回歸分析方法的實證研究設計

呂　浩

廣東第二師範學院 數學學院 統計系

[**摘　要**] 回歸分析作為數量經濟與統計建模的核心工具，廣泛應用于金融市場行為解釋、宏觀經濟變數關係識別以及教育績效評估等多個領域。本文以金融經濟教育為綜合研究場景，圍繞“變數間因果機制識別與定量解釋”這一核心問題，設計一套基於多元回歸分析方法的實證研究方案。研究從三個維度展開：首先，在金融教育模組中，探討金融素養、風險認知與理財行為之間的關係；其次，在經濟教育方面，分析宏觀經濟課程掌握程度對學生就業預期與經濟判斷能力的影響；最後，在教育效果評估層面，結合學生背景變數、學習投入與教學模式，構建多層次回歸模型以評估教學干預的實際效果。本文將採用 OLS、Logit 回歸、Probit 回歸、分層回歸、固定效應模型等多種回歸方法，並輔以異方差檢驗、多重共線性診斷與穩健性分析，提升模型解釋力與實證有效性。通過構建理論模型、設計問卷調查與實地採樣，研究力求實現教育政策、金融行為與經濟素養提升路徑之間的量化聯結，為金融經濟教育的發展、資源優化配置與干預措施改進提供資料支撐與決策依據。

[**關鍵字**] 回歸分析；多元回歸模型；Logit 回歸、Probit 模型；實證研究設計；因果推斷

1. 理論框架與假設

1.1 一元回歸分析模型

一元線性回歸模型考察的是單一引數與因變數之間的關係，其數學運算式為：

$$y = \beta_0 + \beta_1 x + \epsilon\#$$

這個等式被稱為一元線性回歸理論模型。等式中 β_0，β_1 為未知參數，β_0 是回歸常數，β_1 為回歸係數，ϵ 表示其他隨機因素的影響，是一個隨機變數。這個式子看起來很簡單，但是它充分表達了 x 和 y 之間密切相關，但由於 ϵ 的存在不能唯一確定。x 和 y 之間的關係由兩個部分來描述：一個部分反映了 x 的變化引起的 y 的線性變化，而另一個部分考慮了所有其他隨機因素的影響。

一元回歸分析一共有以下 3 項基本假設：

（1）引數是非隨機變數

（2）$\begin{cases} E(\epsilon_i)=0 \\ D(\epsilon_i)=\sigma^2 \end{cases}$（高斯 - 瑪律可夫條件）

（3）$\epsilon_i \sim N(0,\sigma^2), \epsilon_1,\ldots,\epsilon_n$ 是相互獨立的（為了方便對參數做區間估計和假設檢驗）

利用 n 組觀測值去估計 β_0，β_1 的值，將估計值記作 $\hat{\beta}_0$，$\hat{\beta}_1$，則稱

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x\#$$

為 *y* 關於 *x* 的一元線性經驗回歸方程。採用普通最小二乘估計 OLS，對每一個樣本觀測值 (x_i,y_i)，考慮使觀測值 y_i 與其回歸值 $\hat{y}$ 的離差越小越好，綜合得考慮 *n* 個離差值，定義離差平方和為

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\ \#$$

找到 β_0，β_1 的估計值 $\hat{\beta}_0$，$\hat{\beta}_1$ 令 Q 最小，得到 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_i$ 稱為回歸值或擬合值，$e_i=y_i-\hat{y}_i$ 是殘差

由此得出的回歸方程不能立即使用 , 要對其進行檢驗，以確定它是否真正捕捉到了 *y* 和 *x* 之間的統計規律性。為此，通常需要正態性假設 $\epsilon_i \sim N(0,\sigma^2)$，因為所使用的 *t* 檢驗，*F* 檢驗方法都只適用於正態總體。下面研究對回歸係數的顯著性檢驗——*t* 檢驗。

T 分佈：設 $X \sim N(0,1), Y \sim X^2(n)$, 且 X 與 Y 相互獨立，則稱 $t=\frac{X}{\sqrt{Y/n}}$ 服從自由度為 *n* 的 *t* 分佈。*t* 檢驗是一種假設檢驗，用於評估構建的統計量是否服從 *t* 分佈。它的前提假設是樣本是正態分佈或近似正態分佈。在回歸分析中，*t* 檢驗用於評估回歸係數的顯著性，即確定引數對因變數的影響是否具有統計學意義。

原假設：$H_0:\ \beta_1=0$，備擇假設：$H_1:\ \beta_1\neq 0$；

構造 t 統計量：$\frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/L_{xx}}} \sim t(n-2)$ 其中 $L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$，$\sigma^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$；給定顯著性水準 α，雙側檢驗的臨界值 $t_{\alpha/2}$；當統計量 t 的值被計算並發現 $|t|\geq t_{\alpha/2}$ 時，它就會落入拒絕域，從而拒絕原假設，認為 β_1 顯著不為 0，這意味著因變數 *y* 對引數 *x* 的一元線性回歸是有效的。

下面介紹對回歸方程的顯著性檢驗——*F* 檢驗。設 $X \sim x^2(n), Y \sim x^2(m)$，其中 *x* 與 *y*

是獨立的隨機變數，則稱 $F=\frac{X/n}{Y/m}$ 服從自由度為 n，m 的 F 分佈代表了兩個獨立隨機變數的比值的抽樣分佈，每個比值除以各自的自由度，且 x 與 y 在比值中的位置不可互換。具體來說，它是兩個獨立的卡方分佈隨機變數的比值除以各自的自由度後的抽樣分佈。F 分佈是一種非對稱分佈，廣泛應用於各種統計分析。如在方差分析、回歸方程的顯著性檢驗中發揮著至關重要的作用。

F 檢驗是一種假設檢驗方法，用於判斷的統計量是否符合 F 分佈。在回歸分析中，F 檢驗根據平方和分解方程來直接檢驗回歸效果，從而評估回歸方程的顯著性。

$$SST = SSR + SSE\#$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\left(\hat{y_i}-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\hat{y_i}\right)^2\#$$

總平方和（SST）表示因變數的波動情況；回歸平方和（SSR）是由回歸方程決定的，歸因於引數 X 的波動；誤差平方和（SSE）是表示引數 X 無法解釋的波動，是由 X 以外的不可控制的因素引起的。這樣，SST 就被分解為 SSR 和 SSE 兩部分，SSR 占 SST 的比例越大，說明回歸效果越好。據此構造 F 統計量：

原假設：H_0：$\beta_1=0$，備擇假設：H_1：$\beta_1\neq 0$；

構造 t 統計量：$\frac{SSR/1}{SSE/(n-2)}\sim F(1,n-2)$；

給定顯著性水準 α，雙側檢驗的臨界值 F_α；

計算統計量的值，當統計量的值 $F\geq F_\alpha$，落入拒絕域，拒絕原假設，認為回歸方程顯著。

1.2 多元回歸分析模型

隨機變數 y 與一般變數 $x_1, x_2, \ldots x_p$ 的線性回歸模型為

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\epsilon\#$$

其中，p=1 時就是上面介紹的一元線性回歸。同樣地，假定$E(\epsilon)=0, D(\epsilon)=\sigma^2$，稱 $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\epsilon$ 為理論回歸方程。

有 n 組觀測資料，則線性回歸模型可表示為

$$\begin{cases} y_1=\beta_0+\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_px_{1p}+\epsilon \\ y_2=\beta_0+\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_px_{2p}+\epsilon \\ \cdots\cdots \\ y_n=\beta_0+\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_px_{np}+\epsilon \end{cases}\#$$

和一元線性回歸一樣，多元線性回歸也要進行一些假定，這些假定是一元情況的拓展和延伸，這些假定主要都是為了方便地進行參數估計。

（1）線性關係假設：因變數與引數之間存在線性關係

（2）獨立性假設：觀測值之間相互獨立。即每個觀測值的誤差項與其他觀測值的誤差項 $\epsilon_j(i=j)$ 不相關。引數之間相互獨立。完全的線性關係是不存在的，也就是說，不存在某個引數可以完全表示為其他引數的線性組合的情況，否則，就會出現多重共線性問題。

（3）正態性假設：誤差項 ϵ_j 服從正態分佈

$$\begin{cases} \epsilon_i \sim N(0,\sigma^2) \\ \epsilon_1,\epsilon_2,\ldots\epsilon_i \text{相互獨立} \end{cases}$$

（4）方差齊性假設：誤差項 ϵ_j 的方差在所有觀測值上是恒定的，即具有同方差性，即 $Var(\epsilon_i)=\sigma^2, i=1,2,\ldots n$。

（5）無多重共線性假設：變數之間沒有嚴重的線性相關性。引數之間存在高度線性相關會導致回歸係數估計不穩定，標準誤差增大，從而影響對引數與因變數之間關係的準確評估。

多元線性回歸的未知參數的估計遵循與單一變數情況相同的原則，採用最小二乘法和最大似然估計法等方法，最小二乘法用於確定參數 $\beta_0,\beta_1,\ldots,\beta_p$ 的估計值 $\hat{\beta}_0,\hat{\beta}_1,\ldots\hat{\beta}_p$ 目標是最小化離差平方和。

$$Q\left(\hat{\beta}_0,\hat{\beta}_1,\ldots\hat{\beta}_p\right)=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\ldots-\hat{\beta}_px_{ip}\right)^2\#$$

現在要做的就是求極值，由於 Q 是關於 $\hat{\beta}_0,\hat{\beta}_1,\ldots\hat{\beta}_p$ 的非負二次函數，因而它的最小值總是存在，求 Q 對於每一個未知參數的偏導，再令偏導數為 0，得到各未知參數的估計值。稱

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\ldots\hat{\beta}_px_p\#$$

為回歸經驗方程。最大似然估計法是一種基於概率論和統計學原理的參數估計方法。其基本概念是確定參數值，使觀察給定樣本資料的概率最大化。對於多元回歸分析模型，假設隨機誤差項 ϵ_j 服從正態分佈 $N(0,\sigma^2)$ 時，則可以根據樣本資料聯合概率密度函數構建似然函數。然後通過最大化這個似然函數來獲得參數估計。F 檢驗從整體上看引數是否對隨機變數 y 有明顯影響。

原假設為：$H_0:\ \beta_1=\beta_2=\ldots=\beta_n=0$

備擇假設為 H_1：至少有一個 $\beta_j\neq 0, j=1,2,\cdots,p$

通過計算 F 統計量 $F=\frac{SSR/p}{SSE/(n-p-1)}$，並與給定顯著性水準下的 F 分佈臨界值進行比較，若 $F>F_\alpha(p,n-p-1)$，則拒絕原假設，認為回歸方程是顯著，說明回歸方程是顯著的，

引數全體對因變數 y 產生線性影響。

下面通過 t 檢驗來判斷每個引數對因變數的影響是否顯著。用於檢驗每個引數對因變數的單獨影響是否顯著。對於每個回歸係數 β_j，原假設為 $H_0: \beta_j = 0$，備擇假設為 $H_1: \beta_j \neq 0$。計算 t 統計量 $t_j = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}}$，其中 $\hat{\beta}_j$ 是 β_j 的估計值，$S_{\hat{\beta}_j}$ 是 $\hat{\beta}_j$ 的標準誤差。然後與給定顯著性水準下的 t 分佈臨界值進行比較，若 $|t_j| > t_{\alpha/2}(n-p-1)$，則拒絕原假設，認為引數 x_j 對因變數 y 有顯著影響。與簡單的線性回歸中的樣本決定係數類似，樣本決定係數定義為：

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \#$$

R^2 的取值範圍在 0 到 1 之間，越接近 1 表示模型與資料的擬合程度越高。與 F 檢驗相比，樣本決定係數能更清晰、更直觀地反映模型的擬合效果。但是，它不能作為嚴格的顯著性檢驗。

$$R = \sqrt{R^2} \#$$

稱 R 為 y 關於 $x_1, x_2, \ldots x_p$ 的樣本複相關係數，與簡單相關係數的不同是，複相關係數只取正值，在實際應用中常常用來衡量作為一個整體的 $x_1, x_2, \ldots x_p$ 與線性關係的大小。具體 R^2 或 R 需要達到多少才算通過擬合優度檢驗，這要視具體情況而定。

1.3 主成分分析

主成分分析（PCA）作為一種降維技術，在資料分析中被廣泛應用。PCA 的基本原理是將 n 維特徵投影到 k 維子空間，由此產生的 k 維空間由一組稱為主成分的正交特徵表徵。這些主成分是通過對原始 n 維特徵進行線性變換得到的。PCA 的運行機制基於在原始特徵空間中依次識別相互正交的軸，其選擇標準由資料分佈內在地決定。具體來說，主軸沿著原始資料集中方差最大的方向定向。通過最大化與所有之前選定的軸正交的子空間中的剩餘方差來確定後續軸。通過這種反覆運算過程，可以系統地建立起一整套 n 個正交軸。經驗觀察表明，資料方差的主要部分通常集中在前 k 個主軸上，而與其餘軸相關的方差通常在統計上並不重要。因此，通過保留這 k 個主軸可以有效捕捉資料集的基本方差結構，從而達到降維的目的。這種方法實質上是保留了顯示大量方差的維度，同時剔除了方差貢獻最小的維度，從而提高了資料表示的效率。

資料集 $X = \{x_1, x_2, x_3 \ldots x_n\}$，需要降到 k 維。第一，去平均值（即資料中心化），計算每個特徵的平均值並做中心化處理；第二，計算協方差矩陣 $\frac{1}{n}XX^T$，注：這裡除或不除樣本數量 n 或 n-1，對得到的特徵向量沒有影響；第三，特徵值分解，對協方差矩陣進行特徵分解，得到特徵值和特徵向量；第四，特徵排序，按降冪對特徵值進行排序，選擇最

大的特徵值，然後將相應的特徵向量作為行向量，形成特徵向量矩陣 P；第五，轉換資料，將資料轉換到 k 個特徵向量構建的新空間中，得到變換後的資料集 Y=PX。

原始資料集 $X=\{x_1，\ x_2，\ x_3...x_n\}$，需要降到目標維度 k 維。第一，對資料去均值，從資料集中減去每個特徵的平均值；第二，協方差矩陣求解，計算標準化資料的協方差矩陣；第三，正弦值分解（SVD），對協方差矩陣進行 SVD 分解，得到的特徵值與特徵向量；第四，主成分選取，按特徵值降冪排列，選擇其中最大的個，然後將其對應的個特徵向量分別作為列向量組成特徵向量矩陣；第五，將資料轉換到個特徵向量構建的新空間中。

2. 回歸分析對金融風險影響的實證研究設計

2.1 資料來源與整理

本文所研究的 2007-2023 年商業銀行的年度資料及 2011-2022 年上市公司數字普惠金融發展指數的挖掘，所獲取的 335 家銀行資料主要來源國泰安資料庫、各商業銀行年報、北京大學數字普惠金融指數及國家統計局等。

由於銀行資料涵蓋 07-23 年，而上市公司普惠金融指數資料為 2011-2022 年，兩者時間跨度存在差異。為確保資料的一致性與分析的有效性，本研究選取 2011-2022 年這一共同時間段 20 家銀行的資料進行深入分析。資料清洗結果如表 2-1 所示。

表 2-1 信用風險實驗樣本資料

A	Year	不良貸款率	省級數位化程度	資本充足率	加權風險資產淨額（億元）	成本收入比
平安銀行	2011	0.40	127.06	11.51	7947.02	38.41
平安銀行	2012	0.30	184.78	11.37	8955.93	39.41
平安銀行	2013	0.48	201.53	9.90	11704.10	40.77
平安銀行	2014	0.65	240.95	10.86	13804.30	36.33
平安銀行	2015	0.81	247.996	10.94	16617.50	31.31
平安銀行	2016	0.88	296.168	11.53	20337.20	25.97
平安銀行	2017	0.70	331.918	11.20	22261.10	29.89
平安銀行	2018	0.92	360.608	11.50	23402.40	30.32
平安銀行	2019	0.95	379.535	13.22	27844.10	29.61
平安銀行	2020	0.79	406.531	13.29	31517.60	29.11
平安銀行	2021	0.84	416.359	13.34	35664.60	28.3
平安銀行	2022	0.78	341.158	13.01	39751.82	27.45

