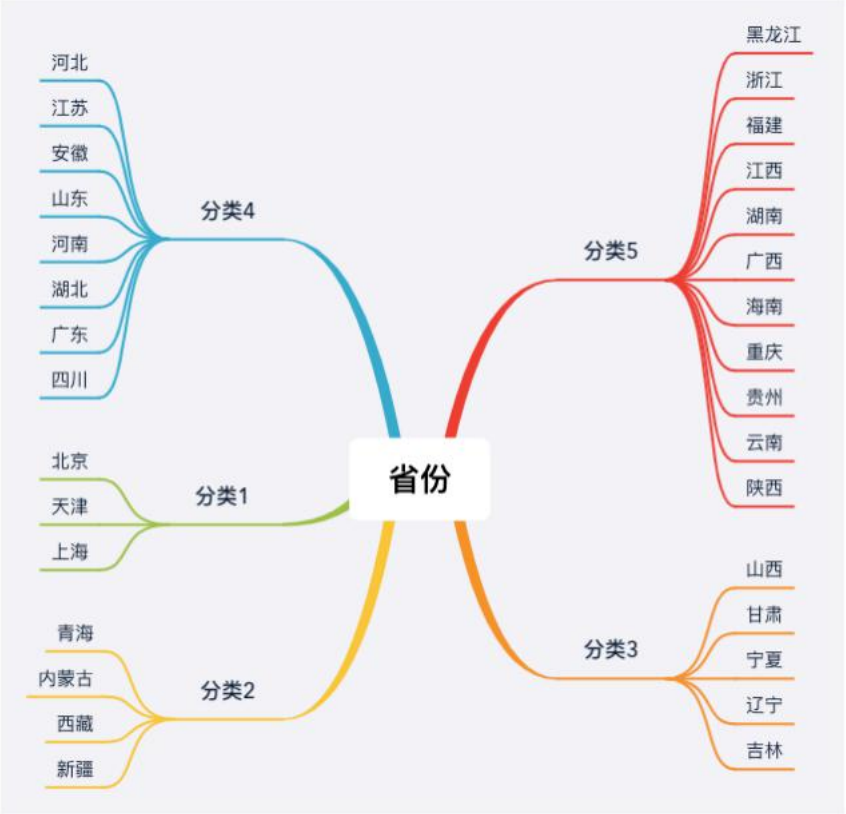
不同省份之间地理环境差异较大，各省也有着各自的经济发展模式和路线。为了能够更加有针对性且较为方便的探讨我国林业经济发展中林总产值的产业结构和因素分析，我们先试图对我国三十一个省份进行分类。

经过研究，我们选取各个省份的森林覆盖率、人均 GDP、人口总量、土地面积以及林总产值作为分类的依据。之所以选择以上几个特征，是因为一个地区的人口数量、土地面积和人均 GDP 能够较好的体现该地区的经济发展模式，而森林覆盖率、林总产值能够大致反映出林业在此地区的状况。我们采用 K-means 聚类法尝试对省份进行分类。其步骤是，预将数据分为 K 组，则随机选取 K 个对

象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心，迭代进行下去。为能够体现不同年份数据的信息，我们采用 1998-2017 的平均值作为该省份的分类值，为了避免不同类型数据单位的差异对分类造成影响我们对数据进行标准化处理。



采用 K-means 聚类法结果如上，由图可知，采用 k=5 是较好的选择。此时各个省份分类如下：



直观上来看，第 1 类的元素北京天津和上海是国际化大都市，土地面积小人口密度大，这些地区以发展城市经济为主要目标，林业的规模较小，同时也缺少大规模发展林业的条件。第 2 类的元素青海、内蒙古、西藏和新疆所在地理位置较为极端，林业发展一定的环境限制，同时地广人稀、经济发展水平和工业化水平较低。而第 3 类相对于第 2 类经济水平稍好，环境气候相对更适合林业的发展。由上分析可知，分类具有较好的解释性。

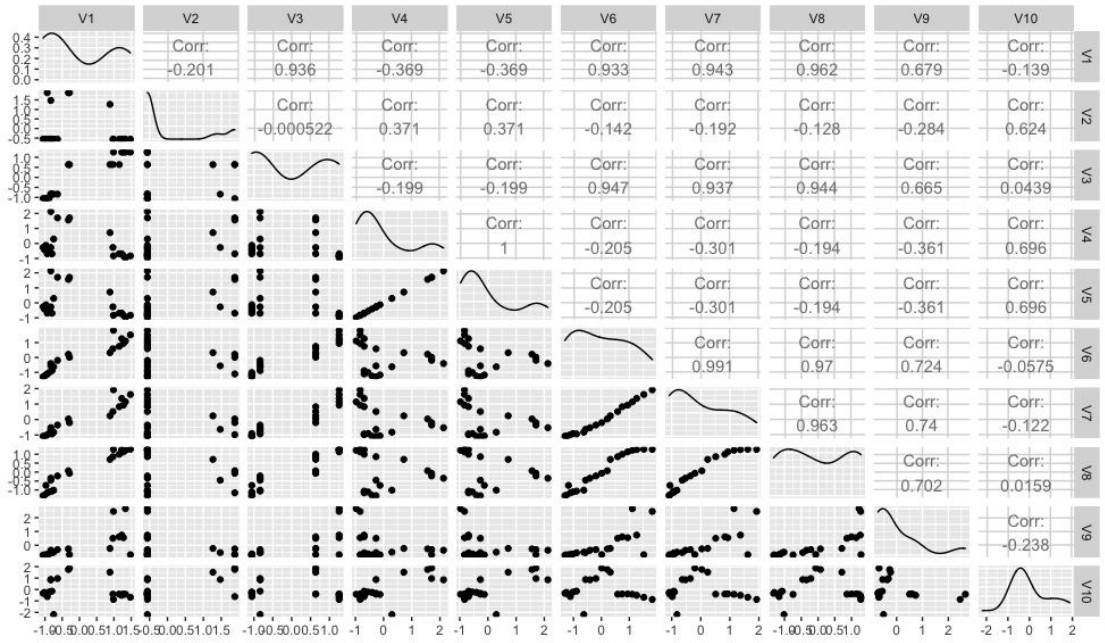
2. 回归分析

为了更好的分析我国林业经济发展，我们将各个省份分类五大类，下面对各个类别进行逐一的分析。为了论述方便，我们定义以下变量：Y 表示林业总产值，CFS 表示林业站的数量，FCR 表示森林覆盖率，FE 表示林业企业数量，Fer 表示林业从业人员，GDP 表示人均生产总值，IP 表示房地产规模，PS 表示人口数量，SIF 林业资金投入，NADS 表示一定规模以上的工业企业数量。因为不同变量之间数值差异很大，为了探究变量对 Y 的影响程度大小，我们对数据进行标准化处理。

未经过标准化处理的数据表示该年度的各省份该变量的总和。标准化数据后线性模型中系数解释为变量 x 变化一个标准差，y 变化多少个标准差。

2.1. 分类 1

我们首先对数据的探索性分析。



从图中可以清晰看出：GDP 和 IP（总资金投入）相关系数达到 0.991，一个地区的总资金投入的多少与人均生产总值高度相关，这与实际情况相吻合。PS（人口数量）与 GDP 和 IP（房地产规模）均高度相关，人口数量多的地方投资较大经济发展较快，这与改革开放以来的中国劳动型发展模式相符。值得注意的是，森林覆盖率和 GDP 相关系数达到 0.947，Y 变量出现了双峰分布。

我们先尝试将所有变量加入回归模型，其中 ε_i 是均值为 0 正态同分布的。为了行文流畅，在下面四类线性回归中不再重复说明。

$$Y_i = \beta_0 + \beta_1 CFS_i + \beta_2 FCR_i + \beta_3 FE_i + \beta_4 Fer_i + \beta_5 FCR_i + \beta_6 IP_i + \beta_7 PS_i + \beta_8 SIF_i + \beta_9 NADS_i + \varepsilon_i$$

```
Call:
lm(formula = Y_i ~ CFS_i + FCR_i + FE_i + fer_i + GDP_i + IP_i +
    PS_i + SIF_i + NADS_i)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20576 -0.03426  0.01358  0.05880  0.26656

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.664e-16  3.381e-02   0.000  1.00000
CFS_i       -3.812e-02  5.268e-02  -0.724  0.48580
FCR_i        3.463e-01  1.318e-01   2.627  0.02529 *
FE_i        -4.506e-01  1.289e-01  -3.495  0.00577 **
fer_i        3.721e-01  1.287e-01   2.892  0.01606 *
GDP_i        1.659e-01  4.712e-01   0.352  0.73215
IP_i        -3.938e-01  4.341e-01  -0.907  0.38559
PS_i         7.316e-01  1.818e-01   4.025  0.00242 **
SIF_i       -9.837e-02  5.403e-02  -1.820  0.09870 .
NADS_i      -1.773e-01  7.541e-02  -2.351  0.04055 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1512 on 10 degrees of freedom
Multiple R-squared:  0.988, Adjusted R-squared:  0.9771
F-statistic: 91.23 on 9 and 10 DF,  p-value: 2.081e-08
```

回归得到 $R^2 = 0.988, R_j^{\ 2} = 0.9771$，尽管拟合效果较好，但 9 个变量显著性 p 值只有两个两颗星和三个一颗星。这说明有些变量不适用于建模。为了看各变量是否存在共线性问题，我们查看 VIF 值。9 个变量中有 6 个的方差膨胀因子大于 10，有两个大于 100，这表明模型中存在很强的多重共线性问题。

CFS	FCR	FE	fer	GDP	IP	PS	SIF	NADS
2.31	14.44	13.81	13.76	184.51	156.59	27.47	2.427	4.73

接下来，我们以 AIC 信息统计量为准则，通过选择最小的 AIC 信息统计量，来进行变量选择。对选择后的模型进行回归得到以下结果：

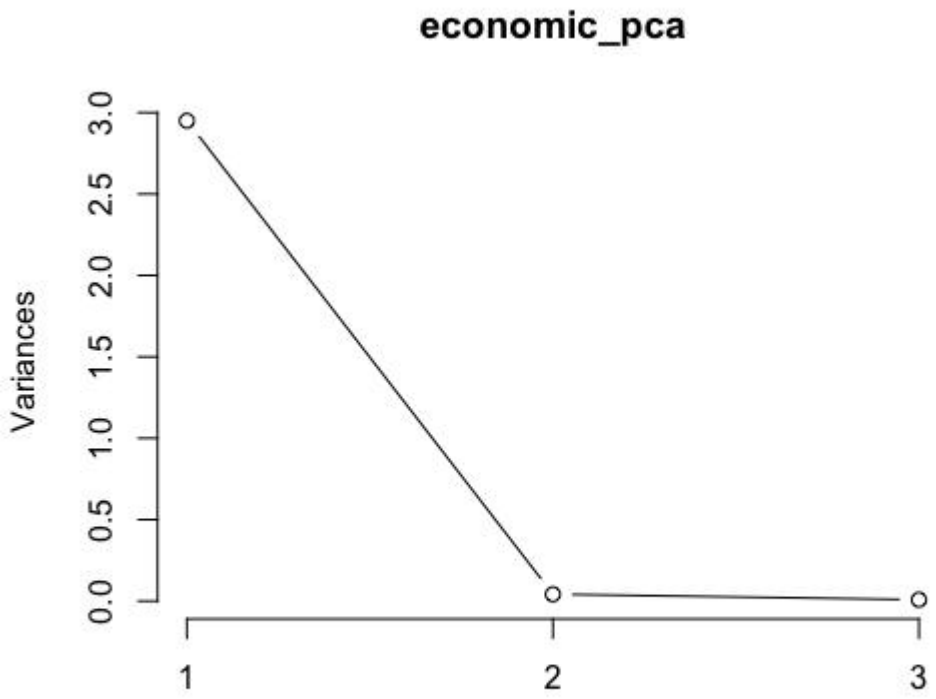
```
Call:
lm(formula = Y_i ~ FCR_i + FE_i + fer_i + IP_i + PS_i + SIF_i +
    NADS_i)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21607 -0.06002  0.01994  0.07438  0.24779

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.667e-16  3.171e-02   0.000 1.000000
FCR_i        3.225e-01  1.106e-01   2.917 0.012908 *
FE_i        -4.299e-01  9.824e-02  -4.376 0.000903 ***
fer_i        3.769e-01  1.165e-01   3.236 0.007145 **
IP_i        -2.425e-01  1.584e-01  -1.531 0.151734
PS_i         7.735e-01  1.571e-01   4.925 0.000351 ***
SIF_i       -9.283e-02  5.008e-02  -1.854 0.088519 .
NADS_i      -2.087e-01  5.868e-02  -3.556 0.003955 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1418 on 12 degrees of freedom
Multiple R-squared:  0.9873,    Adjusted R-squared:  0.9799
F-statistic: 133.3 on 7 and 12 DF,  p-value: 2.035e-10
```