2.2 實證模型設定

通過以上現有的資料，我們對這些資料構建大資料對金融風險影響的檢驗回歸模型。$NPL_{i,t}$ 表示 i 銀行在第 t 年的不良貸款率，$RW_{ORi,t}$ 代表 i 銀行在第 t 年的操作風險加權資產，$DL_{i,t}$ 代表 i 銀行在第 t 年的省級數位化程度，$CAR_{i,t}$ 代表 i 銀行在第 t 年的資本充足率，$NRA_{Wi,t}$ 代表 i 銀行在第 t 年的加權風險資產淨額，$CIR_{i,t}$代表 i 銀行在第 t 年的成本收入比，$\ell_{i,t}$、$\lambda_{i,t}$ 為隨機擾動項，具體如下兩個模型所示，分別探究大資料對信用風險與操作風險的影響。

$$NPL_{i,t} = \beta_0 + \beta_1 DL_{i,t} + \beta_2 CAR_{i,t} + \beta_3 NRA_{Wi,t} + \beta_4 CIR_{i,t} + \ell_{i,t}$$

$$RW_{ORi,t} = \beta_0 + \beta_1 DL_{i,t} + \beta_2 CAR_{i,t} + \beta_3 NRA_{Wi,t} + \beta_4 CIR_{i,t} + \lambda_{i,t}$$

2.3 主要變數定義

在闡述完實證模型設定後，為了更清晰地理解模型中各個變數的含義，以下將對主要變數進行定義。在本文當中為探討大資料對金融風險（信用風險與操作風險）的影響被解釋量分別為不良貸款率和操作風險加權資產，如表 2-2，這兩個指標能較直接地反映金融風險的情況，解釋變數為省級數位化程度，能夠直接反映大資料的發展狀況。此外，控制變數為資本充足率、加權風險資產淨額和成本收入比，以排除干擾因素，更好地輔助此模型的完成。

表 2-2 變數設置表

變數類型	變數名稱	變數符號	變數說明
被解釋變數	不良貸款率	NPL	銀行不良貸款占總貸款餘額的比例。
	操作風險加權資產	RWOR	用於衡量操作風險的大小。
解釋變數	省級數位化程度	DL	反映數位化運用技術的廣泛程度。
控制變數	資本充足率	CAR	銀行資本總額與銀行風險加權資產的比率。
	加權風險資產淨額	NRAW	反映銀行資產組合的風險程度。
	成本收入比	CIR	指銀行營業費用與營業收入的比率。

2.4 描述性統計分析

描述性統計分析主要是對資料集中趨勢、離散程度等特徵進行計算，能說明瞭解資料的基本特徵，為後續的分析奠定基礎。以下是對信用風險和操作風險描述性統計分析表的分析如表 2-3、2-4 所示。

。

表 2-3 信用風險描述性統計分析表

變數	樣本量	均值	標準差	最小值	最大值
NPL	264	1.30	0.23	0.53	1.75
DL	264	313.67	6.25	61.76	460.70
CAR	264	12.03	0.067	9.9	13.34
NRAW	264	24059.4	600.88	7947.02	39751.8
CIR	264	29.36	0.27	12.38	42.77

不良貸款率均值為 1.30，標準差為 0.23，可說明銀行不良貸款率平均水準相對穩定，但不同銀行間也存在著一定的差異。省級數位化程度均值為 313.67，標準差為 6.25，表明各地區數位化發展程度有差異，但波動較小。

資本充足率標準差較小（0.067），反映銀行資本充足率整體穩定。加權風險資產淨額標準差較大，體現了銀行資產組合風險程度有波動。成本收入比均值標準差為 0.27，說明銀行營業費用與營業收入的比率相對穩定。

表 2-4 操作風險描述性統計分析表

變數	樣本量	均值	標準差	最小值	最大值
RWOR	100	3649943260	5473925734	313045000	1.70983E+10
DL	100	392.40	68.479	222.12	462.228
CAR	100	13.63	1.88	9.88	18.02
NRAW	100	26470.48	8468.013	8955.93	35664.6
CIR	100	28.54	4.97	12.38	42.77

操作風險加權資產均值標準差較大，反映出資料波動大，代表不同銀行操作風險大小差異顯著。

2.5 Pearson 相關性分析

表 2-5 信用風險相關性分析表

	DL	NPL	CAR	NRAW	CIR
DL	1				
NPL	0.294**	1			
CAR	0.725**	-0.37	1		
NRAW	0.924**	0.109	0.837**	1	
CIR	-0.282**	-0.177**	-0.221**	-0.227**	1

注：***、**、* 分別代表 1%、5%、10% 的顯著性水準。

如上表 2-5 可以看出，數位化程度與不良貸款率呈正相關，說明數位化程度提高，不良貸款率有上升趨勢，對信用風險有增高的趨勢。資本充足率與不良貸款率呈負相關，表明資本充足率越高，不良貸款率越低。加權風險資產淨額與不良貸款率呈正相關，即加權風險資產淨額增加，不良貸款率可能上升。成本收入比與不良貸款率呈負相關，意味著成本收入比越高，不良貸款率越低。

表 2-6 操作風險相關性分析表

	DL	RWOR	CAR	NRAW	CIR
DL	1				
RWOR	-0.922	1			
CAR	0.572	-0.478	1		
NRAW	0.824	-0.659	0.587	1	
CIR	-0.320	0.291	-0.194	-0.242	1

注：***、**、* 分別代表 1%、5%、10% 的顯著性水準。

省級數位化程度與操作風險加權資產呈負相關，說明數位化程度提高，操作風險加權資產可能有降低的趨勢，對操作風險有正向作用。資本充足率與操作風險加權資產呈負相關，表明資本充足率越高，操作風險加權資產越低。加權風險資產淨額呈負相關，即加權風險資產淨額增加，操作風險加權資產可能降低。而成本收入比與操作風險加權資產呈正相關，意味著成本收入比越高，操作風險加權資產越高。

2.6 多重共線性檢驗

表 2-7 信用風險多重共線性檢驗表

Variable	VIF	1/VIF
CONST		
DL	7.614	0.131
CAR	3.572	0.280
NRAW	11.735	0.085
CIR	1.109	0.902

加權風險資產淨額的方差膨脹因數為 11.735 ＞ 10，說明其與其他引數間可能存在嚴重多重共線性，可考慮將這個變數剔除；省級數位化程度的 VIF 為 7.614，雖小於 10 但相對較高，也可能存在一定共線性問題；資本充足率的 VIF 為 3.572，共線性問題相對較弱；成本收入比（CIR）的 VIF 為 1.109，接近 1，幾乎不存在共線性。

表 2-8 操作風險多重共線性檢驗表

Variable	VIF	1/VIF
CONST		
DL	3.378	0.296
CAR	1.586	0.631
NRAW	3.322	0.301
CIR	1.117	0.895

各變數 VIF 均小於 10，表明這些變數間雖存在一定共線性，但程度相對較輕，對模型影響較小。

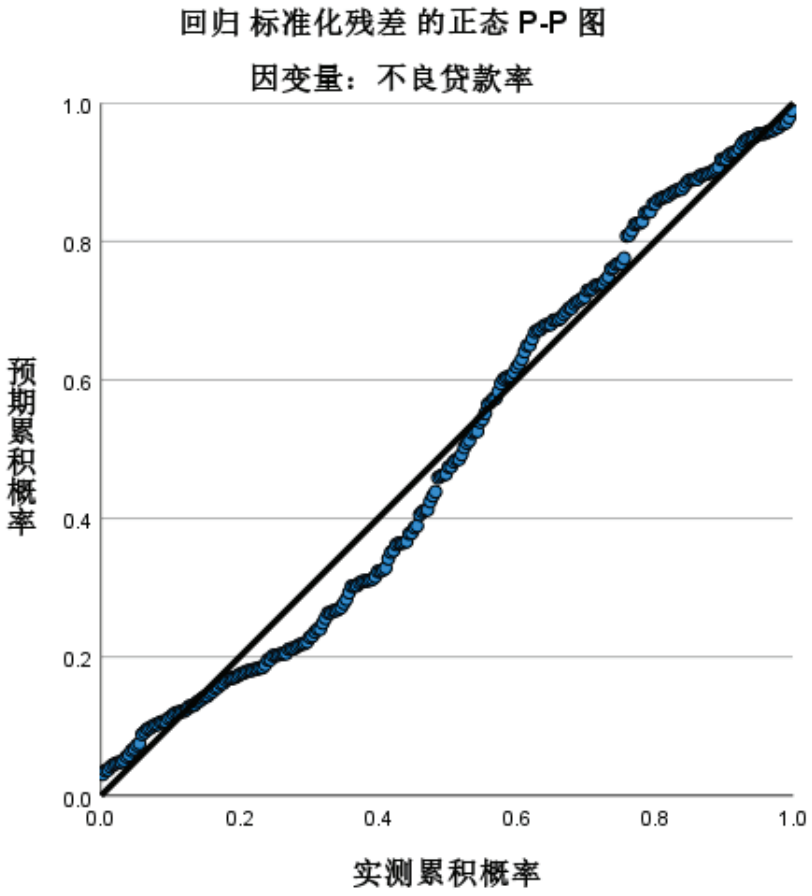


圖 2-1 信用風險標準化殘差的正態 P-P 圖

根據上圖 2-1 正態 P-P 圖所示，橫坐標為實測累積概率，縱坐標為預期累積概率。圖中散點雖然存在一定波動，但整體較為緊密地圍繞在對角線附近。這表明信用風險模型的殘差近似服從正態分佈，從殘差分佈角度說明模型對信用風險資料有較好的擬合效果 。即模型能夠在一定程度上合理地解釋不良貸款率相關資料，模型設定相對合理，未出現明顯因模型設定問題導致殘差分佈異常的情況。

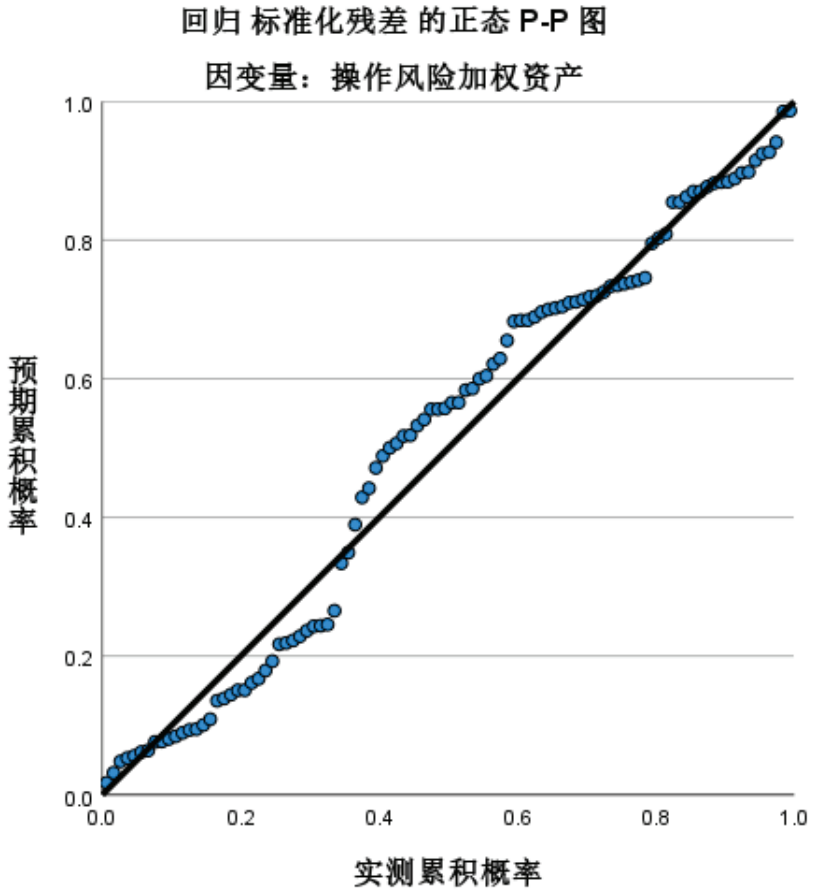


圖 2-2 操作風險標準化殘差的正態 P-P 圖

根據上圖 2-2 所示，該圖中散點同樣是圍繞對角線分佈，但相比信用風險圖，散點波動稍大一些。不過總體上，大部分散點還是分佈在對角線附近，說明操作風險模型的殘差也近似服從正態分佈，模型對操作風險加權資產資料有一定的擬合能力 。但相較于信用風險模型，操作風險模型可能存在一些細微的不完美之處，不過並不影響整體上認為模型能對操作風險資料進行有效解釋。

綜上所述，大資料通過影響普惠金融發展、銀行資產風險及資本管理等方面，對信用風險產生顯著作用。綜合模型整體來看，大資料相關變數（數位化程度）與不良貸款率的呈正相關關係，與“隨著互聯網大資料發展迅猛，數位化的運用加強，可能降低了商業銀行的不良貸款率”這一假設不相符。這可能暗示著，雖然大資料在理論上具備提升風險評估精准度、降低信用風險的潛力，但在實際情況中，由於大資料應用初期銀行風險管控體系尚未完全適應，或者大資料帶來的金融業務創新使信用風險的結構發生變化等原因，導致了數位化程度提高時不良貸款率反而上升。模型良好的擬合效果表明，這種關係不是由於模型設定錯誤或殘差異常導致的虛假關係，而是在現有資料和模型框架下真實存在的一

種關聯，這就需要進一步深入研究大資料在信用風險領域應用的具體機制和影響路徑，以更好地利用大資料來管理信用風險。

綜上所述，實際情況與“隨著商業銀行數位化程度的提升，其資料處理與分析能力增強，可能降低操作風險加權資產。”這一假設相符，省級數位化程度與操作風險加權資產的負相關關係具有重要意義。這表明大資料相關因素在操作風險領域可能發揮著積極的作用，例如大資料技術可能通過更精准的風險監測、更高效的業務流程管理等方式，降低了操作風險加權資產。然而，模型擬合的不完美也意味著，大資料對操作風險的影響可能還存在其他未被模型完全揭示的方面，或許存在一些調節變數或者複雜的交互作用尚未被納入模型。以更全面、準確地探究大資料對操作風險的影響機制，從而為銀行等金融機構利用大資料優化操作風險管理策略提供更可靠的依據。

從以上分析結果可以看出模型對操作風險加權資產的解釋程度較高。說明大資料在操作風險研究中具有較強的解釋力，能夠有效捕捉影響操作風險的關鍵因素。整體回歸模型顯著，即這些引數聯合起來對操作風險加權資產有顯著影響。綜上，大資料對操作風險有著多維度的影響。一方面，大資料帶來的數位化程度提升對降低操作風險加權資產有顯著的積極作用；另一方面，大資料能夠説明銀行更清晰地認識業務收入變化與操作風險之間的關係，輔助銀行制定更有效的操作風險管控策略。

2.7 研究結論總結

大資料支援下的普惠金融發展通過省級普惠金融水準綜合指數體現，與不良貸款率呈顯著負相關，有助於降低信用風險。但互聯網支付規模與不良貸款率正相關，可能因大資料推動下互聯網支付發展使信用風險暴露。且引數間存在共線性問題，影響模型準確性。操作風險是指在組織出現流程無序、人為操作失誤、內部控制失靈等現象後，遭受損失的風險，（大資料背景下銀行業風險）數位化程度與操作風險加權資產呈顯著正相關，導致原因可能是數位化轉型初期系統不穩定、資料安全風險、人員適應問題以及業務拓展創新帶來的風險。模型擬合效果和穩定性好，但與降低操作風險的假設相悖。

在大資料環境下，資料量龐大且來源多樣，資料品質把控難度極大，資料缺失、異常值以及變數共線性等資料品質問題，會嚴重影響模型準確性和穩定性。與此同時，難以全面評估大資料對不同金融風險之間傳導機制的影響，比如操作風險的變化可能通過業務流程影響信用風險，可當前研究較少涉及此類風險間的聯動關係 。