对比之前的回归结果，我们可以看到变量的显著性变强，我们查看 VIF 值。

FCR	FE	fer	IP	PS	SIF	NADS
11.55	9.12	12.82	23.71	23.31	2.37	3.25

七个变量中有 4 个方差膨胀因子大于 10，模型的共线性问题仍然存在。

面对多重共线性问题时，我们一般除了变量选择的方法之外，常用的有岭回归和 Robust 回归的方法，值得一提的是这两种方法对数据量大的情形较为有效。这里由于数据量较少，所以我们尝试使用 PCA 方法进行降维。我们首先尝试了对所有变量进行 PCA 分析得到前三个主成分并进行回归，发现结果并不好，同时模型也缺乏解释性。经过思考，我们选择将相关性较高 GDP、IP（房地产规模）和 PS（人口数量）放在一类提取第一主成分。



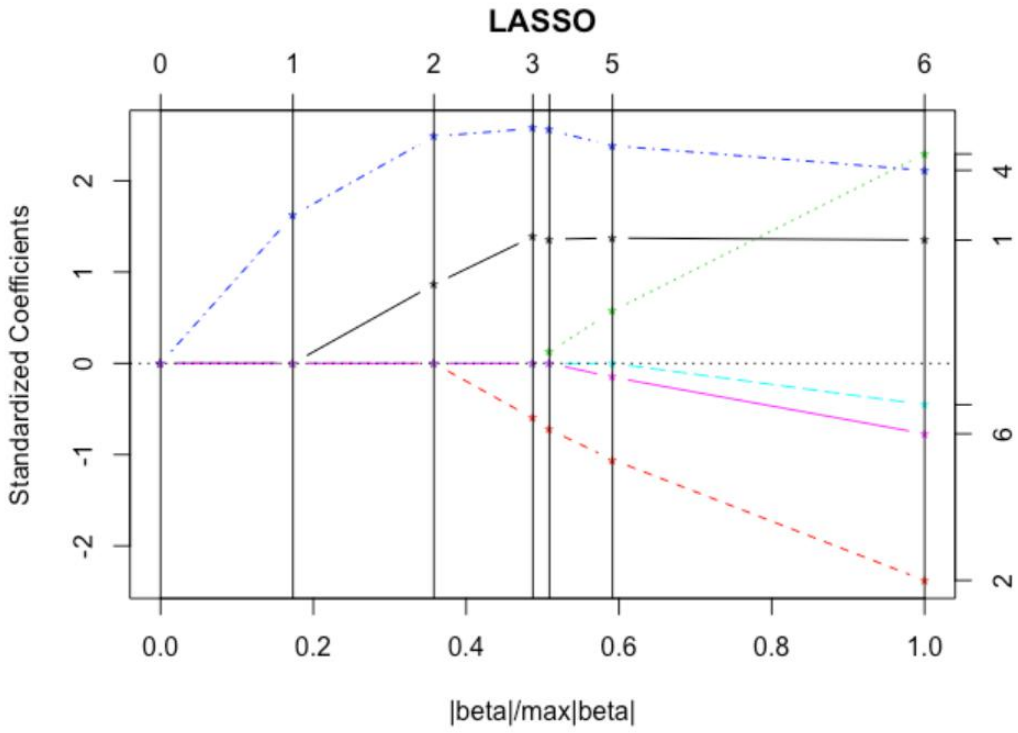

```
Standard deviations (1, .., p=3):
[1] 1.71750036 0.20508918 0.09017166

Rotation (n x k) = (3 x 3):
          PC1        PC2        PC3
V1 0.5797111 -0.3106536 -0.7532790
V2 0.5782367 -0.4944978  0.6489332
V3 0.5740883  0.8117674  0.1070345
```

毫不意外的，第一主成分占走了绝大多数方差，第一主成分解释的总体总方差比例为: $(1.71750036^2)/3 = 0.9832692$ ，我们把它解释为整体经济发展因子，可用来描述一个地区的经济发展总量的强弱。

```
Standard deviations (1, .., p=2):
[1] 1.1192716 0.8644253

Rotation (n x k) = (2 x 2):
         PC1        PC2
V1 0.7071068 -0.7071068
V2 0.7071068  0.7071068
```

我们用整体经济发展因子代替对应的三个变量，发现仍存在变量不显著和多

重共线性的问题。起初我们自然的想到用 FE（林业企业数量）和 Fer（林业从业人员）放到一类做主成分分析。第一主成分解释的总体总方差比例为 $1.3536596^2/2 = 0.9161972$，我们把它解释为林业规模因子，用来描述一个地区林产业的规模。加入林业规模因子后我们进行回归发现结果并不好：共线性问题仍然存在并且由于损失了部分信息，模型拟合度下降。我们选择不加入林业规模因子，而是使用 lasso 进行变量选择。



根据上图我们选择加入变量 FE（林业企业数量）、fer（林业从业人员）和 econ 进行回归：

```
Call:
lm(formula = Y_i ~ FE_i + fer_i + econ_i)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49316 -0.08837  0.03099  0.09587  0.39482

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.482e-16  5.292e-02   0.000  1.00000
FE_i        -5.252e-01  1.513e-01  -3.471  0.00315 **
fer_i        3.983e-01  1.505e-01   2.646  0.01761 *
econ_i       4.341e-01  4.997e-02   8.686 1.88e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2367 on 16 degrees of freedom
Multiple R-squared:  0.9528,    Adjusted R-squared:  0.944
F-statistic: 107.7 on 3 and 16 DF,  p-value: 8.008e-11
```

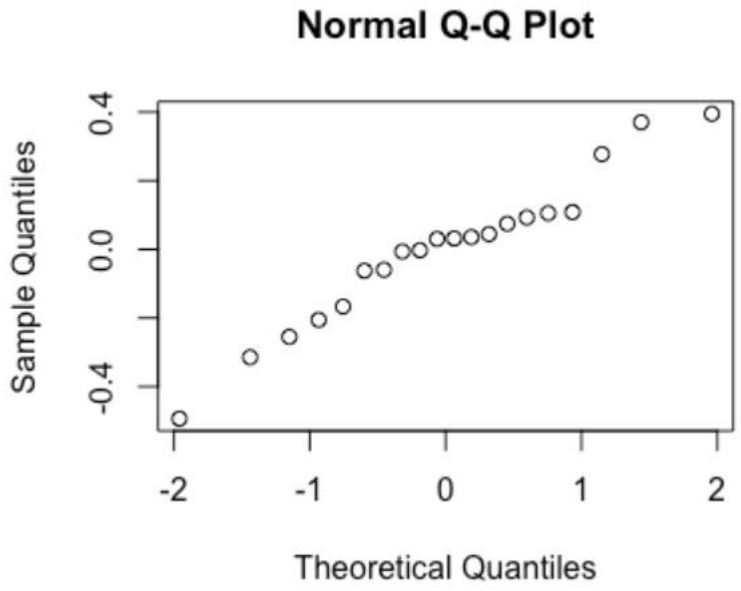
我们查看 VIF 值：

FE	Fer	Econ
7.77	7.69	2.50

相比于之前有较好的改观。我们再来进行同方差检验：

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.8303803, Df = 1, p = 0.36216
```

进行同方差检验发现 p = 0.36216>0.05 满足原假设。

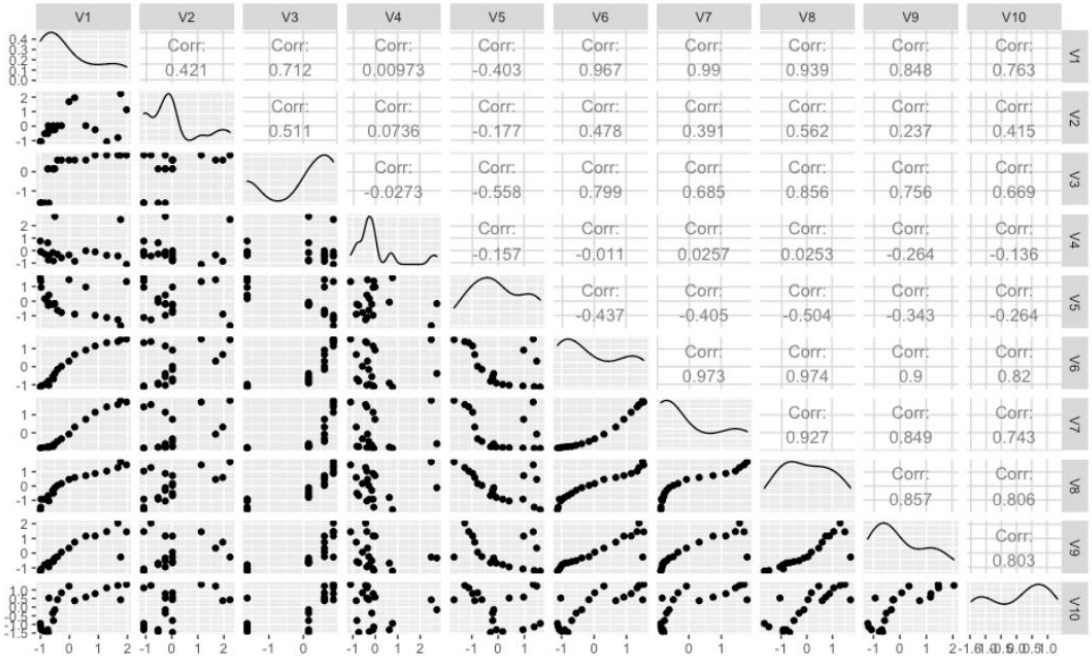


我们通过 Q-Q 图检验残差的正态性，图像近似一条过原点的直线，结果较令人满意。最终得到：

$$\frac{Y_i-2712451}{2343861}=-2.482\times10^{-16}-0.5252\frac{FE_i-6.9}{7.1}+0.3983\frac{Fer_i-15068.98}{8923.049}$$
$$+18.3085\left(0.5797\frac{GDP_i-190398.8}{101131.9}+0.4341\frac{\mathrm{IP}_i-303951848}{220494605}+0.5741\frac{\mathrm{PS}_i-4828.7}{1035.774}\right)$$

2.2. 分类 2

类似上面的分析步骤，我们首先进行探索性分析。



从图中可以看出，PS（人口数量）、GDP、IP（房地产规模）、SIF（林业资金投入）和 NADS（一定规模以上的工业企业数量）之间相关性较高，这与预期相符。而上一类中 NADS 与前三者相关性不高的原因在于北津上这样的国际都市由于土地价格昂贵、污染环境等因素，大规模企业较少。而上一类中 SIF 与前三者相关性不高的原因在于北津上林业规模较小，经济投入相对于整体经济水平势微。

我们先尝试将所有变量拿来预测：

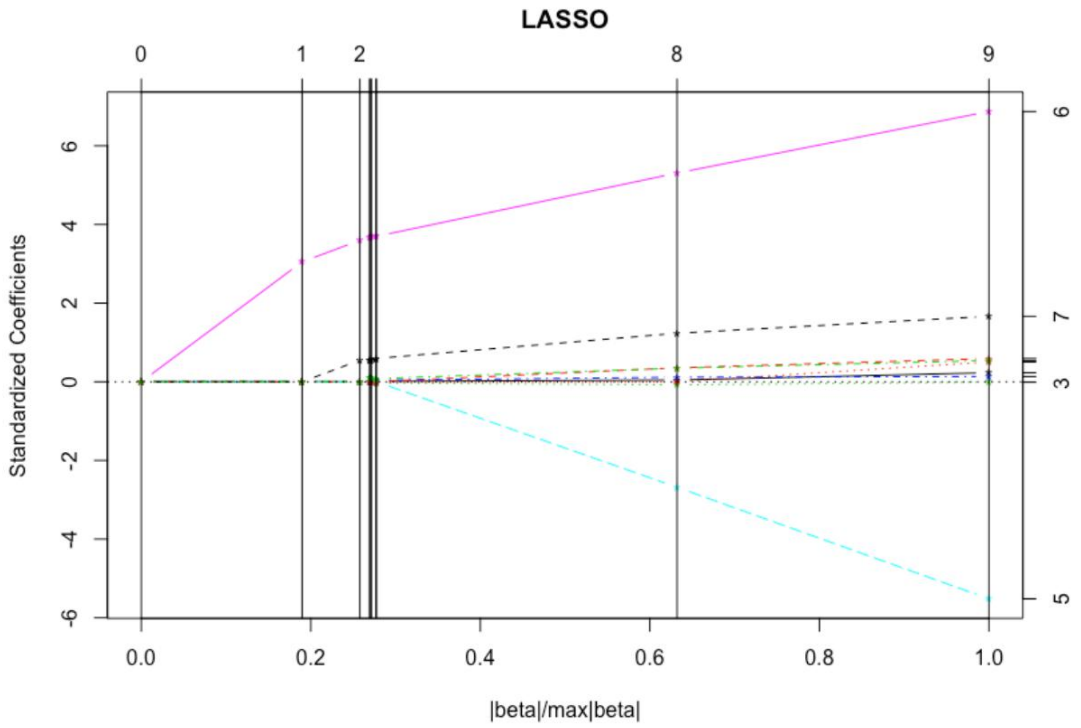
$$Y_i = \beta_0 + \beta_1 CFS_i + \beta_2 FCR_i + \beta_3 FE_i + \beta_4 Fer_i + \beta_5 FCR_i + \beta_6 IP_i + \beta_7 PS_i + \beta_8 SIF_i + \beta_9 NADS_i + \varepsilon_i$$

```
Call:
lm(formula = Y_i ~ CFS_i + FCR_i + FE_i + fer_i + GDP_i + IP_i +
    PS_i + SIF_i + NADS_i)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11279 -0.04799  0.01754  0.04797  0.11077

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.438e-16  2.092e-02   0.000  1.00000
CFS_i        5.300e-02  4.265e-02   1.243  0.24234
FCR_i        1.360e-01  5.984e-02   2.272  0.04642 *
FE_i        -2.372e-03  2.882e-02  -0.082  0.93602
fer_i        3.079e-02  2.951e-02   1.043  0.32138
GDP_i       -1.266e+00  2.784e-01  -4.548  0.00106 **
IP_i         1.577e+00  1.723e-01   9.151 3.56e-06 ***
PS_i         3.819e-01  1.393e-01   2.741  0.02079 *
SIF_i        1.157e-01  9.125e-02   1.268  0.23348
NADS_i       1.243e-01  4.577e-02   2.716  0.02170 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09356 on 10 degrees of freedom
Multiple R-squared:  0.9954,    Adjusted R-squared:  0.9912
F-statistic: 240.1 on 9 and 10 DF,  p-value: 1.75e-10
```

有些变量性 p 值较大，说明有些变量不适用于建模。值得注意的是，IP 的显著性很好，系数很大达到了 1.577，这与青海、内蒙古、西藏、新疆不适合发展工业，投资对林业影响较大的事实相符。但 SIF（林业资金投入）并不显著，这可能是因为很多投资造成的林业影响是间接的，没有被记入 SIF 中。为了看各变量是否存在共线性问题，我们查看 VIF 值。发现有 4 个变量的方差膨胀因子大于 10，又一个大于 100.这说明模型中有很强的共线性问题。

CFS	FCR	FE	fer	GDP	IP	PS	SIF	NADS
3.95	7.77	1.80	1.89	168.25	64.44	42.12	18.07	4.54

接下来，我们以 AIC 信息统计量为准则，通过选择最小的 AIC 信息统计量，来进行变量选择，调用 R 语言中 step( )函数后发现结果并不好，模型的共线性没有明显改善。我们尝试用 lasso 进行变量选择。



选择加入变量 GDP 和 IP 进行回归。

```
Call:
lm(formula = Y_i ~ GDP_i + IP_i)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23999 -0.12625  0.03343  0.08691  0.29819

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.395e-16  3.254e-02   0.000     1.00
GDP_i       6.099e-02  1.451e-01   0.420     0.68
IP_i        9.310e-01  1.451e-01   6.415  6.4e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1455 on 17 degrees of freedom
Multiple R-squared:  0.9811,    Adjusted R-squared:  0.9788
F-statistic: 440.1 on 2 and 17 DF,  p-value: 2.289e-15
```

回归结果显示 $R^2=0.9811, R_j^{\ 2}=0.9788$，令人满意。

我们查看 VIF 值：

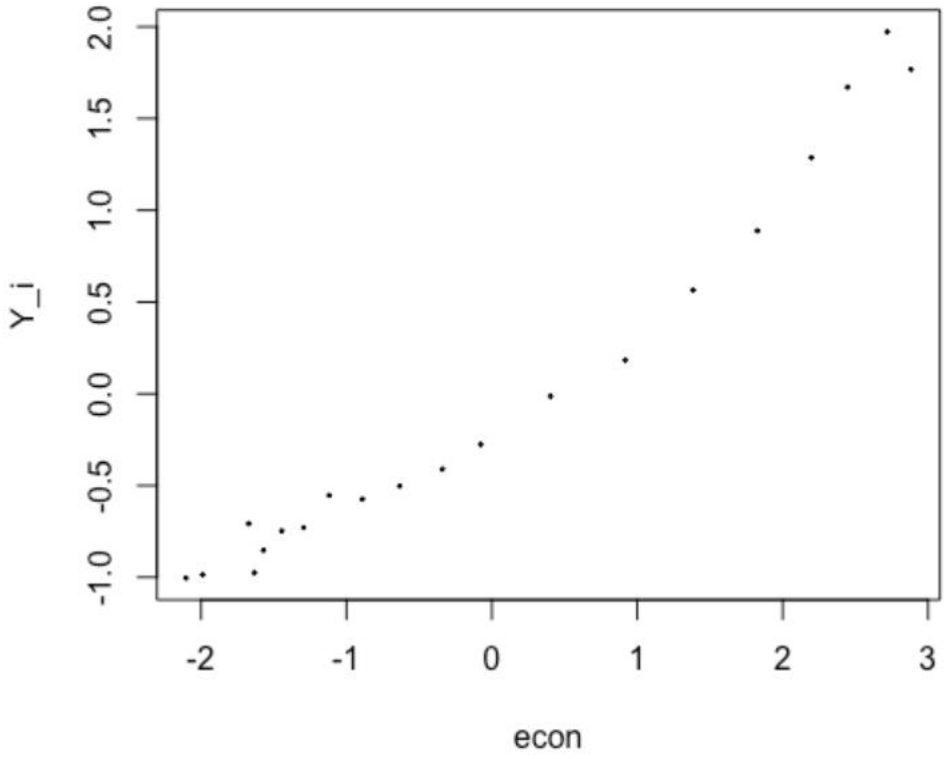
GDP	IP
18.89	18.89

尽管只有两个变量，两变量 VIF 值均大于 10，模型存在共线性问题。

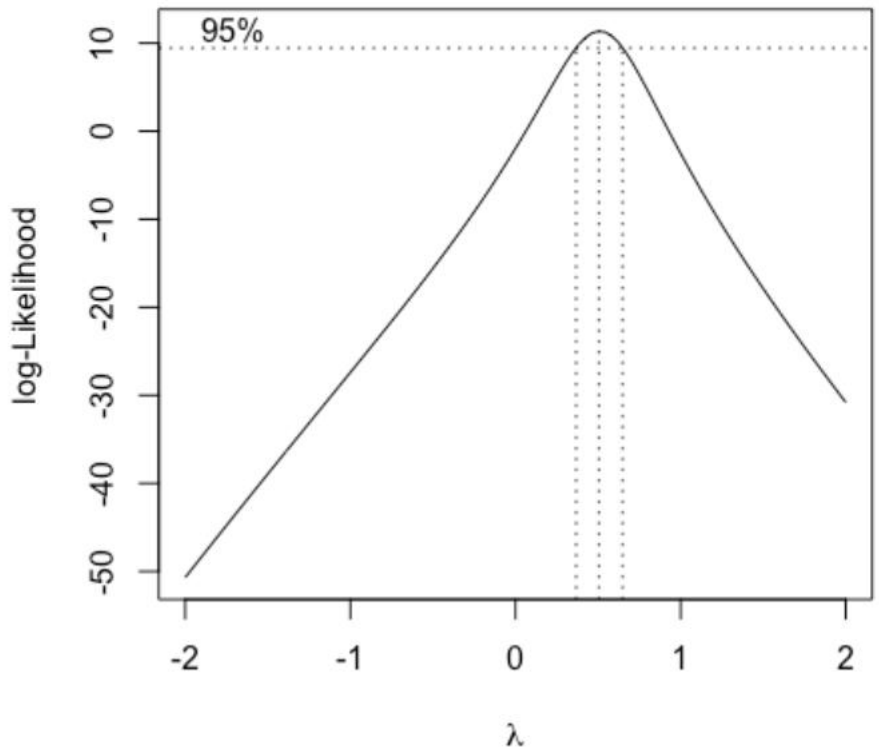
为了与全文保持一致，我们不妨尝试对 PS（人口数量）、GDP、IP（房地产规模）进行主成分分析提取经济发展因子 econ。

```
Standard deviations (1, .., p=3):
[1] 1.7077420 0.2707640 0.1015083

Rotation (n x k) = (3 x 3):
         PC1          PC2         PC3
V1 0.5835741 -0.009286067  0.8120068
V2 0.5740828  0.711938586 -0.4044409
V3 0.5743433 -0.702180332 -0.4207999
```

毫不意外的，第一主成分占走了绝大多数方差，第一主成分解释的总体总方差比例为:$1.7077420^2/3=0.9721276$。然后我们通过散点图看下经济发展因子 econ 和 Y 之间的关系。

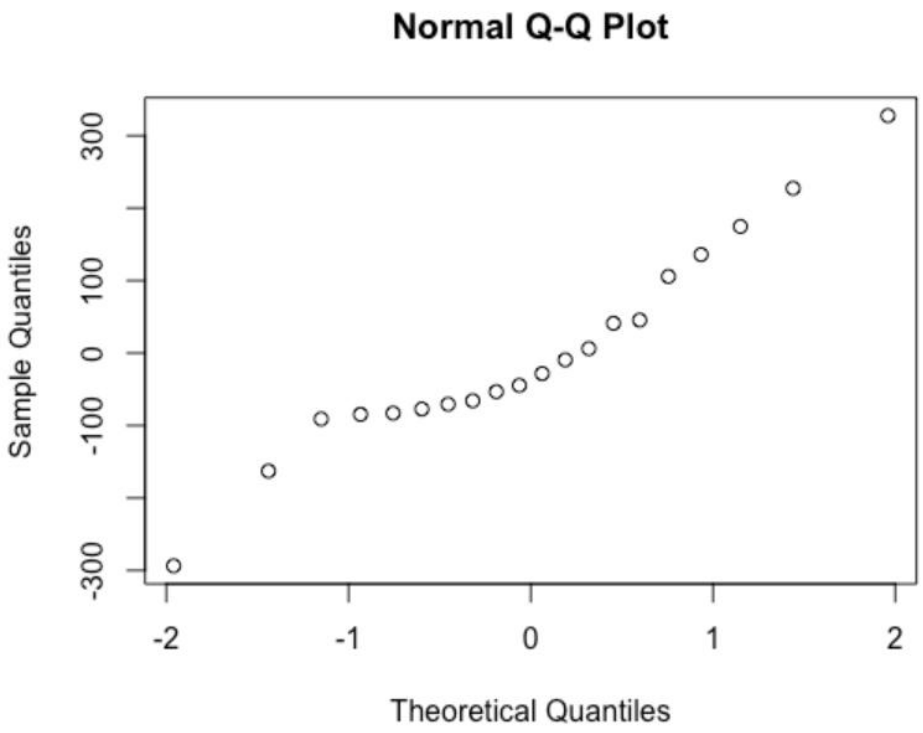


由图我们知道用线性模型拟合效果不好，所以我们选择使用 Box-Cox 变换，得到下图。当$\lambda=0.50505051$时取最大。值得一提的是，由于 Box-Cox 变换要求 y 大于零以及基于 Box-Cox 的原理，我们对 Y 不再标准化处理。