資料獲取不全，樣本不具備代表性，例如研究省級金融風險時可能存在省份資料缺失或者資料不可靠的現象，影響到資料的泛化性。資料更新滯後，更新頻率低，沒有有效反映金融市場的動態變化。由於採集資料量大，對採集的資料進行品質檢查需要一定的優先

原則，可以將對利率風險影響程度較低的指標排除。選擇主要風險因素時可參考相關風險報告尋找有價值的資訊[5]。研究方法也具有局限性，採用傳統回歸分析方法，可能無法充分挖掘大資料的複雜特徵和潛在關係。回歸分析對資料的線性假設較為嚴格，而金融風險資料往往具有非線性、高維度等特點，傳統方法難以有效處理。

金融的核心在風控，風控的核心在資料。 資料時代，資料是商業銀行的重要資產，商業銀行必須重視對大資料技術的應用，它可能是商業銀行未來賴以生存和發展的核心競爭力。

大資料時代，金融機構的資料十分龐雜，維度十分複雜。若在金融風險管理過程中，增強金融資料處理、分析自動化能力是確保資料安全和品質的前提條件，採用先進的資料分析方法和技術以及模型改進等技術必不可少，其中主成分分析和因數分析等模型降維分析方法對於實現處理大資料量、高維資料分析有極大説明。在模型改進假設方面，一方面要不斷發現並完善金融市場運行機制，滿足業務實際需求，對一些與實際情況不符的，不符合現實的簡單模型假設進行改進，這樣可以加強金融風險模型的解釋度和預測精度。例如，在金融風險操作風險模型構建過程中，由於模型本身的操作風險資料往往複雜紊亂、資料量巨大，無法滿足實際需求，採用主成分分析方法可在龐大資料集的資訊篩選上進行快速精准的選擇，能準確剔除對模型運行造成干擾因素，有效簡化模型複雜程度，加快模型分析速度，確保操作風險管理模型評估風險能夠快而准。

充分利用聯邦學習技術，高校配合金融機構在保護資料隱私的情況下開展聯合培訓。金融行業大資料技術的大規模應用，會對行業人才需求提出新的要求，應強化金融科技人才的培養，提升金融機構從業人員大資料分析運用及風險防控等方面的能力。可以由銀行與院校結成聯盟，借助聯邦學習技術進行聯合培訓，在培訓的課程內容上設立初級資料處理、資料分析工具的操作、高級演算法實操等系列內容，使學員充分掌握資料清理、資料採擷、機器學習演算法等方面的內容。在風險防控的課程設計中，圍繞著金融風險識別、風險計量、風險估計、風險控制等方面的內容，對多種不同類型風險的特徵、方式等方面進行講解。通過對課程體系內容的研習，培養出既有豐富金融知識，又能在激烈的市場競爭環境下讓金融機構達到可持續發展目標所需的有生力量。

基於區塊鏈的大資料金融共用平臺可實現對金融資料的安全隱私保護，借助區塊鏈去中心、不可更改、可追溯等特點，進而加強金融機構的資料共用與風險管理戰略共用。近年來大資料背景下金融風險呈現多樣化、複雜化發展趨勢，金融業之間相互聯繫，任一金融機構單方面難以獨力解決金融風險問題，因此，金融機構之間加強資料、風險管理策略共用，是金融風險協同治理的重要目標。

此外，監管部門要構建適合大資料時代適用的監管體制，並完善收集、存儲和使用的

大資料準則。在金融領域系統性風險保護方面，強調的是連續追蹤金融機構的大資料使用事例以識別可能的危險苗頭。為推動資訊流動，在區塊鏈技術條件下構建一個適合金融大資料的共用結構，以及加解密技術加上嚴格的控制手段保障資料安全，保護個人隱私，促進金融體系的穩定與發展。

大資料金融帶來的問題：大資料技術應用導致的演算法風險和模型偏差風險等新型風險在金融領域引起，這些風險會對金融機構的決策效率和品質產生影響，乃至對金融行業的整個風險平穩造成影響，給金融的平穩以及金融風險的決策帶來挑戰。

演算法在金融機構的建設過程中，設計的思路有所缺失，由此形成了演算法設計缺陷。有些演算法模型在規劃中對金融市場存在很大的判斷誤差，市場的真實運行機制過於簡單，缺乏對金融市場運行機理的認識。當市場出現一定程度的變化時，演算法並不能迅速及時準確判斷市場的變化。造成在投資行為上錯誤的行為決策，致使金融機構蒙受巨大的經濟損失。

其次，模型偏差的不良影響也會誘發金融機構對風險的評估失真，做出錯誤的投資決策。金融機構構建投資組合模型時，如果對資產的風險收益特徵進行偏誤的估計，會引起投資組合的配置失衡。還可能被金融機構利用進行監管套利，部分金融機構為使監管指標滿足監管要求，構建風險管理模型，故意篡改模型參數，使風險評估的結果滿足監管要求，但實際的風險情況卻沒有有效控制。

綜合來看，為有效應對大資料在金融風險管理應用過程中產生的演算法風險和模型偏差等問題，需多方協同參與解決。要做好資料治理，做好資料治理機制的構建，確保資料的完整性和品質，通過資料的清洗、去噪、標準化來縮小資料偏差對模型產生的影響。金融機構在開發演算法構建模型過程時，要構建跨領域團隊，包括金融專家、資料科學家、演算法工程師，引入倫理關切，從演算法設計的角度來保障演算法的公平性、公正性，利用多元化的資料集對模型進行訓練，防止演算法對部分人群產生的偏向性。全面考慮到金融市場環境的複雜性，盡可能從根源上降低設計漏洞。對於演算法黑箱問題，研究人員要研究可解釋的 AI，從而能讓演算法的決策變得更加透明，在最大程度上增加金融機構對演算法的信任度。監管層要做好演算法、模型的統一標準制定和定期的審查金融機構所使用演算法模型等工作，打擊監管套利現象。

除此之外，建立健全管理與合規，積極順應監管機構制定的大資料時代監管政策，規範對資料的收集、存儲、使用的合規要求，保證大資料應用的合規合法。最後，建立常態化的監測機制，定期對模型的性能及模型的偏差情況進行定期監測與評估，通過模型的優化和完善以降低模型的偏差，削弱模型偏差對決策效果的影響。通過上述措施，有利於金

融機構對大資料應用所帶來的新風險採取正確措施進行控制，能夠確保將大資料在金融風險控制中的作用更好發揮出來。

3. 回歸分析對經濟發展的實證研究設計

3.1 描述性統計

資料來自國家統計局（https://www.stats.gov.cn/），具體包含：2013-2022 年中國經濟資料集，該資料集共 10 個樣本點，每個樣本均包含 8 個指標，包括 GDP（億元）、從業人員數量（萬人）、進出口額（億美元）、財政支出（億元）、消費總額（億元）、企業數量（個）、產業結構占比（%）以及研發投入（萬元）等。本文以中國省 31 個省級市作為研究單位，以 2013-2022 年這十年間的相關經濟資料為研究物件，所使用的資料來源於《國家統計局省級年鑒》。國家統計局與年鑒資料具有官方權威性，覆蓋指標全面，確保研究結論的可靠性，為消除資料的量綱影響，使分析結果更具有經濟意義，本文對物質資本存量、勞動力、規模以上工業企業數量的資料進行了對數處理，見图 3-1。

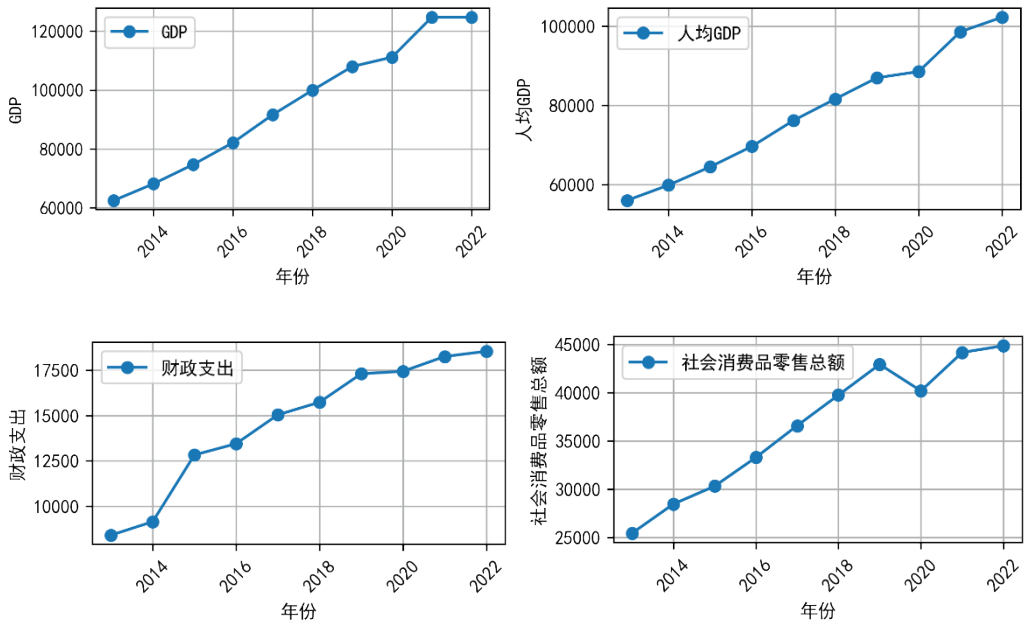


圖 3-1 2013-2022 年廣東省四個經濟指標趨勢圖

在區域經濟研究領域，GDP 作為衡量地區經濟增長水準的核心指標，具有無可替代的重要地位。一般而言，GDP 數值越高，直觀反映出該地區經濟發展水準越高，經濟活力和發展成效越顯著。緊密圍繞本研究主題，本文將中國各省級市的 GDP 設定為被解釋變數。從業人員數量（萬人）：反映勞動力資源的投入情況。進出口額（億美元）：衡量經濟的外向型程度。財政支出（億元）：反映政府對經濟的支持力度。消費總額（億元）：衡量內需

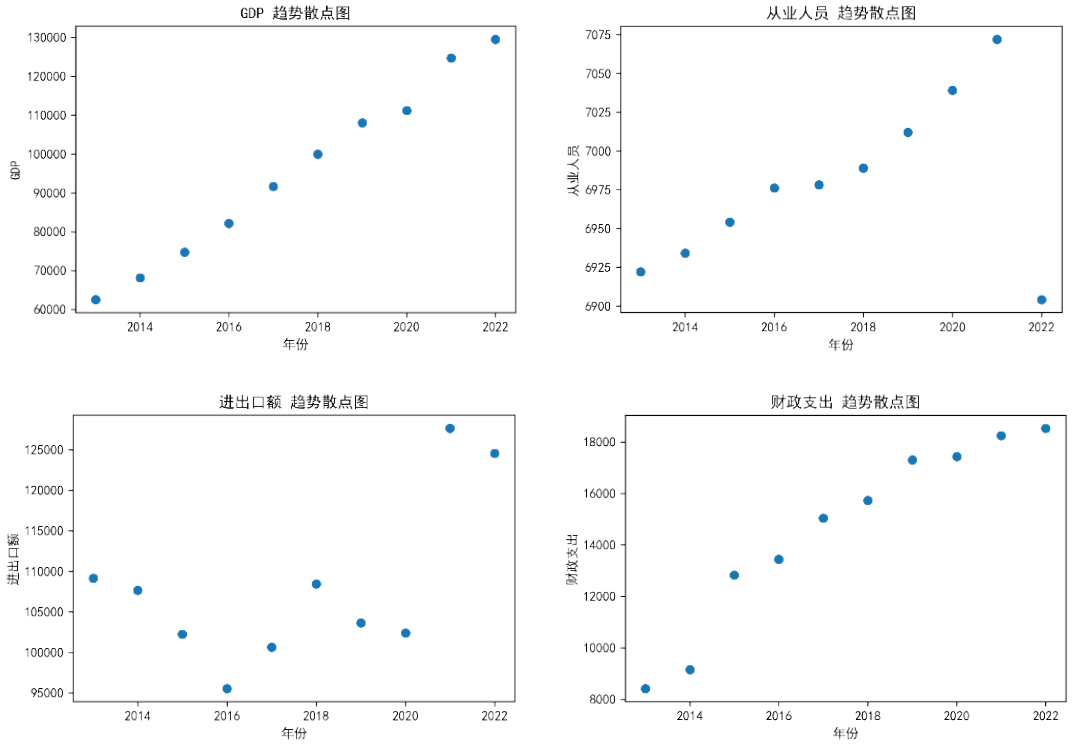
對經濟的拉動作用。企業數量(個):反映經濟活力和市場主體的活躍度。產業結構占比(%):反映產業結構的優化程度。研發投入(萬元):評估技術創新對經濟發展的貢獻。

基於新古典經濟理論所構建的經濟增長分析框架，以及空間計量經濟學在處理區域經濟空間相關性方面的獨特視角，參考國內外學者在此領域的豐碩研究成果，結合本研究針對中國各省級市經濟增長的具體分析需求，初步篩選出一系列經濟指標作為被視作影響中國各省級市經濟增長的主要因素，作為解釋變數引入後續研究：參考前人的做法，選取產業狀況、資本因素、貿易因素、創新驅動四大類八個評價指標，主要包括以下兩級的內容（見:表 3-1）。

表 3-1 中國各區域經濟發展的綜合分析的指標體系

指標性質	指標名稱	指標說明	指標單位
經濟水準	GDP	區域國內生產總值	億元
產業結構	工業發展	規模以上工業企業數	個
	三產占比	各省市第三產業增加值 / 各省市 GDP	%
資本因素	人口資本	各省市從業人員數	萬人
	物質資本	社會消費品零售額	億元
貿易因素	進出口額	各省市進出口額	億美元
	財政支出	各省市的財政支出	億元
創新驅動	R&D 經費	區域全社會研究與試驗發展經費	萬元

對 2013 至 2022 年的資料進行預處理，以廣東省為例（見图 3-2）:



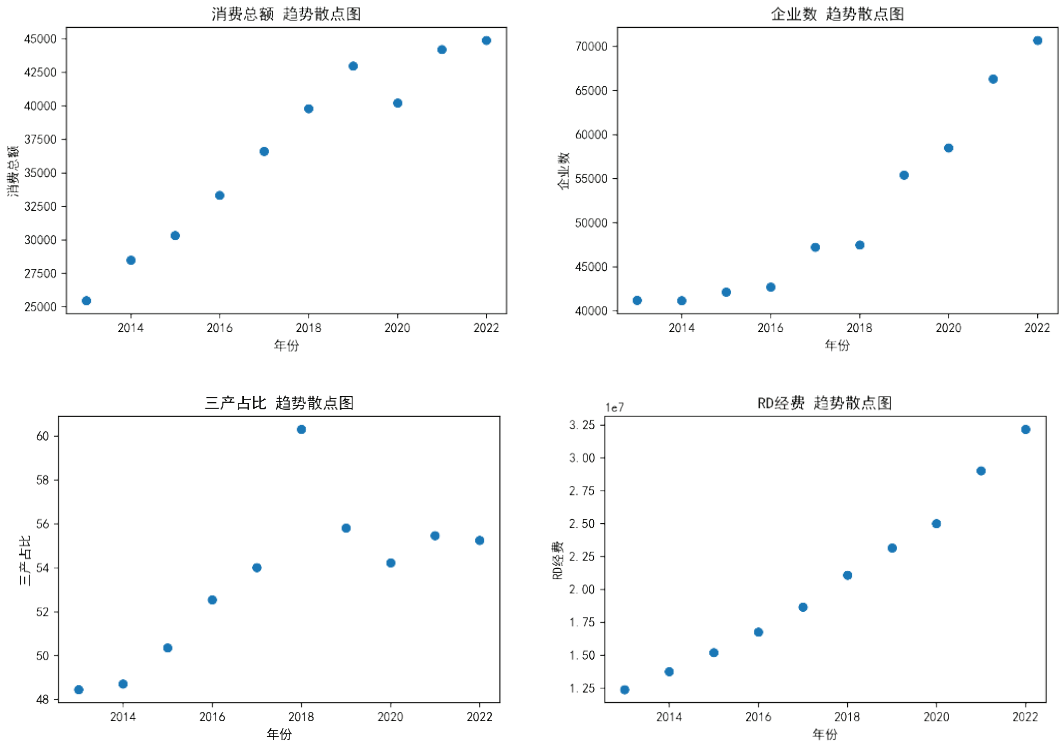


圖 3-2 廣東省各經濟指標趨勢圖

在开始深入分析之前，对数据进行描述性统计是十分必要的。这一步能够帮助我们初步了解数据的基本特征，包括数据的集中趋势、离散程度以及分布形态等，从而为进一步的数据分析奠定基础。

對於數值型資料，如 GDP、從業人員數量、進出口額、財政支出、消費總額、企業數量、產業結構占比以及研發投入等，我們計算了每個變數的均值、中位數、標準差、最小值和最大值（見表 3-2）。這些統計量能夠從不同角度反映資料的分佈情況。例如，平均值和中位數可以揭示資料的中心傾向，而標準差則可以衡量資料的分散程度。最小值和最大值則可以讓人瞭解資料的範圍。

以 GDP 為例，其均值為 28276.77，中位數為 22511.75，標準差為 24121.93，最小值為 828.2，最大值為 129513.6。這表明各地區的經濟發展水準存在一定的差異，標準差 24121.93 反映了資料的離散程度，而最小值 828.2 和最大值 129513.6 則展示了資料的極端值範圍。

表 3-2 2013-2022 年中國經濟指標描述性統計資料

指標	count	mean	std	min	25%	50%	75%	max
GDP	310	28276.77	24121.93	828.20	12110.58	22511.75	37309.7	129513.6
從業人員	310	2342.12	1636.88	180	1228.75	2006.5	3220.5	7072
進出口額	310	14944.64	23841.59	31.05	1891.59	5113.96	12409.68	127659
財政支出	310	5712.9	3184	922.48	3697.49	5090.7	7244.98	18533.08
消費總額	310	11362.8	9560.32	322.2	4217	8800.25	15765.62	44882.9
企業數	310	12746.24	14340.71	76	3579.75	6623.5	16534.75	70702
三產占比	310	51.43	8.64	34.66	46.56	50.2	53.73	89.6
RD 經費	310	4181208	5574814	2602	775087.5	2415111	4780620	32177548

運用皮爾遜相關係數，初步探究各經濟因素間的線性關係強度，為後續回歸分析奠定基礎。根據熱力圖 (見图 3-3) 中的數值和顏色深淺，可以識別出以下經濟因素之間存在顯著的相關性：

（1）從業人員與 GDP：相關係數為 0.86，呈較強正相關。意味著從業人員數量增加，通常能為生產、服務等經濟活動注入更多人力，推動產出增長，促進 GDP 提升。

（2)進出口額與 GDP:相關係數為 0.82 ,正相關明顯 。進出口業務拓展能帶動貿易 物流、製造等多領域發展，增加經濟總量，拉動 GDP 上升。

（3）財政支出與 GDP：相關係數為 0.94，高度正相關。財政支出投向基礎設施建設、公共服務、產業扶持等領域，可刺激經濟活動，帶動 GDP 增長。

（4）消費總額與 GDP：相關係數為 0.99，接近完全正相關。消費是拉動經濟的重要馬車，消費總額增加反映市場需求旺盛，促進企業生產和服務供給，推動 GDP 增長。

（5）RD 經費與 GDP：相關係數為 0.96，高度正相關。R&D 經費投入增加，有助於推動技術創新、產業升級，提高生產效率和產品附加值，進而促進 GDP 增長。

（6）企業數與 GDP：相關係數為 0.92，正相關性強。企業數量增多，意味著市場主體增加，生產、創新等經濟活動更活躍，創造更多價值，促進 GDP 提升。

（7）年份與 GDP：相關係數為 0.26，相關性低。說明年份與 GDP 之間不存在緊密的直接關聯趨勢，不能簡單判定對 GDP 有促進或抑制作用。

（8）三產占比與 GDP：相關係數為 0.09，相關性微弱。表明三產占比變化對 GDP 的直接影響不顯著，難以直接界定其對 GDP 的作用方向。

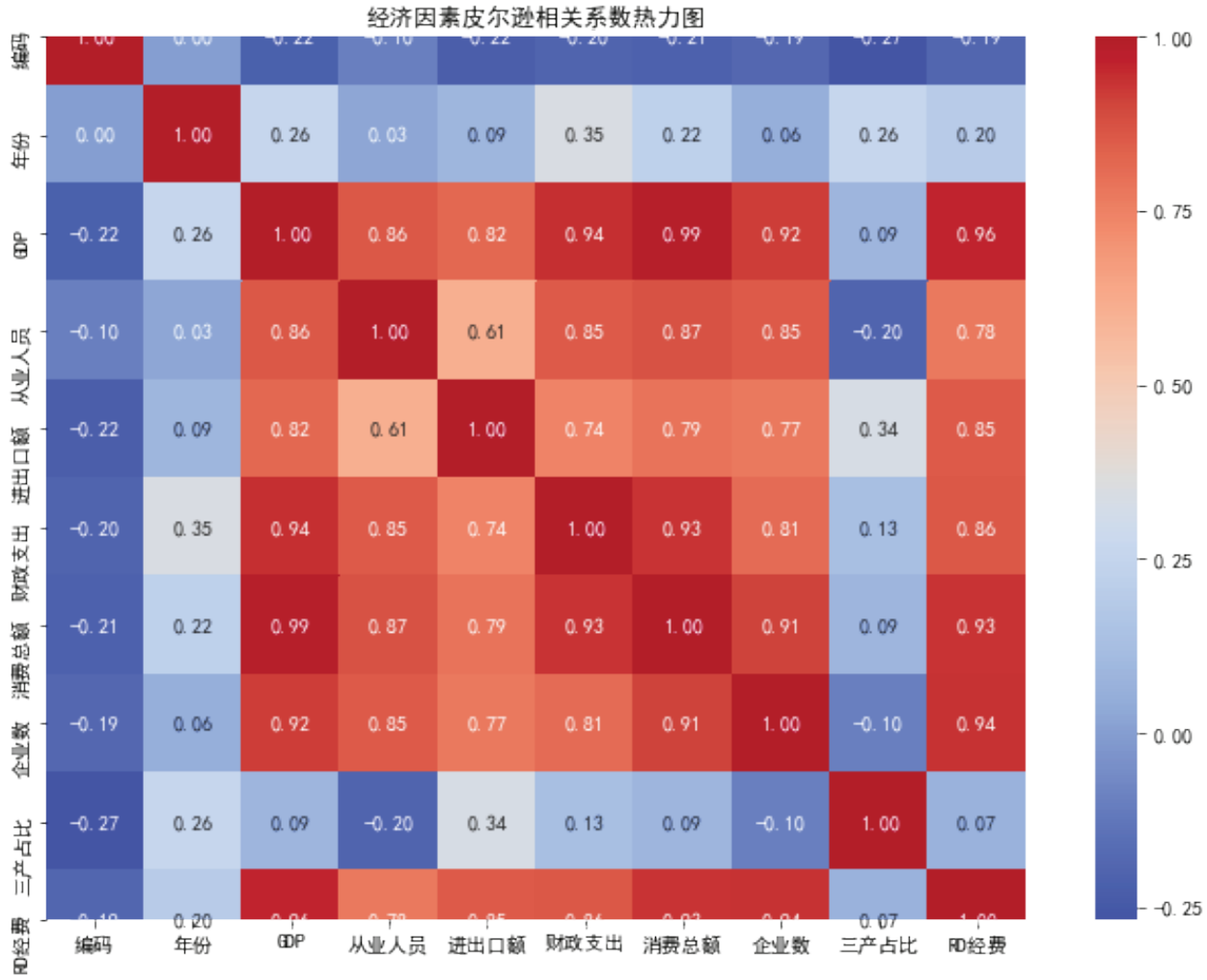


圖 3-3 經濟因素皮爾遜相關係數熱力圖

3.2 中國經濟發展視覺化分析

通過調用 Python 的 matplotlib 標準庫，我們將 2013 年至 2022 年的時間軸設為橫坐標，選取近十年的資料作為縱坐標的衡量標準，繪製出多幅清晰直觀的折線圖。這些圖表不僅精准呈現了資料的動態變化趨勢，還借助折線的波動直觀展現了資料背後的潛在規律與故事，為深入分析提供了強有力的視覺支援。具體而言，在分析影響中國經濟因素時，我們利用折線圖展示了我國近十年間各經濟指標的演變情況（見圖 3-4）。

從圖中指標趨勢可見，2013—2022 年我國經濟呈現總量增長與結構優化並行的發展態勢：GDP 穩步攀升，進出口額、財政支出、消費總額等規模持續擴大，彰顯經濟活躍度與內外需求動力；企業數量增長、從業人員規模穩定，體現市場活力與就業支撐；RD 經費投入遞增、三產占比波動優化，則反映創新驅動與產業結構向高端化轉型的高品質發展特徵。

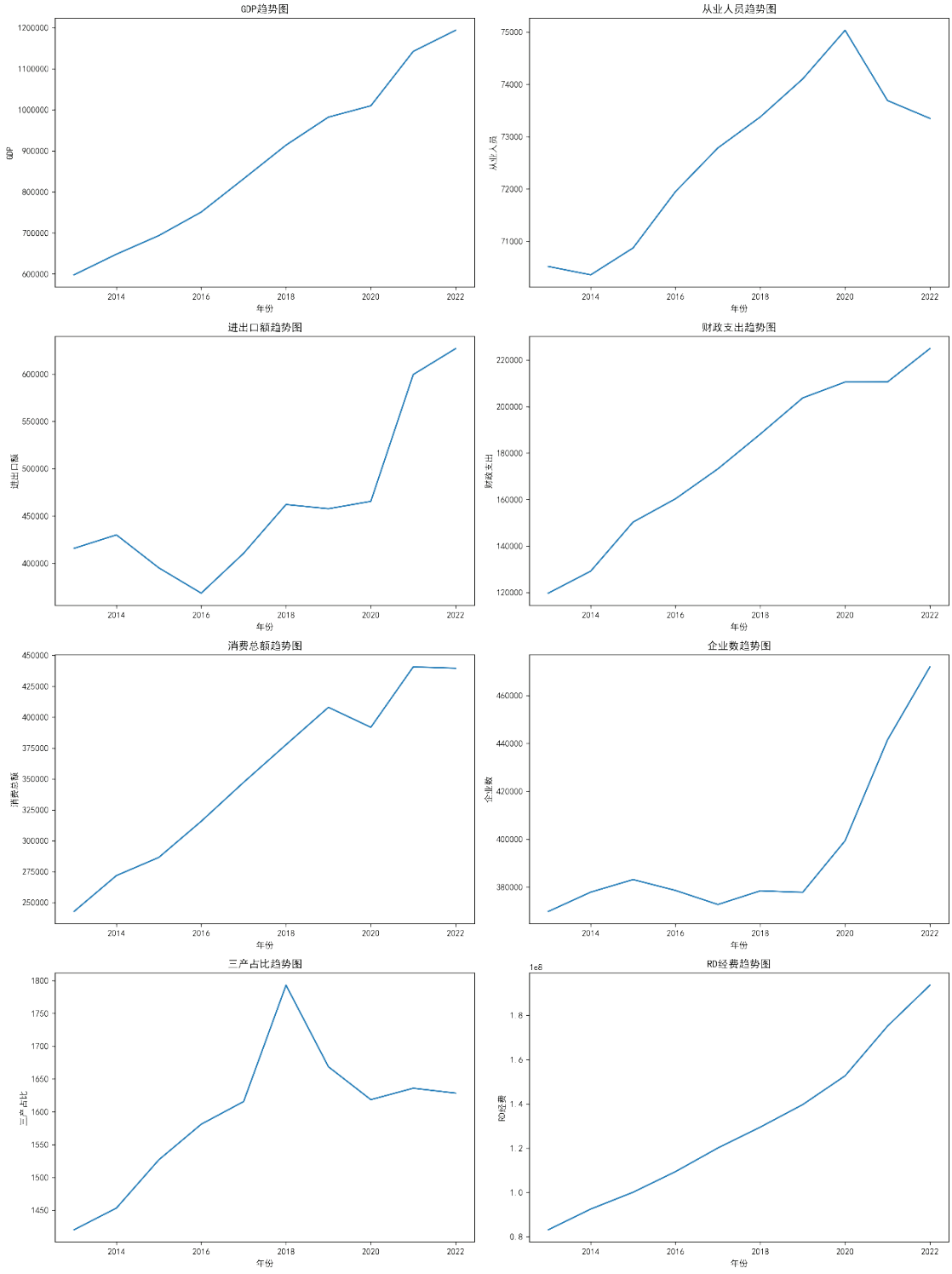


圖 3-4 中國近十年間各經濟指標演變折線圖

根據地區所屬區域（東部、中部、西部、東北）對資料分組，計算各區域每年各指標總和，再分別繪製折線圖（見圖 3-5）。

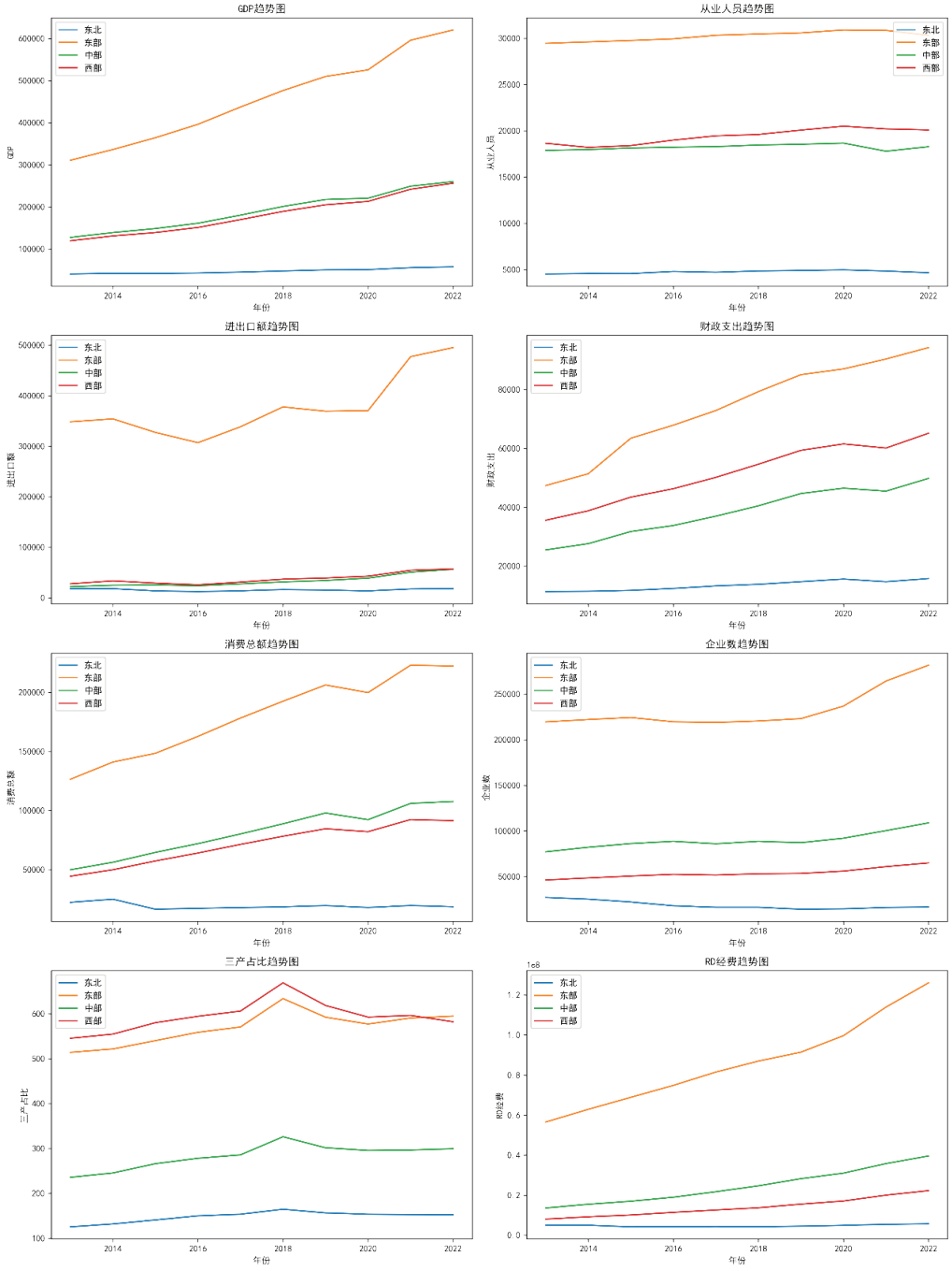


圖 3-5 四大經濟區域折線圖

從圖看，四大經濟區域發展差異顯著。東部地區在 GDP、進出口額、企業數、RD 經費等指標上優勢突出，經濟總量大、對外貿易活躍，創新投入高，引領經濟發展；中部與西部地區在財政支出、消費總額等指標上呈穩步增長趨勢，發展動力較強，但整體經濟規模、外貿活躍度和創新投入仍與東部有差距；東北地區多數指標增長幅度相對較緩，尤其在 GDP、企業數、RD 經費等方面表現較弱，經濟發展動力不足問題較為突出。整體呈現