```
Call:
lm(formula = (y_i)^(0.50505051) ~ econ)

Residuals:
    Min      1Q  Median      3Q     Max
-293.68  -78.66  -36.47   60.55  327.61

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2399.00      32.45   73.93   <2e-16 ***
econ          651.93      19.50   33.44   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 145.1 on 18 degrees of freedom
Multiple R-squared:  0.9842,    Adjusted R-squared:  0.9833
F-statistic:  1118 on 1 and 18 DF,  p-value: < 2.2e-16
```

显然的，进行同方差检验发现 p = 0.23801>0.05 满足原假设。

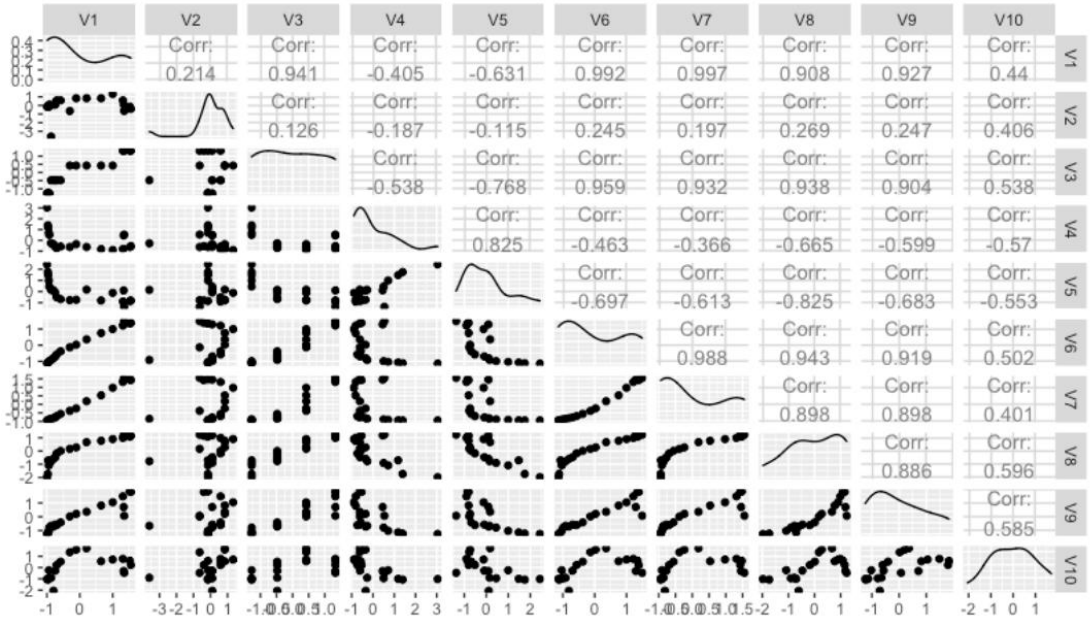


我们通过 Q-Q 图检验残差的正态性，图像近似一条过原点的直线，结果较令人满意。最后我们得到：

$$Y_i^{0.50505051} = 2399.00 + 651.93\left(0.5836\frac{GDP_i - 93628.15}{65619.59} + 0.5741\frac{IP_i - 53551204}{60671648} + 0.5743\frac{PS_i - 5372.35}{316.7442}\right)$$

2.3. 分类 3

我们首先对数据的探索性分析。



从图中我们可以发现，FCR（森林覆盖率）、GDP、IP（房地产规模）、PS（人口数量）、SIF（林业资金投入）之间高度相关，这之所以其他几类不同应该是地区的经济发展模式决定的。这侧面印证了分类的合理性。

我们先尝试将所有变量拿来预测。

```
Call:
lm(formula = Y_i ~ CFS_i + FCR_i + FE_i + fer_i + GDP_i + IP_i +
    PS_i + SIF_i + NADS_i)

Residuals:
      Min        1Q    Median        3Q       Max
-0.045295 -0.015504 -0.001216  0.015514  0.031424

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.919e-17  7.143e-03   0.000  1.00000
CFS_i       -2.284e-03  1.023e-02  -0.223  0.82773
FCR_i       -6.973e-03  3.561e-02  -0.196  0.84867
FE_i         2.391e-02  2.244e-02   1.066  0.31165
fer_i        1.607e-02  2.175e-02   0.739  0.47712
GDP_i        3.495e-01  1.352e-01   2.585  0.02716 *
IP_i         5.315e-01  1.107e-01   4.803  0.00072 ***
PS_i        -1.264e-02  4.100e-02  -0.308  0.76420
SIF_i        1.846e-01  2.899e-02   6.369 8.15e-05 ***
NADS_i      -2.210e-02  1.422e-02  -1.554  0.15114
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03195 on 10 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.999
F-statistic:  2068 on 9 and 10 DF,  p-value: 3.815e-15
```

回归得到，$R^2=0.9995, R_j^2=0.999$ 尽管拟合效果极好，但 9 个变量显著性 p 值只有 2 个三颗星和 1 个一颗星。这说明有些变量不适用于建模。为了看各变量是否存在共线性问题，我们查看 VIF 值。

CFS	FCR	FE	fer	GDP	IP	PS	SIF	NADS
1.95	23.61	9.38	8.81	340.20	228.00	31.30	15.65	3.76

接下来，我们以 AIC 信息统计量为准则，通过选择最小的 AIC 信息统计量，来进行变量选择。对选择后的模型进行回归得到以下结果：

```
Call:
lm(formula = Y_i ~ FE_i + GDP_i + IP_i + SIF_i + NADS_i)

Residuals:
      Min        1Q    Median        3Q       Max
-0.049261 -0.013628 -0.001274  0.019522  0.039093

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.678e-17  6.275e-03   0.000  1.00000
FE_i         3.692e-02  1.238e-02   2.983  0.00989 **
GDP_i        2.778e-01  8.493e-02   3.271  0.00558 **
IP_i         5.704e-01  8.564e-02   6.661 1.08e-05 ***
SIF_i        1.934e-01  2.197e-02   8.805 4.41e-07 ***
NADS_i      -2.059e-02  1.213e-02  -1.698  0.11169
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02806 on 14 degrees of freedom
Multiple R-squared:  0.9994,    Adjusted R-squared:  0.9992
F-statistic:  4822 on 5 and 14 DF,  p-value: < 2.2e-16
```

结果较令人满意，但我们查看 VIF 值：

FE	GDP	IP	SIF	NADS
3.70	174.00	176.93	11.64	3.55

发现 GDP 和 IP 的方差膨胀因子远大于 10，为了与前文统一，我们通过主成分分析提取整体经济发展因子 econ。

```
Standard deviations (1, .., p=3):
[1] 1.69905641 0.32996602
[3] 0.06580089

Rotation (n x k) = (3 x 3):
         PC1        PC2        PC3
V1 0.5864652  0.2021398  0.7843456
V2 0.5774456  0.5747158 -0.5798779
V3 0.5679922 -0.7929951 -0.2203261
```

第一主成分解释的总体总方差比例为: $(1.69905641^2)/3 = 0.9622642$ ，令人满意。我们用得到的 econ 代替 GDP 和 IP 进行回归：

```
Call:
lm(formula = Y_i ~ FE_i + econ_i + SIF_i + NADS_i)

Residuals:
     Min       1Q   Median       3Q      Max
-0.10471 -0.03998 -0.01169  0.04285  0.16269

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.176e-16  1.574e-02   0.000  1.00000
FE_i         1.571e-01  2.159e-02   7.278 2.70e-06 ***
econ_i       4.700e-01  2.433e-02  19.314 5.23e-12 ***
SIF_i        3.274e-01  4.556e-02   7.186 3.15e-06 ***
NADS_i      -6.860e-02  2.119e-02  -3.238  0.00552 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07041 on 15 degrees of freedom
Multiple R-squared:  0.9961,    Adjusted R-squared:  0.995
F-statistic: 954.4 on 4 and 15 DF,  p-value: < 2.2e-16
```

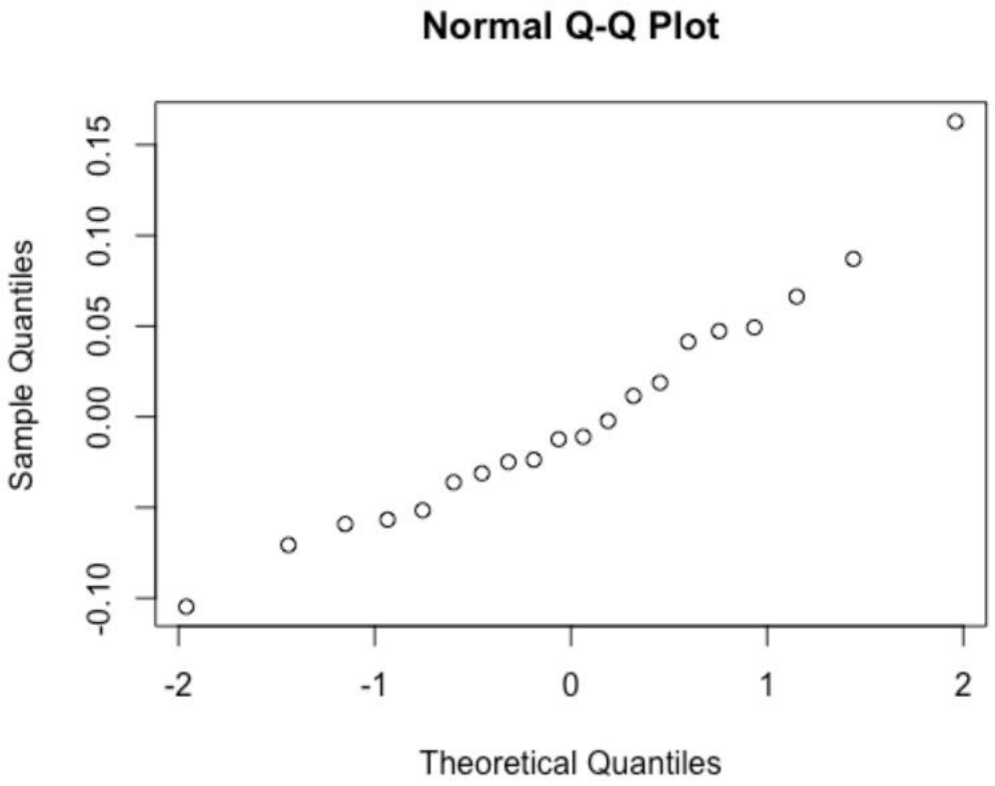
查看 VIF 值：

FE	econ	SIF	NADS
1.79	6.55	7.96	1.72

方差膨胀因子都小于 10，结果令人满意。我们再来进行同方差检验：

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.529766, Df = 1, p = 0.060276
```