東部領跑，中部西部追趕，東北發展動力待提升的區域經濟格局。

四大經濟區域呈現出的發展差異，對國家資源配置的均衡狀況、產業協同所有的效率以及共同富裕的推進過程均產生影響，東部地區在創新以及產業等方面處於領先地位，而中西部和東北地區則各自存在著一定的短板，對影響 GDP 的各項指標展開剖析後發現，RD 經費投入顯得更為關鍵。這是因為它可推動科技創新的發展，促使產業實現升級，提升生產率，是推動經濟高品質發展的核心動力所在。相比企業數、進出口等，創新驅動對 GDP 持續增長的底層支撐作用更顯著。

從繪製的散點圖图 3-6 來看，從業人員、進出口額、財政支出、消費總額、企業數、RD 經費等變數與 GDP 之間大致呈現出正相關趨勢，意味著隨著這些變數數值的增加，GDP 有上升的傾向，體現了勞動力投入、對外貿易規模擴大、政府財政資金投入、國內消費增長、企業數量增多以及研發投入加大等因素對經濟增長的積極推動作用。例如，消費總額散點圖中，散點較為明顯地呈現從左下到右上的分佈態勢，表明消費對經濟增長的有力拉動。然而，三產占比與 GDP 的散點圖分佈相對分散，反映出產業結構與經濟增長的關係更為複雜，說明除了第三產業占比外，還有其他因素影響 GDP，比如產業間協同發展等。

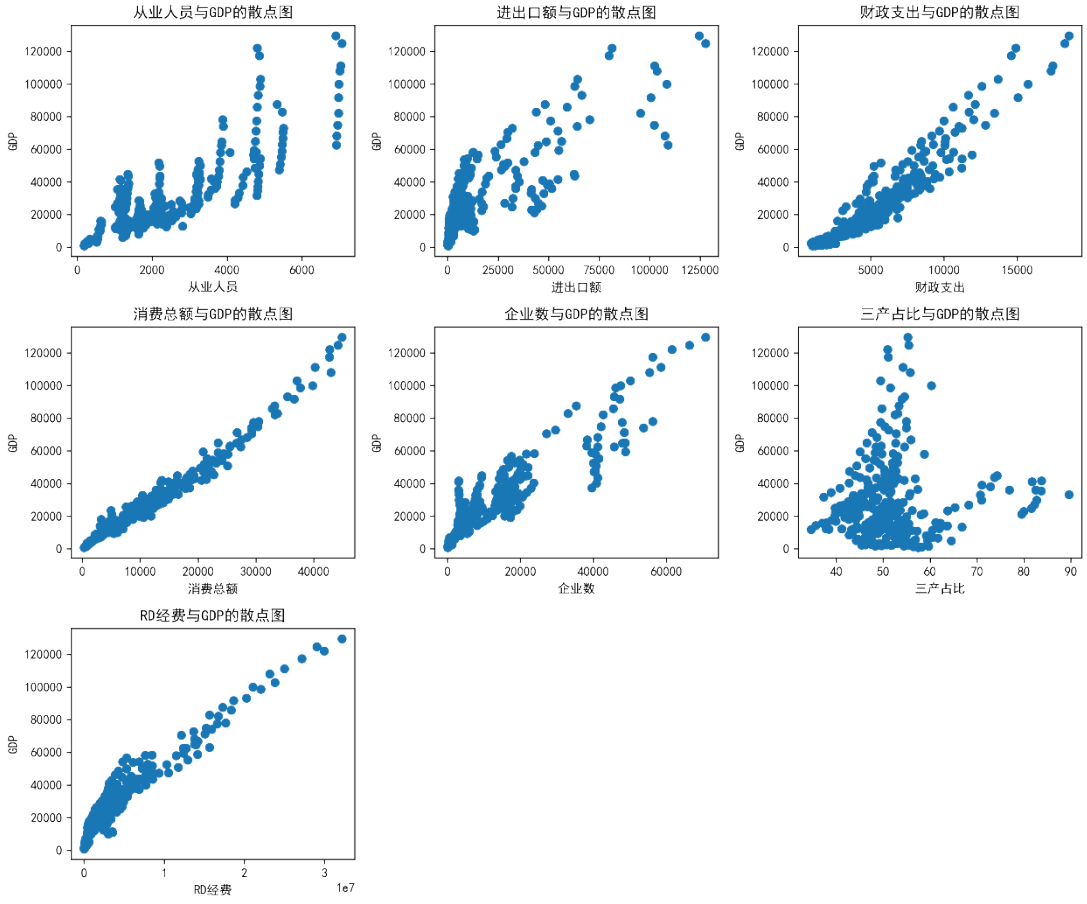


圖 3-6 各變數與 GDP 關係散點圖

根據這四類統計指標，構建了 2022 年中國各省發展資料的空間分佈圖。從图 3-7 上可以直觀分析到，東部沿海地區如北京、上海、廣東、江蘇、浙江等地的 GDP 數值明顯高於全國平均水準。相比之下，中西部地區如西藏、青海、甘肅、新疆等地和東北三省的 GDP 數值較低，在圖中的顏色較淺。整體來看，中國各省份的 GDP 發展存在明顯的區域差異，中國東部沿海地區經濟發展水準較高，而中西部和東北地區經濟發展相對滯後。

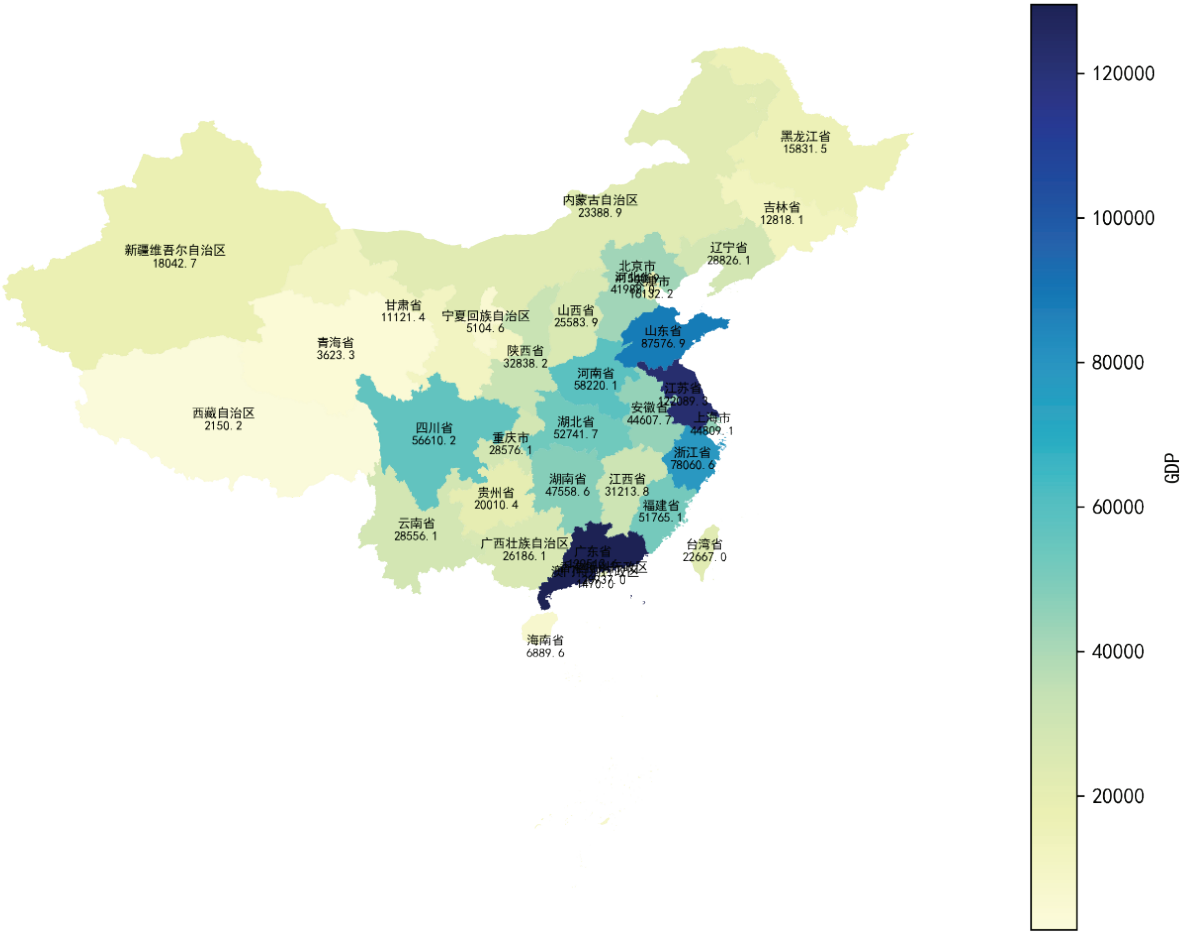


圖 3-7 區域 GDP 地圖

從图 3-8 中可以看到，2022 年廣東、江蘇、浙江、山東等地是第一個梯度的，從業人員數量最多，顏色較深，說明這些地區勞動力密集，經濟活躍；中部地區如河南、湖北、湖南屬於第二梯隊，從業人員數量中等。西藏、青海、遼寧、吉林、黑龍江等地從業人員數量最少，處於第三梯隊。中國各省份從業人員分佈呈現“東高西低”特徵，東部地區勞動力資源豐富，中西部和東北地區相對不足。隨著產業轉移和區域合作推進，這種分佈格局可能發生變化。

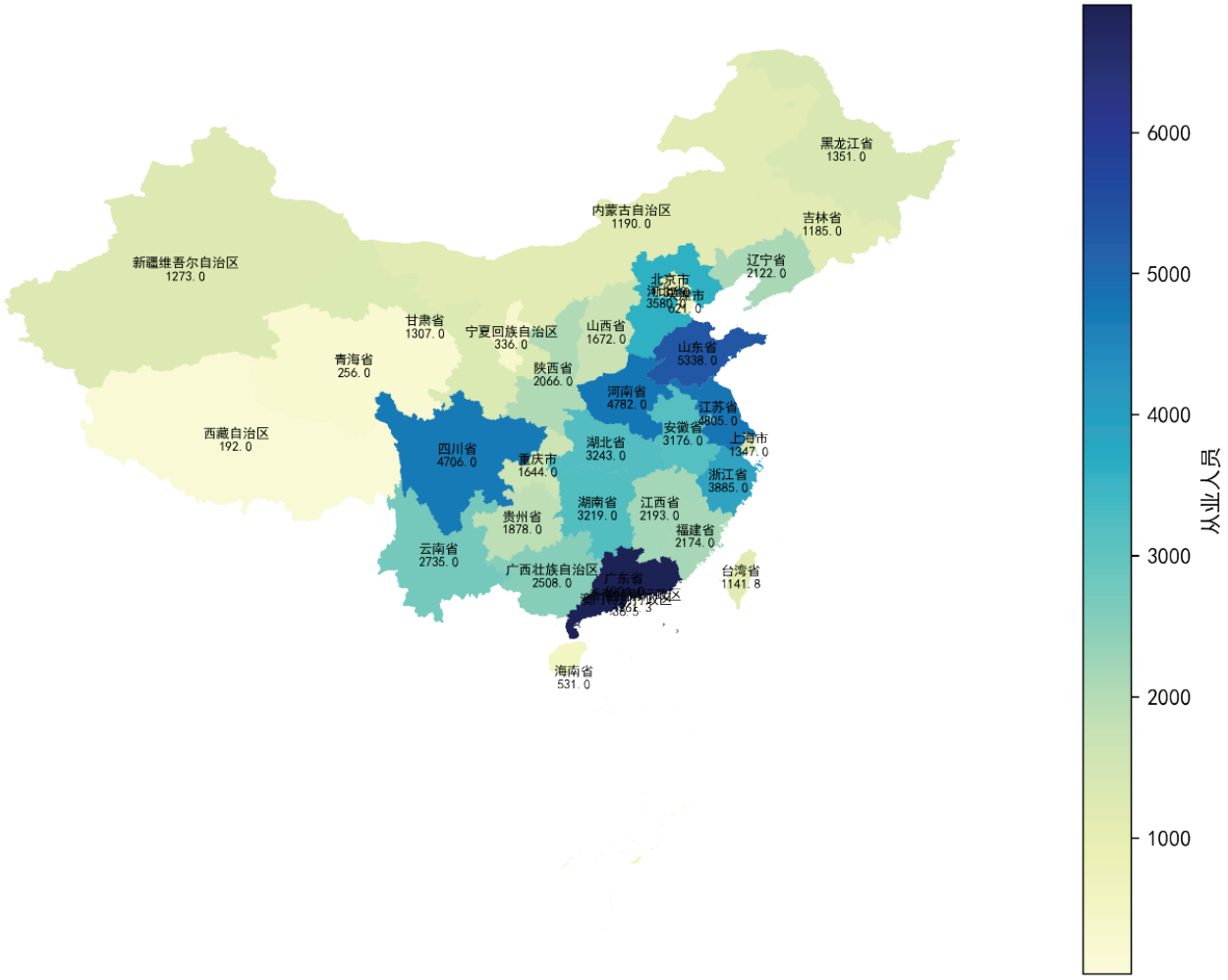


圖 3-8 區域從業人員地圖

图 3-9 顯示了中國進出口額的顯著地區差異，廣東的進出口貿易總額最高，達到了 129513.6 美元，江蘇、上海、北京、山東和浙江緊隨其後。福建、遼寧、湖北、四川等中部地區省份的貿易活躍度一般，而東北地區和西北地方的貿易活躍度最低。這種空間分佈格局顯示了進出口貿易發展的明顯東西梯度。沿海省份保持著先進的貿易網路，而中部、西部和東北地區則相對落後。隨著時間的推移，“一帶一路”倡議和中西部地區經濟的進一步開放可能會改變這種地理格局。

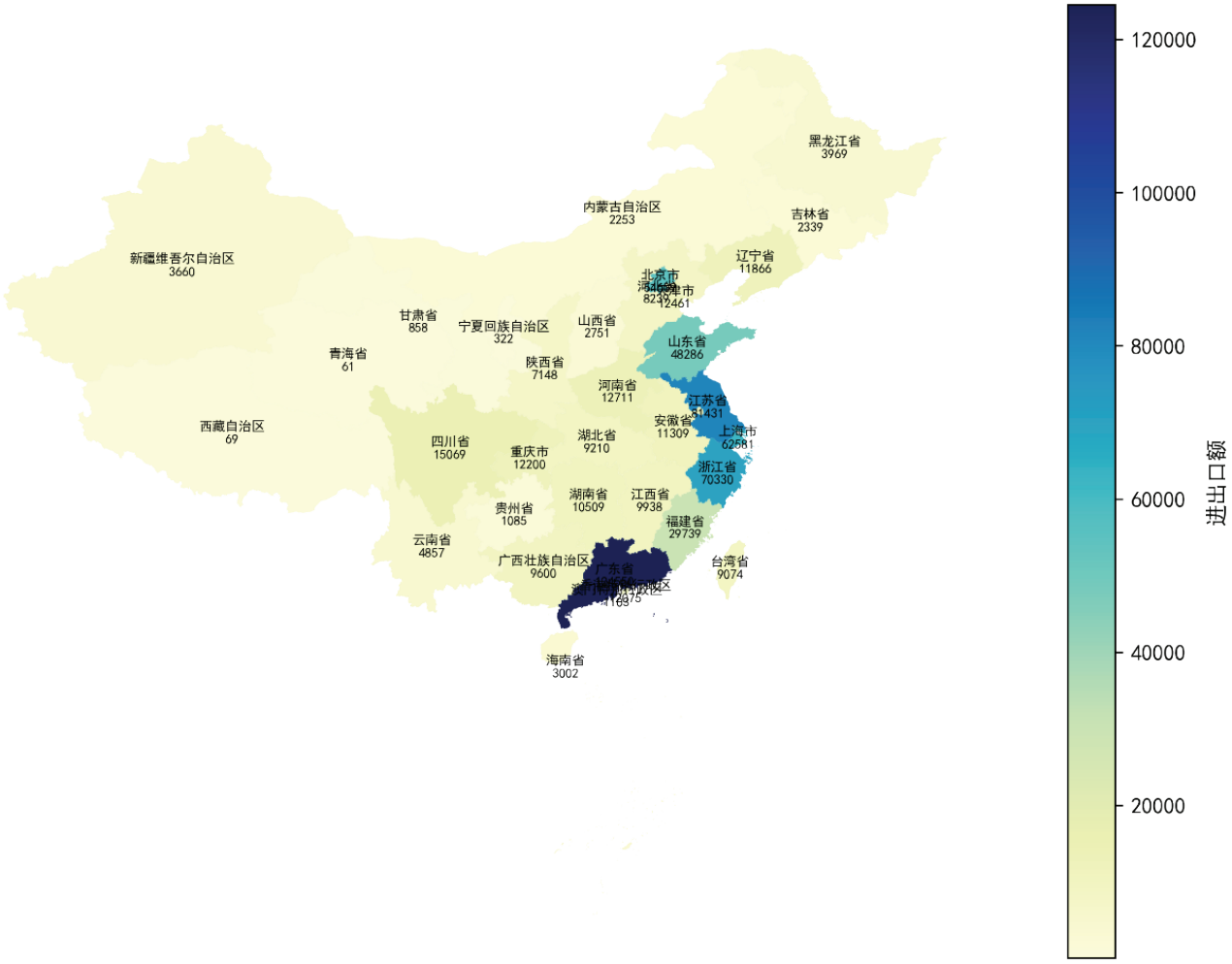


圖 3-9 區域進出口額地圖

图 3-10 顯示中國各省財政支出呈現明顯的區域差異特徵。具體來看，廣東、江蘇、河北、北京等省份的財政支出規模位於全國前列；浙江、山東、四川、湖北、湖南等省份的財政支出總額較高；西藏、青海、寧夏等西部地區財政支出水準偏低；這一分佈格局全國各省份財政支出額分佈不均衡的特徵。

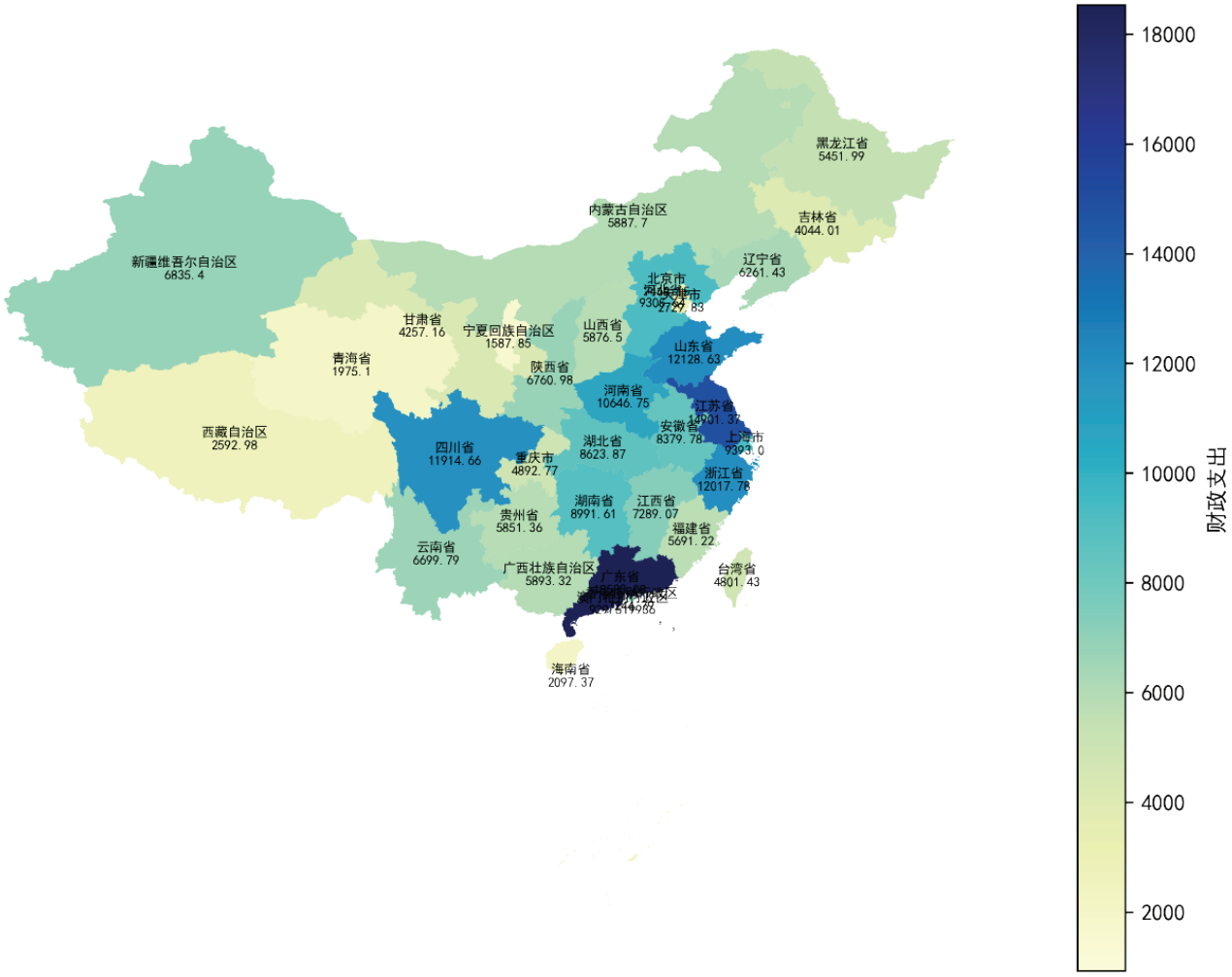


圖 3-10 區域財政支出地圖

從图 3-11 上可以直觀看到，廣東、江蘇、浙江、上海、山東等省份的消費總額最高，表明這些地區的消費水準高；四川、福建、湖南、河南等省份消費總額較高；中西部地區如西藏、青海、甘肅、新疆等省份的消費總額相對較低，東北地區包括遼寧、吉林、黑龍江的消費總額也相對較低。整體呈現“東高西低”的分佈特徵。

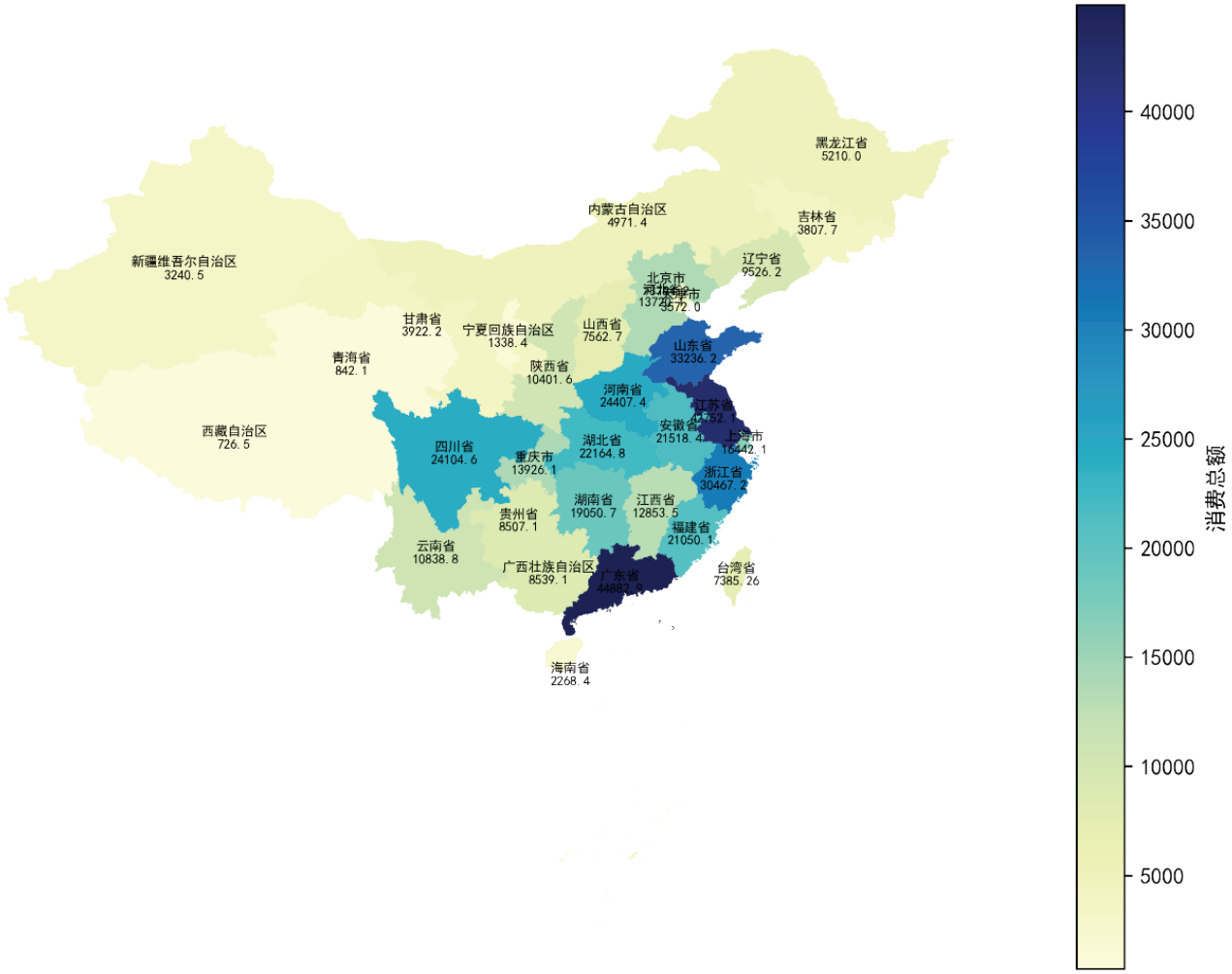


圖 3-11 區域消費總額地圖

如图 3-12 顯示，在所有省份中，廣東、江蘇、浙江、上海和山東的企業數量最多。緊隨其後的是河南、安徽、福建和湖南。遼寧、陝西、重慶和貴州的工業企業數量相對較少。相比之下，西藏和青海的工業企業數量最少。

，

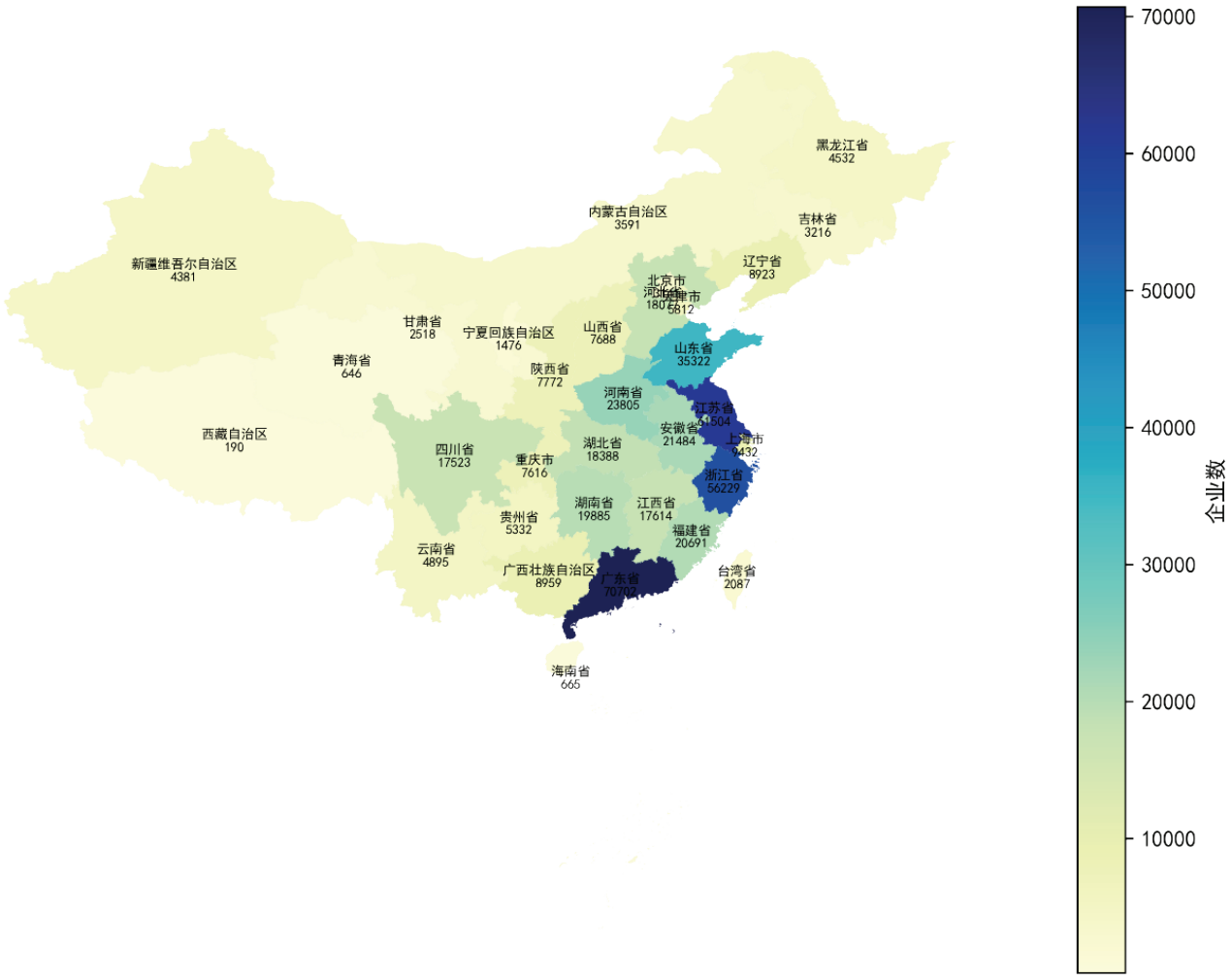


圖 3-12 區域工業企業數地圖

如图 3-13 所示，北京和上海的第三產業占比最高，接近 80%，這表明這些地區的經濟以服務業為主，金融、科技、文化等高端服務業發展迅速。廣東、江蘇、浙江等沿海發達省份的第三產業占比也較高，超過 50%，顯示出這些地區在製造業基礎上，服務業的快速發展。而中西部地區如四川、重慶等省市的第三產業占比中等，表明這些地區在服務業發展上仍有較大潛力。西藏、青海等省份的第三產業占比較低。