P 值大于 0.05，接受原假设。

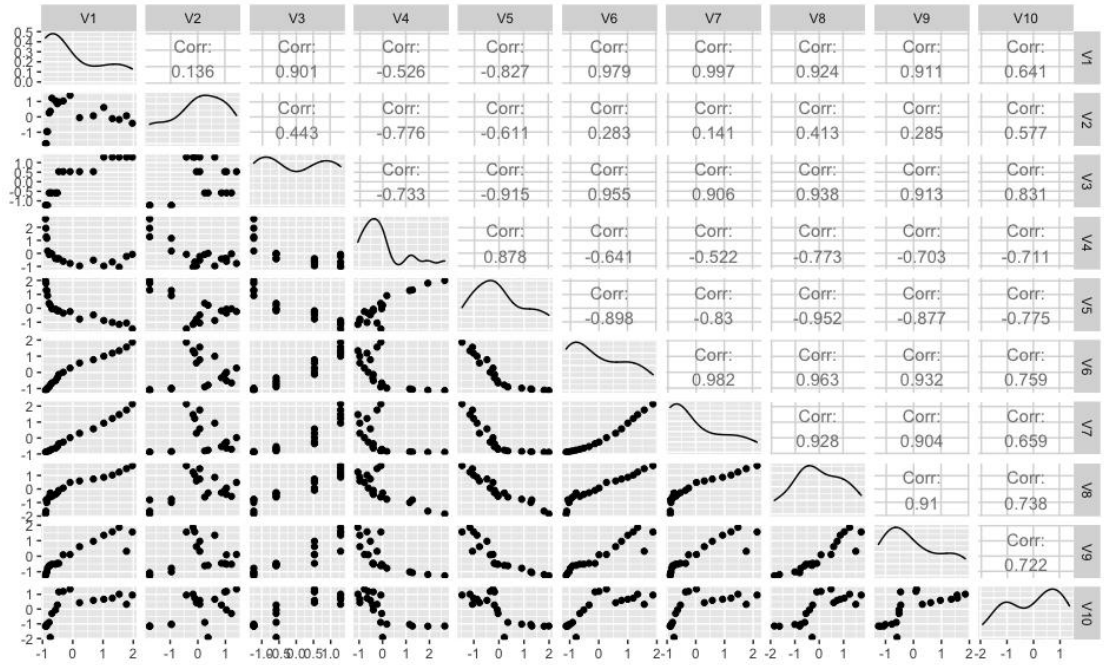


我们通过 Q-Q 图检验残差的正态性，图像近似一条过原点的直线，结果较令人满意。最后我们得到：

$$\frac{Y_i-18071197}{15468270}=3.176\times10^{-16}+0.1571\frac{FE_i-369.4}{89.27155}+0.3274\frac{\mathrm{SIF}_i-1184550}{902539.6}$$
$$-0.06860\frac{\mathrm{NADS}_i-23360}{8110.578}+0.47\left(0.5865\frac{GDP_i-115783.7}{76393.16}+0.5774\frac{IP_i-186178121}{186881981}+0.568\frac{PS_i-13645.1}{366.1108}\right)$$

2.4. 分类 4

我们首先对数据的探索性分析。



从图中我们可以发现，FCR（森林覆盖率）、GDP、IP（房地产规模）、PS（人口数量）、SIF（林业资金投入）之间高度相关，这与上一类相同。值得注意的是，FCR 与所有其他自变量的相关性都很强。

我们先尝试将所有变量拿来预测：

```
Call:
lm(formula = Y_i ~ CFS_i + FCR_i + FE_i + fer_i + GDP_i + IP_i +
    PS_i + SIF_i + NADS_i)

Residuals:
      Min        1Q    Median        3Q       Max
-0.085951 -0.037290  0.005704  0.036442  0.078413

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.593e-16  1.508e-02   0.000 1.000000
CFS_i        4.747e-02  4.539e-02   1.046 0.320263
FCR_i       -4.524e-02  7.196e-02  -0.629 0.543660
FE_i        -1.281e-01  7.519e-02  -1.703 0.119374
fer_i        1.464e-01  1.062e-01   1.378 0.198260
GDP_i        1.747e-01  2.092e-01   0.835 0.423263
IP_i         1.077e+00  1.750e-01   6.157 0.000107 ***
PS_i        -1.672e-01  1.149e-01  -1.455 0.176211
SIF_i        5.316e-02  5.832e-02   0.911 0.383532
NADS_i      -8.375e-02  3.904e-02  -2.145 0.057516 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06744 on 10 degrees of freedom
Multiple R-squared:  0.9976,    Adjusted R-squared:  0.9955
F-statistic:   463 on 9 and 10 DF,  p-value: 6.675e-12
```

只有一个变量是显著的，这说明有些变量不适用于建模。同上的，以 AIC 信息统计量为准则，选择最小的 AIC 来进行变量选择。对选择后的模型进行回归得到以下结果

```
Call:
lm(formula = Y_i ~ FE_i + IP_i + PS_i + NADS_i)

Residuals:
      Min        1Q    Median        3Q       Max
-0.114845 -0.029292  0.000609  0.030832  0.135331

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.161e-16  1.407e-02   0.000   1.0000
FE_i        -1.189e-01  4.435e-02  -2.681   0.0171 *
IP_i         1.165e+00  7.181e-02  16.227 6.37e-11 ***
PS_i        -2.053e-01  9.350e-02  -2.196   0.0442 *
NADS_i      -5.971e-02  2.379e-02  -2.510   0.0240 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06292 on 15 degrees of freedom
Multiple R-squared:  0.9969,    Adjusted R-squared:  0.996
F-statistic:  1196 on 4 and 15 DF,  p-value: < 2.2e-16
```

效果较好，我们查看 VIF 值：

FE	IP	PS	NADS
9.44	24.74	41.96	2.72

发现 GDP 和 IP 的方差膨胀因子大于 10，为了与前文统一，我们通过主成分分析提取整体经济发展因子 econ。

```
Standard deviations (1, .., p=3):
[1] 1.7075093 0.2719239 0.1023202

Rotation (n x k) = (3 x 3):
         PC1        PC2        PC3
V1 0.5832451 -0.1421439 -0.7997627
V2 0.5763507 -0.6213832  0.5307569
V3 0.5724029  0.7705050  0.2804939
```

第一主成分解释的总体总方差比例为:$(1.7075093^2)/3 = 0.9718627$，令人满意。我们用得到的 econ 代替 IP 和 PS 进行回归：

```
Call:
lm(formula = Y_i ~ FE_i + econ_i + NADS_i)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17271 -0.09025  0.01412  0.06984  0.21630

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.670e-16  2.795e-02   0.000 1.000000
FE_i         1.718e-01  4.250e-02   4.041 0.000946 ***
econ_i       6.715e-01  2.560e-02  26.234 1.41e-14 ***
NADS_i      -7.344e-02  4.696e-02  -1.564 0.137430
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.125 on 16 degrees of freedom
Multiple R-squared:  0.9868,    Adjusted R-squared:  0.9844
F-statistic: 399.9 on 3 and 16 DF,  p-value: 2.988e-15
```

我们发现加入整体经济发展因子 econ 后，NADS 变得并不显著，我们尝试去掉 NADS，发现 R^2 和 R_j^2 几乎不变。

```
Call:
lm(formula = Y_i ~ FE_i + econ_i)

Residuals:
      Min        1Q    Median        3Q       Max
-0.223162 -0.107428  0.007281  0.095408  0.220125

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.697e-16  2.912e-02   0.000        1
FE_i         2.017e-01  3.951e-02   5.105 8.80e-05 ***
econ_i       6.516e-01  2.314e-02  28.158 1.05e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1302 on 17 degrees of freedom
Multiple R-squared:  0.9848,    Adjusted R-squared:  0.983
F-statistic: 551.7 on 2 and 17 DF,  p-value: 3.46e-16
```

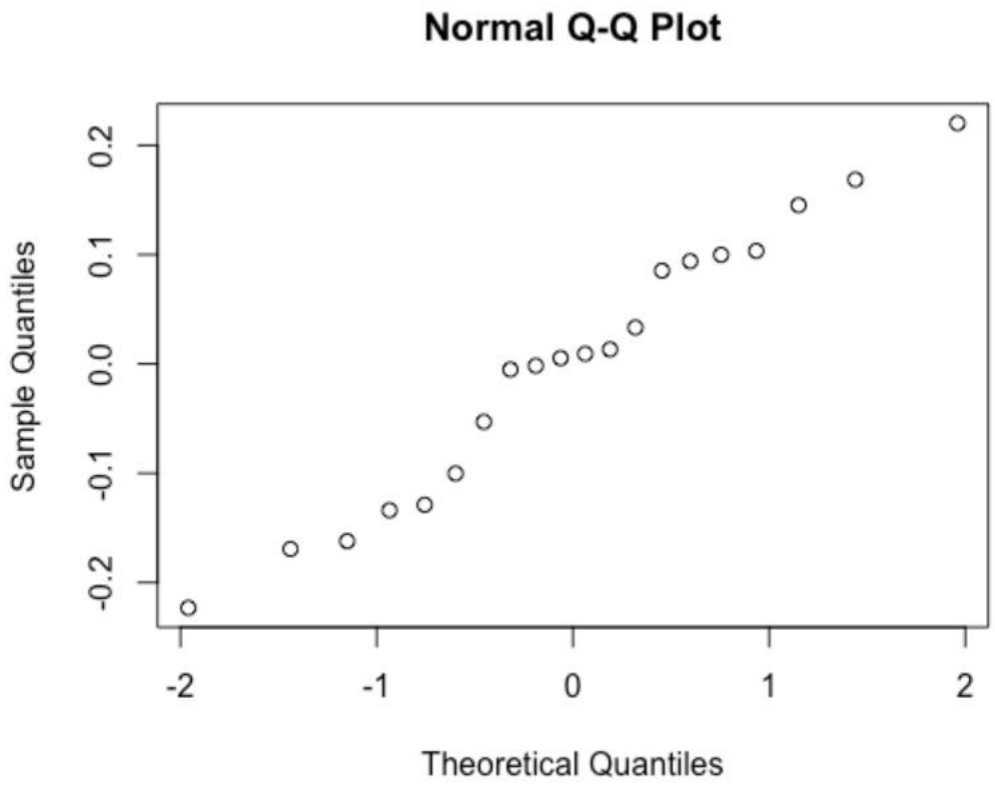
查看 VIF 值：

FE	econ
1.75	1.75

方差膨胀因子都小于 10，结果令人满意。我们再来进行同方差检验：

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.296708, Df = 1, p = 0.12965
```

P 值大于 0.05，接受原假设。

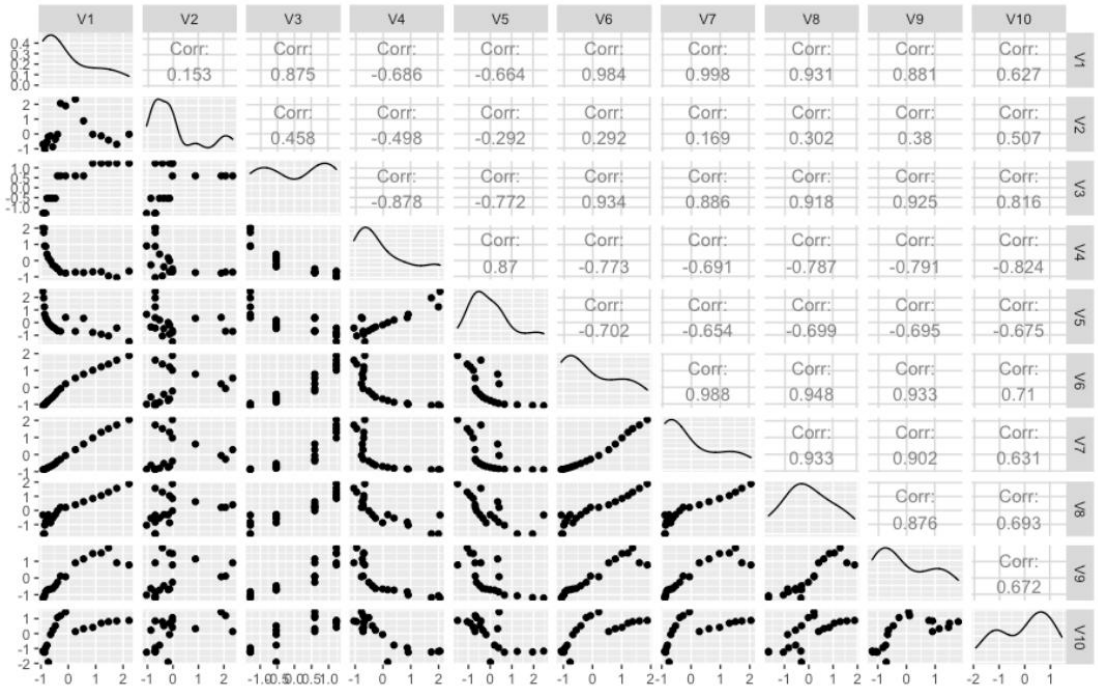


我们通过 Q-Q 图检验残差的正态性，图像近似一条过原点的直线，结果较令人满意。最后我们得到：

$$\frac{Y_i - 112199194}{112877290} = -1.697\times10^{-16} + 2.017\frac{FE_i - 737.15}{297.3885}$$
$$+ 0.6516\left(0.5832\frac{GDP_i - 222011.5}{149674.8} + 0.5764\frac{IP_i - 967124561}{1008124051} + 0.5724\frac{PS_i - 63241.9}{2031.344}\right)$$

2.5. 分类 5

我们首先对数据的探索性分析。



从图中我们可以发现，FCR（森林覆盖率）、GDP、IP（房地产规模）、PS（人口数量）、SIF（林业资金投入）之间高度相关。这种情况与前面也有出现。

我们先尝试将所有变量拿来预测：

```
Call:
lm(formula = Y_i ~ CFS_i + FCR_i + FE_i + fer_i + GDP_i + IP_i +
    PS_i + SIF_i + NADS_i)

Residuals:
      Min        1Q    Median        3Q       Max
-0.047888 -0.009951  0.001096  0.011730  0.031663

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.980e-17  5.843e-03   0.000  1.00000
CFS_i        6.441e-03  1.341e-02   0.480  0.64129
FCR_i       -5.792e-02  2.878e-02  -2.012  0.07188 .
FE_i        -1.692e-03  2.627e-02  -0.064  0.94990
fer_i       -4.673e-02  1.374e-02  -3.401  0.00676 **
GDP_i        1.494e-01  1.418e-01   1.053  0.31691
IP_i         9.873e-01  1.279e-01   7.720 1.61e-05 ***
PS_i         3.997e-03  2.207e-02   0.181  0.85989
SIF_i       -1.305e-01  2.090e-02  -6.246 9.55e-05 ***
NADS_i      -6.661e-03  1.337e-02  -0.498  0.62900
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02613 on 10 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9993
F-statistic:  3091 on 9 and 10 DF,  p-value: 5.114e-16
```

回归得到，$R^2 = 0.9996$，$R_j^2 = 0.9993$，尽管拟合效果较好但 9 个变量显著性 p 值只有两个三颗星和 1 个两颗星。这说明有些变量不适用于建模。同上的，以 AIC 信息统计量为准则，选择最小的 AIC 来进行变量选择。对选择后的模型进行回归得到以下结果:

```
Call:
lm(formula = Y_i ~ FCR_i + fer_i + GDP_i + IP_i + SIF_i)

Residuals:
      Min        1Q    Median        3Q       Max
-0.049405 -0.012548  0.003797  0.009470  0.034693

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.691e-17  5.118e-03   0.000   1.0000
FCR_i       -5.989e-02  2.150e-02  -2.785   0.0146 *
fer_i       -4.531e-02  8.346e-03  -5.429 8.88e-05 ***
GDP_i        1.732e-01  6.286e-02   2.755   0.0155 *
IP_i         9.653e-01  4.739e-02  20.370 8.39e-12 ***
SIF_i       -1.272e-01  1.621e-02  -7.848 1.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02289 on 14 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9995
F-statistic:  7251 on 5 and 14 DF,  p-value: < 2.2e-16
```

结果较令人满意，查看 VIF 值：

FCR	fer	GDP	IP	SIF
16.77	2.53	143.30	81.44	9.53

发现 GDP 和 IP 的方差膨胀因子远大于 10，为了与前文统一，我们通过主成分分析提取整体经济发展因子 econ。同时 FCR 的方差膨胀因子也大于 10，为了全文统一，我们尝试通过主成分分析提取林业规模因子 FD。

```
Standard deviations (1, .., p=3):
[1] 1.7069072 0.2758405 0.1018816

Rotation (n x k) = (3 x 3):
         PC1        PC2        PC3
V1 0.5820533 -0.3040043  0.7541852
V2 0.5790555 -0.4961940 -0.6469051
V3 0.5708841  0.8132483 -0.1127760
```

整体发展因子 econ 的总体总方差比例为$1.7069072^2/3=0.9711774$，令人满意。

```
Standard deviations (1, .., p=2):
[1] 1.3673522 0.3610375

Rotation (n x k) = (2 x 2):
          PC1        PC2
V1 -0.7071068  0.7071068
V2 -0.7071068 -0.7071068
```

林业规模因子 FD 的总体总方差比例为$1.3673522^2/2=0.934826$，同样令人满意。

我们用整体发展因子 econ 和林业规模因子 FD 替代对应变量做回归发现存在变量不显著的问题，我们再次根据 AIC 进行变量选择，得到以下结果：

```
lm(formula = Y_i ~ FCR_i + econ_i)

Residuals:
      Min        1Q    Median        3Q       Max
-0.311214 -0.095607  0.008949  0.047613  0.262572

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.679e-16  3.189e-02   0.000  1.00000
FCR_i       -2.677e-01  8.669e-02  -3.088  0.00668 **
econ_i       7.227e-01  5.079e-02  14.230 7.12e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1426 on 17 degrees of freedom
Multiple R-squared:  0.9818,    Adjusted R-squared:  0.9797
F-statistic: 458.4 on 2 and 17 DF,  p-value: 1.627e-15
```

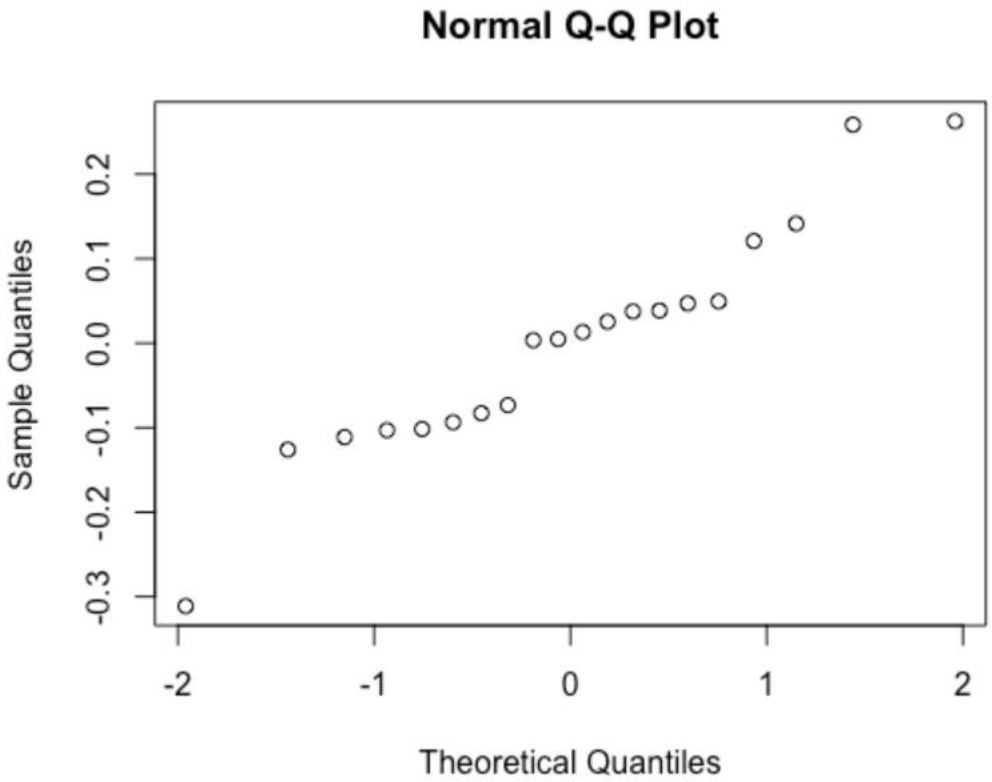
查看的 VIF 值：

FCR	econ
7.02	7.02

再次进行同方差检验:

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.368746, Df = 1, p = 0.066444
```

发现 p>0.05，满足原假设。



我们通过 Q-Q 图检验残差的正态性，图像可以大致近似一条过原点的直线，但存在一定的畸形。最后我们得到：

$$\frac{Y_i-102573948}{97372094}=-1.679\times10^{-16}-0.2677\frac{\mathrm{FCR}_i-490.83}{52.2111}$$
$$+0.7227\left(0.5821\frac{GDP_i-254529.9}{177889.7}+0.5791\frac{IP_i-660095418}{695437675}+0.5709\frac{PS_i-43958.5}{1061.514}\right)$$

3. 基于模型的分析和建议

通过对地区分类，我们得到了各类的回归模型。我们可以看到，不同类别的模型有着较大的差异，这与我们分类的初衷相符。但值得关注的是，在所有类别中，整体经济发展因子都起着重要的作用，并且各个大类中经济发展因子中各变量的系数都较为接近。这也侧面体现了林业与整体经济发展之间的密切关系。本节我们将基于以上模型具体分析并给出对应的建议。

3.1. 分类 1

在分类 1 中，我们最终整合出的整体经济发展因子包含了 GDP、IP（房地产）和 PS（人口数量），不包含 NADS（一定规模以上的工业企业数量）和 SIF（林业资金投入），究其原因，可能是北津上这样的国际都市由于土地价格昂贵、污染环境等因素，大规模企业较少，且林业规模较小，自己投入相对于整体经济水平势微，因此 NADS 和 SIF 与前三者的相关性不高。同样地，我们整合出了林业规模因子，包含 FE（林业企业数量）和 Fer（林业从业人员），用以描述一个地区林产业的规模。在回归中，我们发现森林覆盖率和林业规模因子对林业总产值的影响不够显著，主要的影响因子还是整体经济发展因子。我们认为，由于北津上国际都市的林业规模较小，生态优势不突出，因此林业总产值主要依靠经济总水平一并带动。在政策上，我们建议北津上地区补齐生态短板，控制环境污染，维护现有的林业资源，开发对生态破坏相对较少的林业第三产业，如旅游业等。对于林业企业和林业从业人员，需做到系统培训和精简编排，提高从业人员技术水平，使林业规模因子能够对林业总产值发挥积极作用。

3.2. 分类 2

在分类 2 中，我们最终整合出的整体经济发展因子包含了 PS（人口数量）、GDP、IP（房地产）、SIF（林业资金投入）和 NADS（一定规模以上的工业企业数量），但其中 SIF 的影响并不显著，究其原因，可能因为很多投资造成的林业影响是间接的，没有被记入 SIF 中。 综合来看，青海、内蒙古、西藏、新疆地广人稀，林业生态资源丰富。但是早年间过度采伐严重，生态修复过程艰难。青海、内蒙古、西藏、新疆等地本就有较大比重的经济总值依赖于林业产出，从投资和劳力对产出的影响来看，在未来相当时间内，林业投资仍是林业产出增长的主动力。实现可持续发展，维持森林生态稳定则是亟待解决的问题。

3.3. 分类 3

在分类 3 中，FCR（森林覆盖率）、GDP、IP（房地产）、PS（人口数量）、SIF（林业资金投入）之间高度相关，这之所以其他几类不同应该是地区的经济发展模式决定的。东北地区向来是中国重点林区，林业投资大，后期防沙林修建项目众多，因此林业资金、总资金和森林覆盖率之间存在高度相关性可以理解。辽宁、吉林的天然林较多，早年发生过长期的乱采乱伐，甘肃、宁夏、山西的沙漠化较严重，综合来看，森林资源维持和修复是这片地区的重点，林业资金投入、林业站、林业企业在生态维护中起到主要带动作用，因此对林业总产值影响显著。在政策上，2014 年起，林业部门全面深化改革。2017 年，我国实现了天然林商业性采伐全面禁止。我们认为这片地区应当继续强化天然林的维护，由于修复工程难度较大，因此需要提高林业从业人员技术性，均衡发展林业第一二三产业，减少对生态的破坏。