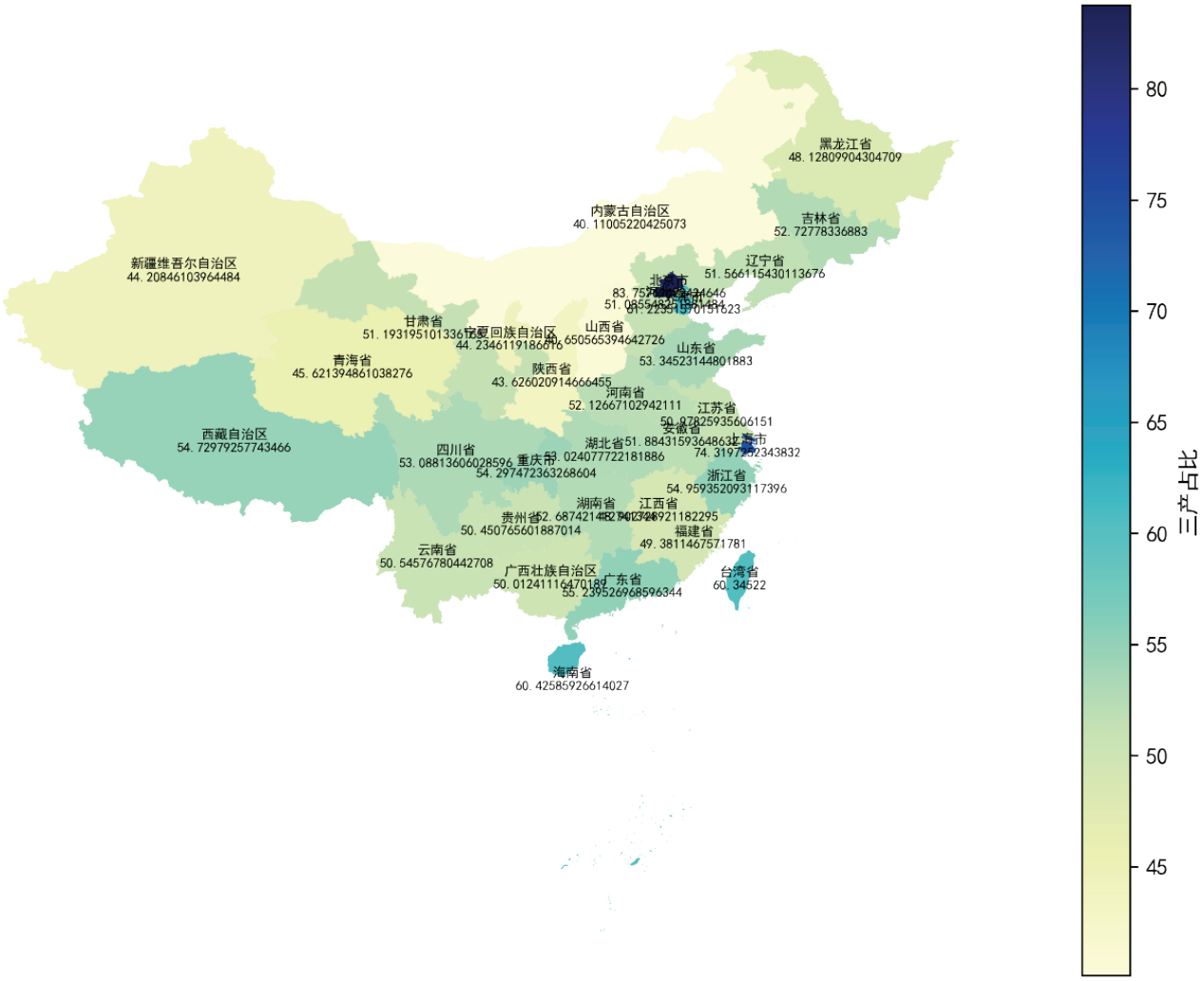


圖 3-13 區域第三產業占比地圖

從图3-14上可以看到各省R&D經費投入情況。廣東省投入最高，顯示其重視科技創新。江蘇，浙江緊隨其後，投入也較高。西藏投入最低，與其經濟水準和產業結構相關。總的來看，沿海發達省份研發投入普遍高於中西部欠發達地區。這一分佈反映了區域經濟發展差異，表明需要提升中西部科技創新能力，促進區域均衡發展。同時，這也指出了未來需要加強中西部地區科技創新能力，促進區域協調發展的重要性。

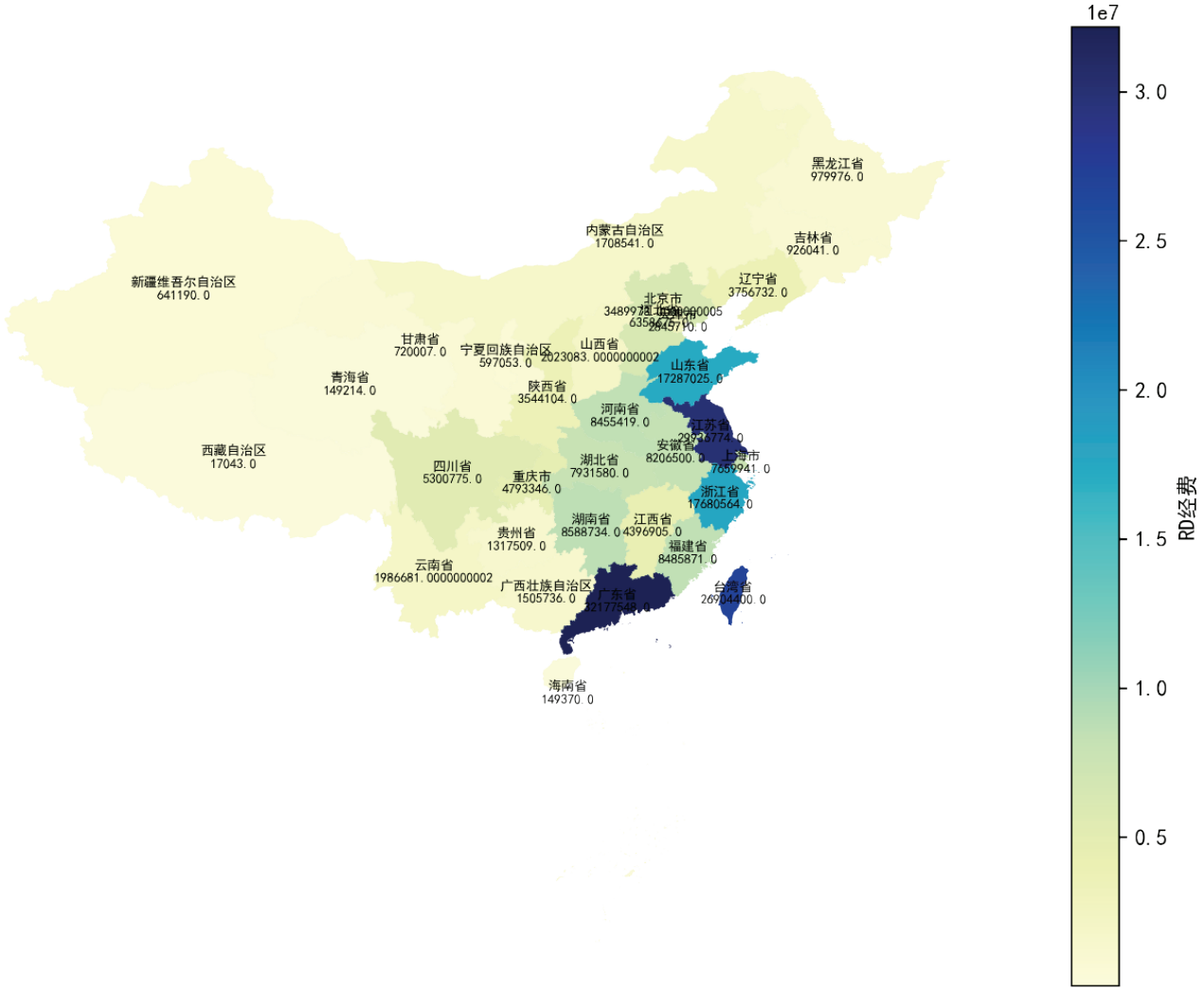


圖 3-14 區域全社會研究與試驗發展經費地圖

3.3 多元回歸模型構建

本研究採用多元線性回歸模型來分析影響中國經濟發展的各種因素。該模型旨在量化各因素對經濟發展程度的具體影響，從而揭示其對經濟增長的運行機制。

由多元回歸模型公式：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n+\epsilon$$

構建模型，模型設定如下：

$$GDP=\beta_0+\beta_1\times\text{從業人員}+\beta_2\times\text{進出口額}+\beta_3\times\text{財政支出}+\beta_4\times\text{消費總額}+\beta_5\times\text{企業數}+\beta_6\times\text{三產占比}+\beta_7\times R\&D\text{ 經費}+\epsilon$$

其中，因變數為 GDP，用以衡量各地區的經濟發展水準；引數包括從業人員數量、進

出口額、財政支出、消費總額、企業數量、產業結構占比和研發投入，這些變數分別從勞動力投入、國際貿易、政府調控、內需市場、市場主體活力、產業結構優化以及技術創新等不同維度對經濟發展產生影響。模型中的係數至表示各引數對因變數的影響程度，而為誤差項，用以捕捉模型中未包含的其他因素對因變數的影響。

基於上述模型，我們利用統計軟體 Python 對資料進行多元回歸分析，得處以下結果：從回歸結果图 4-1 來看，模型的 R-squared 值為 0.991，表示模型解釋了 GDP 變化中 99.1% 的方差，說明模型的解釋能力強，擬合效果較好。調整後的 R-squared 同樣為 0.991，說明即使考慮了變數數量的影響，模型的解釋能力依然很高，進一步驗證了模型的解釋能力。值為 4888，Prob(F-statistic) 為 1.44e-306，遠小於 0.05。模型整體顯著，至少有一個解釋變數對 GDP 有顯著影響。

通過建立模型，得出結果：GDP=2766.3745-0.094 從業人員 +0.0262 進出口額 +1.6814 財政支出 +1.2431 消費總額 -0.0655 企業數 -68.8742 三產占比 +0.0014RD 經費，=0.991。

條件數遠超過 100 的臨界值，且注釋明確提示“可能存在嚴重多重共線性”。這意味著解釋變數間存在高度線性相關，會導致係數估計值不穩定、標準誤放大，雖部分變數顯著，但係數實際經濟意義可能失真，需通過變數篩選，如剔除高度相關變數或使用正則化方法改進模型。從殘差圖上可以直觀看到，殘差未呈現隨機均勻分佈，而是隨著預測值增大，殘差的波動範圍明顯擴大（右側殘差點分散更開）。這表明模型可能存在異方差性，即誤差項的方差不恒定。殘差未圍繞“0”隨機分佈，且部分區域殘差集中偏離零值，說明模型對部分資料的擬合效果較差。接下來我們可採用主成分分析法剔除高度相關變數和加權最小二乘法（WLS）重新估計模型。

OLS Regression Results

Dep. Variable:	GDP	R-squared:	0.991
Model:	OLS	Adj. R-squared:	0.991
Method:	Least Squares	F-statistic:	4888.
Date:	Tue, 01 Apr 2025	Prob (F-statistic):	1.44e-306
Time:	14:47:23	Log-Likelihood:	-2833.0
No. Observations:	310	AIC:	5682.
Df Residuals:	302	BIC:	5712.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2766.3745	1249.314	2.214	0.028	307.912	5224.837
从业人员	-0.0948	0.258	-0.368	0.713	-0.602	0.412
进出口额	0.0262	0.013	2.068	0.039	0.001	0.051
财政支出	1.6814	0.134	12.546	0.000	1.418	1.945
消费总额	1.2431	0.065	19.154	0.000	1.115	1.371
企业数	-0.0655	0.037	-1.755	0.080	-0.139	0.008
三产占比	-68.8742	23.785	-2.896	0.004	-115.680	-22.069
RD经费	0.0014	0.000	12.976	0.000	0.001	0.002

Omnibus:	11.481	Durbin-Watson:	0.569
Prob(Omnibus):	0.003	Jarque-Bera (JB):	17.581
Skew:	0.247	Prob(JB):	0.000152
Kurtosis:	4.057	Cond. No.	6.71e+07

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.71e+07. This might indicate that there are
strong multicollinearity or other numerical problems.

圖 3-15 回歸分析圖

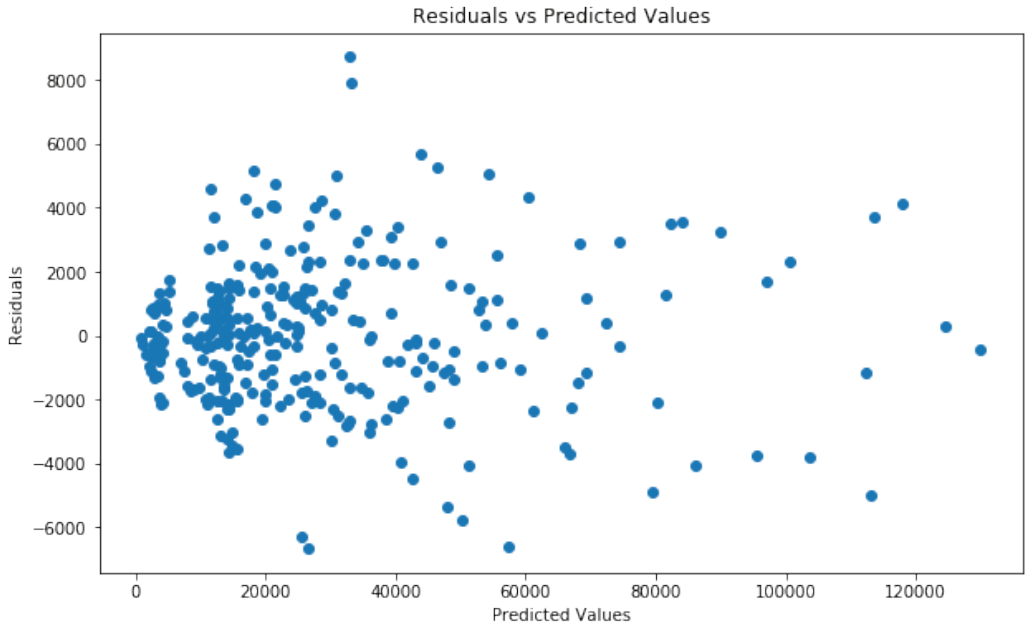


圖 3-16 殘差圖

3.4 加權最小二乘法（WLS）和主成分回歸

WLS 通過加權調整，修正了普通回歸中可能存在的異方差問題，使誤差項滿足同方差假設，提升了參數估計標準誤的準確性，讓變數顯著性檢驗（如 t 檢驗）結果更可靠。在異方差場景下，普通回歸估計量不再是最佳線性無偏估計（BLUE），而 WLS 通過加權使估計量重新滿足 BLUE 性質，參數估計更貼近真實值。模型 R^2 和 R^2 調整後均為 1.000 表明引數對因變數 GDP 的解釋能力極強，幾乎能涵蓋 GDP 的所有波動，體現了模型對樣本資料的高度適配。

根據 WLS 回歸結果的係數，回歸方程為：

GDP=-1.021×10^6+12.4047 從業人員 +0.2975 進出口額 +0.5535 財政支出 +1.1746 消費總額 +0.4044 企業數 −236.3484 三產占比 +0.0015RD 經費

加权最小二乘法（WLS）模型结果：

WLS Regression Results

Dep. Variable:	GDP	R-squared:	1.000
Model:	WLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	3.947e+05
Date:	Tue, 01 Apr 2025	Prob (F-statistic):	2.53e-06
Time:	17:59:00	Log-Likelihood:	-74.925
No. Observations:	10	AIC:	165.8
Df Residuals:	2	BIC:	168.3
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.021e+06	1.63e+05	-6.264	0.025	-1.72e+06	-3.2e+05
从业人员	12.4047	1.789	6.935	0.020	4.708	20.101
进出口额	0.2975	0.014	21.169	0.002	0.237	0.358
财政支出	0.5535	0.106	5.201	0.035	0.096	1.011
消费总额	1.1746	0.055	21.488	0.002	0.939	1.410
企业数	0.4044	0.154	2.632	0.119	-0.257	1.065
三产占比	-236.3484	175.337	-1.348	0.310	-990.761	518.064
RD经费	0.0015	0.000	5.252	0.034	0.000	0.003