3.4. 分类 4

在分类 4 中，我们用 econ 作为整体经济发展因子代替 IP（房地产）和 PS（人口）进行回归，除整体经济发展因子外，FE（林业企业数量）也是显著影响因子。综合来看，河北、江苏、安徽、山东、河南、湖北、广东、四川是粮食大省，农田较多，森林较少，因此林业总产值主要靠整体经济来协同带动。由于本身林业企业的发展就较农业落后，其数量增加也对林业产出有明显促进作用。在政策方面，由于该地区本身以粮食产出为主，林业发展可适当与之进行调剂。但林业发展不仅有经济价值，更有生态价值，起到了保持水土、防风固沙的作用，对农业也有直接的影响。因此，该地区在林业管理方面，应当以维持林业生态为主。

3.5. 分类 5

对于分类 5，FCR（森林覆盖率）、GDP、IP（房地产）、PS（人口数量）、SIF（林业资金投入）之间高度相关。FCR 和整体经济发展因子 econ 对林业总产值的影响较为显著。综合来看，黑龙江、浙江、福建、江西、湖南、广西、海南、重庆、贵州、云南、陕西等地的共同点是山地众多，森林覆盖率高，森林资源相对丰富。因此，除了整体经济发展因子外，森林覆盖率对林业总产值的影响也十分显著。在政策方面，对于林业资源相对丰富的省份，早年黑龙江省的林业政策

就是很好的参考。在林业经济布局和规划上，商品林和经济林同时开发，为了达到更好的管理，可以就地设置大型基地，打造本地绿色产品品牌。林权制度改革、林业信贷担保机制、政府补贴也需要跟进，尤其在经济较为发达的地区，通过贯彻市场经济运行原则，可激发企业热情。

三．空间面板数据模型

1．模型构造

通过观察我们所使用的数据集，我们发现，数据集给出的解释变量包括了林业产业发展情况（林业站数目、企业数量、从业人数、一定规模以上企业数量）、人口资本（林业国家投资额、人口总数）、社会经济情况（房地产、人均 GDP）这三个方面，基本覆盖了可能影响林业产值的所有决定因素，因此我们此处假设预测变量 Y 为林业产值。

实验过程中，我们发现，由于各省份间的差异比较明显，所给的数据集各变量的方差都比较大，运用空间面板模型时，数据矩阵的逆矩阵将会变得很难计算，同时，考虑到数据的非负性以及最大规模（不超过e^{15}），我们将对所有数据进行以下变换：

$$x' = \ln(2 + x)$$

经过对数变换后，由于大部分数据都较大（超过 100），因此对数的大小能较为合适地反映各因素的比例关系，而对于林业站数这一较为特殊的数据集（0 出现的次数较多且数据普遍偏小），则能够有效解决 0 无法取对数的问题，且得到的处理后数据均是正数。

接下来，我们开始筛选有效的解释变量，我们将分别在随机效应模型和固定效应模型下检验，筛选时采用的是逐步回归中的 “backward” 法，置信水平临界值设为$\alpha = 0.10$，筛选过程（即调整代码中的回归方程）略，变量对应关系如下：

X_1:$City\ forestry\ station\ number$ 市林业站数

X_2:$Forest\ coveragr\ rate$ 森林覆盖率

X_3:$Forestry\ enterprises$ 林业企业数量

X_4:$Forestry\ practitioners$ 林业从业人数

X_5:$GDP\ per\ capita$ 人均 GDP

X_6:$Investment\ in\ property$ 房地产

X_7:$Population\ size$ 人口总量

X_8:$State\ investment\ in\ forestry$ 林业国家投资额

X_9:$The\ number\ of\ industrial\ enterprises$ 一定规模以上企业的数量

2. 筛选结果

2.1. 随机效应模型：

```
ML panel with spatial lag, random effects, spatial error correlation

Call:
spreml(formula = formula, data = data, index = index, w = listw2mat(listw),
    w2 = listw2mat(listw2), lag = lag, errors = errors, cl = cl)

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.21    7.69    8.32    8.30    8.96   10.42

Error variance parameters:
    Estimate Std. Error t-value  Pr(>|t|)
phi  4.40045    1.19916  3.6696 0.0002429 ***
rho  0.27250    0.12203  2.2331 0.0255425 *

Spatial autoregressive coefficient:
       Estimate Std. Error t-value  Pr(>|t|)
lambda 0.563993   0.061985  9.0988 < 2.2e-16 ***

Coefficients:
             Estimate Std. Error t-value  Pr(>|t|)
(Intercept) -7.915195   0.856089 -9.2458 < 2.2e-16 ***
X2           0.445661   0.083242  5.3538 8.613e-08 ***
X5           0.414032   0.029699 13.9410 < 2.2e-16 ***
X7           0.944571   0.113591  8.3155 < 2.2e-16 ***
X9           0.138073   0.032652  4.2287 2.351e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> model1$sigma2
$one
[1] 0.4914809

$idios
[1] 0.09100734

$id
[1] 0.4004736
```

参数拟合结果为：

$\lambda = 0.564, \rho = 0.273, \phi = 4.400, \beta = [-7.915, 0, 0.446, 0, 0, 0.414, 0, 0.945, 0, 0.138]$

$$\sigma_\epsilon^2 = 0.400, \sigma_\mu^2 = 0.091$$

其中λ,ρ的 t 检验 $p - value$ 均小于规定值，认为这两个参数显著，即林业产值模型中同时存在空间滞后回归项和空间误差回归项。

根据随机效应模型，我们得到的模型认为林业产值与森林覆盖率、人均 GDP、人口总量、一定规模以上的林业企业数量呈正相关。

2.2. 固定效应模型：

```
Call:
spml(formula = Y ~ X2 + X5 + X7 + X9, data = log_dataset, index = NULL,
    listw = Wn0, model = "within", effect = "individual", lag = T,
    spatial.error = "b")

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-1.9336140 -0.1721253 -0.0088464  0.1708013  1.0718524

Spatial error parameter:
    Estimate Std. Error t-value Pr(>|t|)
rho  0.19732    0.17309    1.14   0.2543

Spatial autoregressive coefficient:
       Estimate Std. Error t-value  Pr(>|t|)
lambda 0.576428   0.082727  6.9679 3.218e-12 ***

Coefficients:
   Estimate Std. Error t-value  Pr(>|t|)
X2 0.387330   0.091382  4.2386  2.25e-05 ***
X5 0.431853   0.109698  3.9367  8.26e-05 ***
X7 0.697481   0.210530  3.3130 0.0009231 ***
X9 0.121732   0.033342  3.6511 0.0002612 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> model3$sigma2
[1] 0.08620207
```

参数拟合结果为：

$$\lambda = 0.576, \rho = 0.197, \beta = [0, 0.387, 0, 0, 0.432, 0, 0.697, 0, 0.122]$$

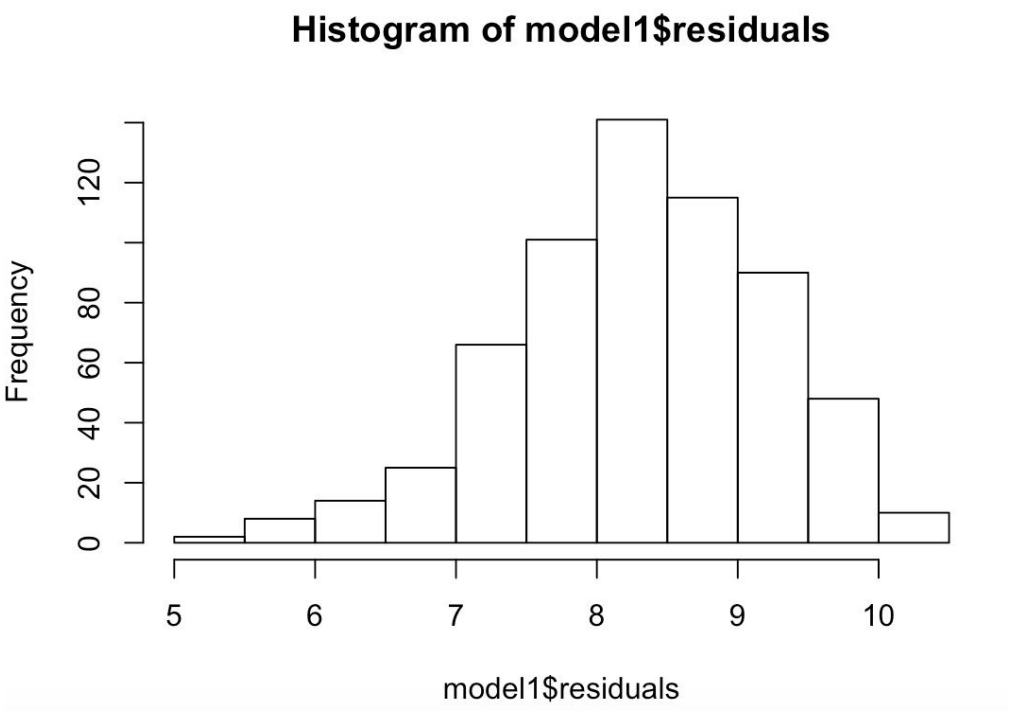
$$\sigma_{\epsilon}^2 = 0.086$$

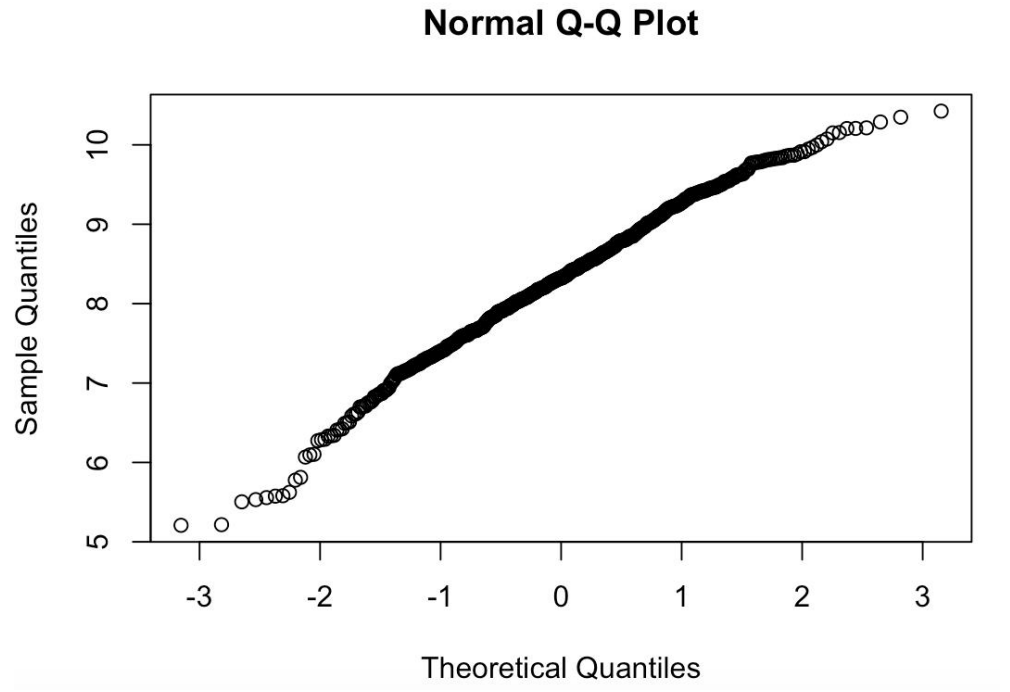
固定效应模型分析结果没有截距项和ϕ这一随机变量，因为我们将μ_i看作一组常数用于区分各个组别（这里即各个省份），这个常数的具体值没有很特殊的意义，只是告诉我们林业产值随时间、地点的变化会发生固定的变化，而这个变化不在我们这次的研究范围内，所以也就不作进一步探讨。

3. 模型结果分析

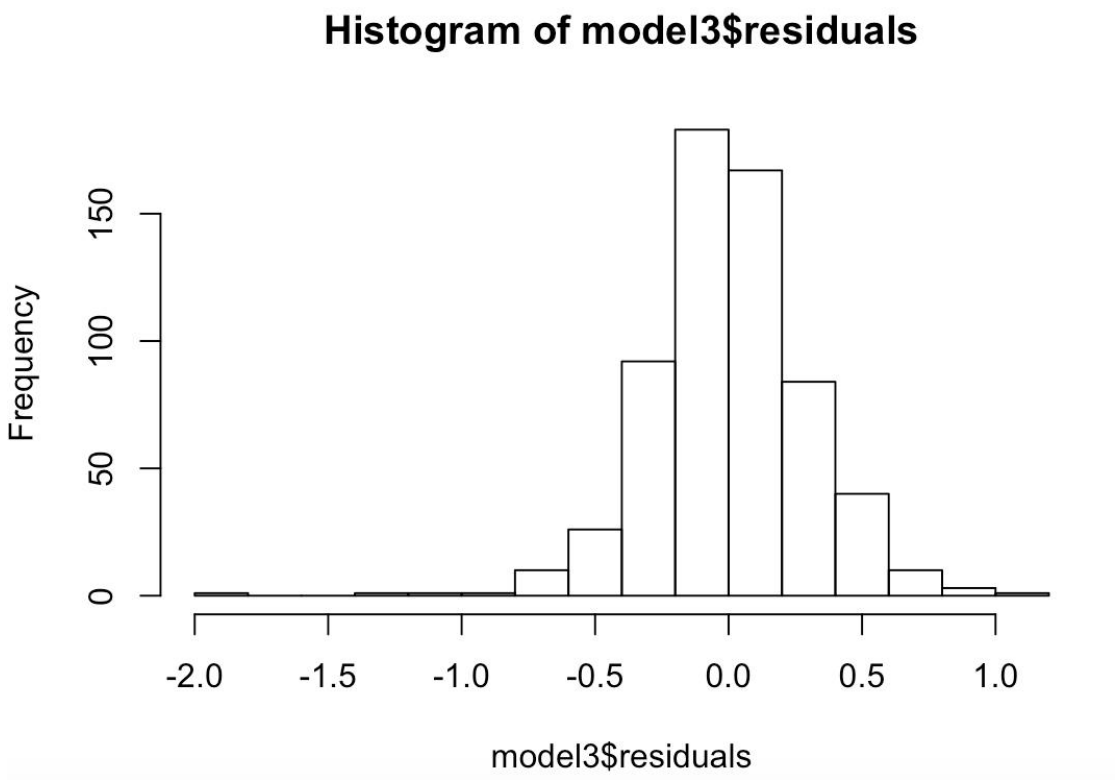
首先，我们绘制了以上两种模型的残差直方图和正态 QQ 图，发现以上两组模型的残差均较为符合正态性假设。（随机效应组的残差均值不为 0 是因为存在未考虑的截距偏差（由个体效应所致），这一部分作为一个常数没有被模型考虑进去，因此我们只需要观察其分布是否大致正态，不需要观察其均值）

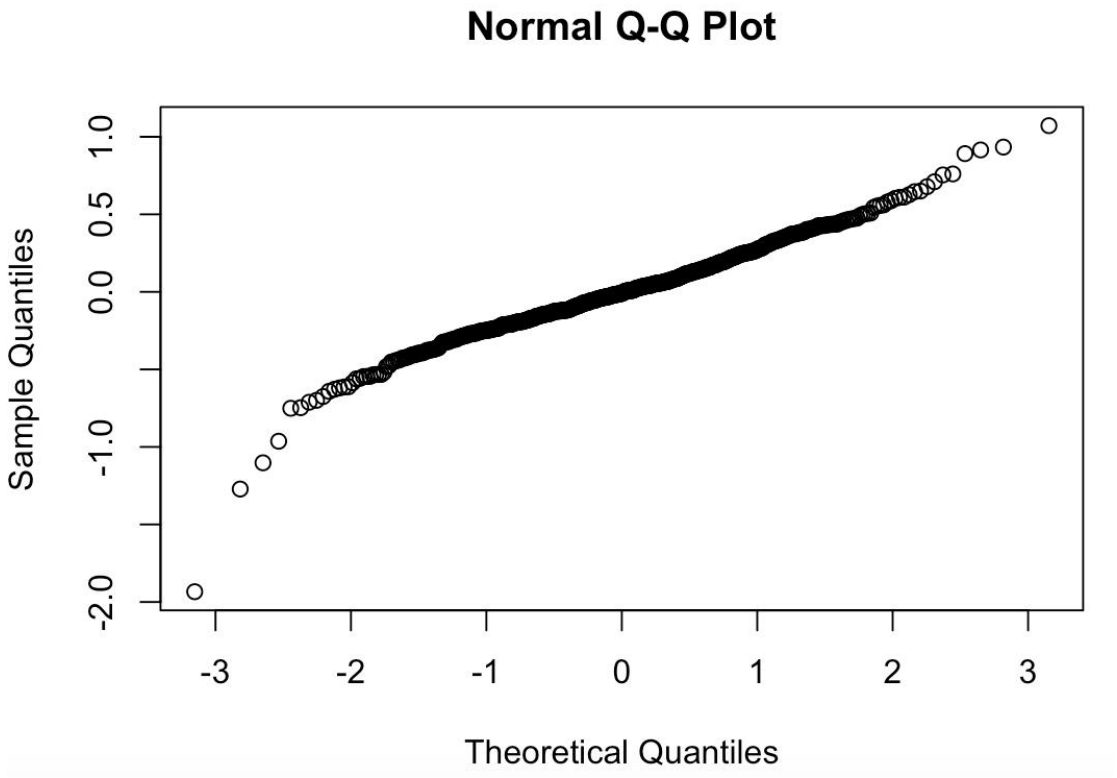
第一组：


Histogram of model1$residuals
Frequency
0
20
40
60
80
120
5
6
7
8
9
10
model1$residuals


Normal Q-Q Plot
Sample Quantiles
5
6
7
8
9
10
-3
-2
-1
0
1
2
3
Theoretical Quantiles

第二组：





其次，通过之前的假设，（阐述一下共线性现象，没准能够解释这里为什么保留下来的是这些解释变量）

同时，在研究的过程中，我们发现数据集中存在一些问题：

在最开始的模拟中，我们得到了林业产值与林业从业人数呈反相关的结论：

```
Call:
spreml(formula = formula, data = data, index = index, w = listw2mat(listw),
    w2 = listw2mat(listw2), lag = lag, errors = errors, cl = cl)

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.06    8.76    9.45    9.40   10.12   11.58

Error variance parameters:
    Estimate Std. Error t-value  Pr(>|t|)
phi  5.98168    1.77584  3.3684 0.0007562 ***
rho -0.93874    0.24886 -3.7722 0.0001618 ***

Spatial autoregressive coefficient:
       Estimate Std. Error t-value  Pr(>|t|)
lambda 0.639586   0.060093  10.643 < 2.2e-16 ***

Coefficients:
             Estimate Std. Error t-value  Pr(>|t|)
(Intercept) -7.180006   0.907717 -7.9100 2.575e-15 ***
X2           0.285487   0.077068  3.7043 0.0002119 ***
X4          -0.062249   0.027044 -2.3018 0.0213483 *
X5           0.368769   0.020506 17.9837 < 2.2e-16 ***
X7           1.015781   0.122821  8.2704 < 2.2e-16 ***
X9           0.040823   0.021810  1.8717 0.0612426 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

以上是随机效应模型的最初拟合结果，可以发现X_4的系数为负且没有在参数筛选的过程中被筛掉。

通过观察X_4（林业从业人数）的原始数据，我们发现了以下问题：

西藏省的 1998~2007 年林业就业人数数据如下：

1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
3538	5747	5747	0	156	156	156	1456	2453	2453

显然一个省的林业从业人员不可能在一年内从 5000 多名降低到 0 名，也不可能从 5000 名减少至 100 名后又迅速增加至 2000 名。

经过对整体表格的观察，我们又发现有不少其他省的从业人员数据也出现过相邻数年从业人员数据完全一致的情况，因此我们初步断定此处数据存在一定的

信息缺失和错误。导致它的原因可能是某些年未进行调查而直接用前一年的数据填补，也可能是数据缺失或文件损坏导致的数据误差。

所以我们对西藏省 2001 至 2004 年的林业从业人员数据进行了修改，将其均填补为 2005 年的数据 1456 名，最终的模型也是对数据进行修改后得出的结果。

4. 模型选择与进一步研讨

本次实验中涉及的解释变量已经基本覆盖了所有与林业产值有关的社会经济因素，剩余的个体效应应当主要为自然因素的不同所导致的差异，而各地区自然条件的差异应当是相对稳定的。所以，从实际问题角度而言，我们应当采用固定效应模型作为空间面板模型拟合的最优模型。

同时，Hausman 检验的结果如下：

```
> test<-sphtest(Y~X2+X5+X7+X9,data=log_dataset,listw=Wn0,spatial.model="sarar", me
thod="GM")
> test

        Hausman test for spatial models

data:  x
chisq = 22.29, df = 5, p-value = 0.0004612
alternative hypothesis: one model is inconsistent
```

根据 Hausman 检验的结果，我们同样应当采用固定效应模型。

所以我们最终得到模型

$$\ln(Y+2) = 0.576(I_T \otimes W_N)\ln(Y+2) + (1_T \otimes I_N)\mu + 0.387\ln(X_2+2) + 0.432\ln(X_5+2) + 0.697\ln(X_7+2) + 0.122\ln(X_9+2) + \epsilon$$

$$\epsilon \sim N(0, 0.086)$$

根据最终选择的模型，我们发现，在综合考虑 9 个解释变量的情况下，为了提高林业产值，我们更应当注意森林覆盖率、人均 GDP、人口总量、一定规模以上的林业企业数量这四项因素。从生态学角度分析，森林覆盖率的提高有助于林业生产规模的扩大；从社会学角度分析，人均 GDP 以及人口总量的增加有助于社会经济带动林业发展，而具有一定规模的林业企业数量的增长也能够在一定程

度上拉动更多民间资本向林业方向流动，从而促进林业产值的增长。

最后我们给出政策上的建议——首先，发展任何产业都需要经济基础，以经济建设为中心仍然需要坚持；其次，发展林业需要一定的自然资源支持，各地需要重视省内绿化程度和森林覆盖率的保持；最后，要发展林业，需要国家和林业企业的支持和带动，要充分发挥大林业企业的优势和作用。

四．结语

我们采用 K-means 聚类法尝试对省份进行分类，发现可将 31 省大致归为 5 类。我们得到了五类各自具有较好解释能力和预测能力的模型，并基于模型分析了各地区的林业发展状况，给出了不同地区相应的政策建议。除此之外，我们也通过分析给出了面向全省的固定效应模型。并针对模型给出了全国性的政策建议和思考。值得一提的是，在所有类别中，整体经济发展因子都起着重要的作用，并且各个大类中经济发展因子中各变量的系数都较为接近。这也侧面体现了林业与整体经济发展之间的密切关系。这与固定效应模型所得到的解释有较好的相容性。

参考文献

[1]熊立春,王凤婷,程宝栋.中国林业产业结构优化及其影响因素分析[J].农业现代化研究,2018,39(03):378-386.

[2]黄韶海,王国峰,邓祥征,陈建成.中国林业生产效率的格局与区域差异分析[J].世界林业研究,2016,29(03):80-85.

[3]徐玮,冯彦,包庆丰.中国林业生产效率测算及区域差异分析——基于Malmquist-DEA 模型的省际面板数据[J].林业经济,2015,37(05):85-88.

[4]王国峰,陈建成,邓祥征.中国林业产业结构与经济增长关系的模型解析[J].林业经济评论,2014,4(01):24-29.

[5]Badi H. Baltagi,Bernard Fingleton,Alain Pirotte. Estimating and Forecasting with a Dynamic Spatial Panel Data Model *[J]. Oxford Bulletin of Economics and Statistics,2014,76(1).

[6]Lung-fei Lee,Jihai Yu. A SPATIAL DYNAMIC PANEL DATA MODEL WITH BOTH TIME AND INDIVIDUAL FIXED EFFECTS[J]. Econometric Theory,2010,26(2).

[7] https://download.csdn.net/download/megantan/10771264