Omnibus:	19.104	Durbin-Watson:	2.650
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11.450
Skew:	-1.988	Prob(JB):	0.00326
Kurtosis:	6.417	Cond. No.	2.07e+11

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.75e-11. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

圖 3-17 WLS 回歸分析圖

各變數對經濟發展的影響如下：

(1) 從業人員數量：其回歸係數為 6.935，P 值為 P>t=0.020（<0.05），表明在其他引數不變情況下，從業人員對 GDP 的影響顯著。這意味著勞動力投入的增加能夠促進經濟的增長，每增加一個單位的從業人員數量，GDP 將增加 6.935 個單位。

（2）進出口額：回歸係數為 0.2975，且 t 值 21.169，P>t=0.002（小於 0.05），說明進出口額對經濟發展具有正向作用。國際貿易能夠通過技術引進、市場拓展以及產業協同等效應，帶動國內產業的升級和經濟的增長。進出口額每增加一個單位，GDP 預計將會增加 0.2975 個單位。

（3）財政支出：其係數為 0.5535，P 值顯著，表明政府的財政支出對經濟發展有積極的推動作用。財政支出在教育、基礎設施建設等領域的投入，能夠提升人力資本素質、改善投資環境，進而促進經濟的增長。

（4）消費總額：回歸係數為 1.1746，t 值 21.488，P>t=0.002（<0.05），P 值顯著，且具有統計學意義，說明消費總額對 GDP 的影響顯著且為正，說明消費是推動 GDP 增長的重要驅動力。消費增長能夠直接刺激內需，促進商品和服務的流通，推動生產的發展，從而帶動經濟的增長。

（5）企業數量：係數為 0.4044，P>t=0.119（>0.05），企業數對 GDP 的影響不顯著，可能暗示企業數量的增加並未帶來生產效率的提升。

（6）第三產業結構占比：其係數為 -236.3484，P>t=0.310（小於 0.05），三產占比對 GDP 的影響不顯著，這可能暗示第三產業的效率較低，或者資源在第三產業中的配置存在問題

（7）研發投入：回歸係數為 0.0015，t 值 5.252，P>t=0.034（小於 0.05），P 值顯著，說明研發投入對經濟發展具有顯著的正向影響。技術創新是推動經濟持續增長的核心動力，研發投入的增加能夠提升生產效率、促進產品升級，增強經濟的競爭力和增長潛力。

在進行主成分分析之前，必須使用方差膨脹因數（VIF）來評估資料中的多重共線性。通常情況下，VIF 值超過 5 表示變數之間可能存在較強的多重共線性，而 VIF 值超過 10 則表示存在嚴重的多重共線性。測試結果如图 3-18 所示：從業人員、財政支出、消費總額、企業數和 RD 經費的 VIF 值均超過 10。這表明這些變數之間存在嚴重的多重共線性。在資料分析或建模過程中，這種嚴重的多重共線性會導致參數估計不穩定、解釋困難等問題。此外 ,進出口額的 VIF 值大於 5 ,意味著可能與其他變數存在較強的多重共線性。相反，三產占比的 VIF 值低於 5，表明與其他變數之間的多重共線性極小。

方差膨胀因子（VIF）测试结果：

	Variable	VIF
0	const	92.897411
1	从业人员	10.561941
2	进出口额	5.430740
3	财政支出	10.802273
4	消费总额	22.841372
5	企业数	16.973578
6	三产占比	2.504077
7	RD经费	22.438851

圖 3-18 方差膨脹因數測試結果

接下來運用主成分分析通過將原始變數轉化為一組互不相關的主成分，有效地解決了多重共線性問題，並且保留了原始資料中大部分的資訊。選擇了累積方差貢獻率達到 95% 的主成分，最終保留了 3 個主成分（PC1、PC2、PC3）。新資料集包含 310 行 7 列資料，保留了原資料的編碼、地區、年份和 GDP 列，以及新生成的 3 個主成分列。這 3 個主成分並不是直接對應原來的某幾個引數，而是原始引數的線性組合。每個主成分是通過對原始引數進行加權求和得到的，權重由主成分分析過程中的特徵向量決定。

如表 3-3 在 PC1 中，三產占比的載荷絕對值相對其他變數非常小，而從業人員、進出口額、財政支出、消費總額、企業數和 RD 經費的載荷值較為接近且絕對值較大，說明這些變數對 PC1 的貢獻比較均衡，且共同正向影響 PC1。可以推測 PC1 反映了整體經濟活動規模相關的綜合因素，因為這些變數都與經濟活動的規模和活躍度有一定關聯。

PC2 中三產占比的載荷值高達 0.892093，遠高於其他變數，表明三產占比對 PC2 起主導作用。而從業人員和企業數的載荷為負，說明它們對 PC2 有反向的影響。可以認為 PC2 主要反映了產業結構方面的因素，特別是第三產業在整體經濟中的占比情況。

在 PC3 中，進出口額的載荷為 -0.525061，財政支出的載荷為 0.502132，兩者絕對值較大且一正一負。這可能意味著 PC3 體現了對外貿易與政府財政支出之間的相互關係，反映了經濟中內外需平衡或者政策調控與對外貿易之間的某種權衡關係。

表 3-3 主成分載荷表

	PC1	PC2	PC3
原始引數	載荷值	載荷值	載荷值
從業人員	0.390092	0.286586	0.430291
進出口額	0.376660	0.311960	-0.525061
財政支出	0.410353	0.039652	0.502132
消費總額	0.429657	0.002627	0.173937

	PC1	PC2	PC3
企業數	0.416144	-0.149732	-0.303462
三產占比	0.031216	0.892093	0.257518
RD 經費	0.422860	0.026851	-0.313632

在完成上述實證分析後，為進一步加強對中國經濟影響因素的理解，下面從四個方面補充具體例子進行深入分析。在分析各因素對經濟發展的影響時，以具體省份數更有說服力。就從業人員數量而言，江蘇省就是一個很有代表性的例子。從 2013 年到 2022 年，江蘇省就業人員數量從 4722 萬人逐步增加到 4805 萬人。隨著就業人數的穩定增長，江蘇省的 GDP 也從 59349.4 億元增至 122089.3 億元。詳細計算顯示，在此期間，每增加 1 萬名從業人員，GDP 平均將增加約 13.2 億元。這與模型得出的從業人員數量對 GDP 有積極影響的結論是一致的，並清楚地表明充足的勞動力投入對經濟增長有顯著的促進作用。

在進出口規模方面，浙江取得了優異成績。2013 年至 2022 年，浙江進出口規模從 33578.87 億美元上升到 70330 億美元，同期 GDP 由 37334.6 億元增加至 78060.6 億元。進出口規模每增加 1 億美元，國內生產總值大約增加 0.65 億元。這體現了國際貿易對浙江經濟增長的重要促進作用，通過進出口項目，浙江可以實現資源的優化配置，拓展市場空間，推動相關產業發展，促進整體經濟的增長。

區域經濟發展不平衡是中國經濟發展的一個重要特徵。以東部的上海和西部的青海為例，兩地的經濟指標差異顯著。2022 年，上海市國內生產總值為 44809.1 億元，進出口額 62581 億美元，RD 經費投入 7659941 萬元；而青海省同年 GDP 僅 3623.3 億元，進出口規模 61 億美元，RD 經費投入 149214 萬元。在產業結構上，上海以金融、科技、高端製造等現代服務業和先進製造業為主導，產業附加值高，創新能力強。眾多金融機構彙聚上海，如浦發銀行總部等，推動金融市場蓬勃發展，為經濟增長提供強大動力。同時，上海在人工智慧、生物醫藥等高科技領域投入大量資源，成果豐碩。而青海省產業結構相對單一，主要依賴資源型產業，如鹽湖化工等。這種產業結構使得青海省經濟過度依賴資源，抗風險能力差，與上海形成明顯差距。上海利用自貿區等政策優勢吸引了大量外資，推動經濟的快速發展。青海受自然條件限制，招商引資難度大，政策措施效果有限。這一對比顯示地區發展不平衡。政策須考慮區域條件，因地制宜，實施差異化發展戰略。

在財政政策方面，近年來，山東省積極優化財政支出結構，加大對教育和基礎設施建設的投入。在教育領域，山東省增加對高校科研專案的資金支援，如對山東大學等高校的重點學科建設給予專項財政撥款，推動高校科研水準提升，為企業輸送了大量高素質人才。

在基礎設施建設方面，加大對交通、能源等領域的投資，如修建高速公路和高鐵網路，改善了投資環境，降低了企業物流成本。這些舉措促進了經濟的增長，使得山東省在 2013-2022 年間經濟保持穩定增長，GDP 從 47344.3 億元增長到 87576.9 億元。在刺激內需消費政策上，四川省採取多項措施刺激消費。首先，實施就業政策，發展電子資訊、文創等新興產業，創造就業崗位，提高居民收入。其次，擴大社保覆蓋，提高醫保報銷比例，減輕醫療負擔，增強消費信心。這些政策效果顯著：全省消費總額從 2013 年 11001 億元增至 2022 年 24104.6 億元，有效拉動經濟增長

在研發投入對經濟發展影響的研究中，以資訊技術行業為例，北京市在該行業的研發投入成果顯著。近年來，北京對資訊技術行業的研發投入持續增加，吸引了大量相關企業和高端人才。以位元組跳動為例，在強大研發投入的支持下，其旗下的抖音、今日頭條等產品不斷進行技術創新和功能升級，不僅在國內擁有龐大用戶群體，還在國際市場上取得巨大成功。這些產品的成功帶動了廣告、電商、直播等相關產業的發展，創造了大量就業機會和經濟效益，為北京市的經濟增長做出重要貢獻。在製造業領域，廣東省佛山市是傳統製造業向高端化轉型的典型。佛山市對製造業的研發投入不斷加大，推動了傳統陶瓷、家電等產業的升級。例如，美的集團加大研發投入，在智慧家電領域取得多項技術突破，產品智慧化水準不斷提高，市場競爭力增強。通過研發創新，美的集團不僅實現自身業績增長，還帶動了整個家電產業鏈的發展，促進了佛山市經濟結構的優化和升級，進一步證明了研發投入在推動行業發展和經濟增長中的核心作用。

3.5 研究結論

本研究借助多元回歸模型對多個經濟影響因素展開考察，這些因素涉及從業人員數量、進出口額、財政支出、消費總額、企業數量、產業結構占比以及研發投入等方面，主要的發現如下所示 :

第一，從業人員數量：研究說明就業人口數量的增長可有力地推動經濟增長，這意味著勞動力投入乃是經濟增長的關鍵基礎。統計資料體現出，每當就業人口增加一個單位，國內生產總值便會出現較大的增長，充足的勞動力可為生產以及服務等各類經濟活動注入更為強大的活力，促進經濟產出的增長。

第二，進出口額：研究結果顯示進出口額的上升對經濟發展有著較大的促進作用，借助國際貿易，進出口額的增長可實現資源配置的優化，引進先進技術，推動產業的升級，是經濟持續增長的關鍵動力。

第三，財政支出：財政支出對於經濟發展有著正向作用，政府投資可有效帶動經濟增長，政府借助教育、基礎設施建設以及科技創新等公共支出方式，可以提高人力資本素質，

改善投資環境，促使經濟結構得到優化，這對長期發展有關鍵作用。

第四，消費總額：消費總額的增長可推動 GDP 提升，這與消費拉動經濟增長的理論相契合。國內消費需求的擴大可激發市場活力，促進商品流通以及生產發展，是經濟增長的主要動力來源。

第五，研發投入：研發投入對經濟發展具有重大而有利的影響，技術創新是經濟持續增長的核心推動力。提高研發支出可提升生產力和競爭力，從而為高水準的經濟發展提供強有力的支援。

第六，第三產業占比：第三產業占比對 GDP 的影響不顯著，第三產業占比提升未顯現預期效果，這可能表明第三產業的效率較低，或者資源在第三產業中的配置存在問題。儘管第三產業在經濟中的比重逐漸增加，但對經濟增長的直接拉動作用有限，需提升服務業效率。

第七，企業數量：企業總量增加未帶來相應效益。資料顯示企業數量與 GDP 增長無顯著正相關，反映需要提升企業品質而非數量，需要進一步關注企業的創新能力和市場競爭力。

第八，區域經濟發展的不平衡性：通過對中國東部、中部、西部和東北地區的經濟資料進行視覺化分析，發現四大經濟區域發展差異顯著，東部地區發展優勢明顯。視覺化分析顯示，東部在總量、外貿、創新等方面領先，中西部和東北地區相對滯後，區域發展不平衡問題突出。這種區域不平衡性反映了中國區域經濟發展的不均衡性，需要通過區域協調發展戰略，促進資源的合理流動和區域間的協同發展。

總的來說，經濟增長的主要驅動因素包括：就業人數、進出口額、財政支出、消費總額、研發投入。但研究存在以下局限：第一，資料範圍有限，難以反映微觀差異和長期變化；第二，未考慮數字經濟等新指標，可能遺漏重要因素；第三，區域分析不夠細緻，缺少城市群研究；第四，模型存在內生性問題，未考察非線性關係。未來研究可從三方面深化：首先，加入數位技術、綠色經濟等新興變數，分析其對傳統產業的影響；其次，使用空間計量模型研究區域互動，結合企業資料考察結構問題；最後，應用機器學習處理複雜線性關係，利用衛星資料等新方法拓展研究。

根據上述研究結果，提出以下政策建議：

第一，優化勞動力市場：完善勞動力市場機制，著力破除勞動力流動壁壘，促進人力資源在區域、行業間的合理配置。加強職業培訓，增加職業教育投入，建立多層次培訓體系。推動產教融合，提升培訓實效性。提高人才素質，培養適應產業升級的技能人才，滿足高品質發展需求。要鼓勵創新創業帶動就業，創造更多高品質的就業崗位，這將同時激發市

場活力和社會創造力，為經濟增長注入新動力。

第二，促進對外貿易：要借助國際市場多元化來優先推動對外貿易，鼓勵企業去開拓新興國際市場，降低對傳統單一市場的依賴程度，並且需優化貿易結構，提升高附加值產品的出口比例，促使貿易結構朝著高端化方向發展，另外還應深化海關、檢驗檢疫等貿易便利化改革，以此提升貿易便利化水準，借助簡化進出口流程，降低貿易成本來實現這一目標。

第三，制定合理財政政策：調整優化財政支出結構，重點增加教育、科研和基礎設施建設的資金投入，提高資金使用效率。其次，嚴格控制一般性行政支出，減少無效開支，確保財政資金重點支持經濟增長和社會發展關鍵領域。同時，要加大財政政策與產業政策的配合力度。通過實施財政補貼、稅收減免政策以及設立專項基金等一系列舉措，有針對性地對戰略性新興產業、高新技術產業和現代服務業進行重點培育，以此促進產業結構向更高層次邁進。

第四，推動內需消費增長：推行就業優先戰略，全力創造更多適配各類勞動力的就業崗位，以此提升居民收入，為消費增長奠定堅實經濟基礎。擴大醫療、養老等社會保障體系覆蓋範疇，不斷提高保障標準，解決居民後顧之憂，增強居民消費信心。強化市場監管力度，構建全方位、多層次的市場監管網路，嚴厲打擊假冒偽劣商品和商業欺詐行為，維護消費者權益。通過完善消費環境、培育消費熱點等措施，充分釋放居民消費潛力。

第五，加快發展現代服務業，重點培育金融、物流、資訊技術等高附加值服務業，提升服務業在經濟結構中的比重和品質。支援服務業創新發展模式，推動傳統服務業向數位化、智慧化方向轉型。

第六，強化政府對研發領域的支持力度：政府需持續加大財政資金向基礎研究以及核心技術攻關領域的傾斜力度，以此提升自身自主創新能力。門設立科研專項基金，積極推動高校、科研院所與企業攜手開展聯合科研攻關行動。進一步完善企業研發費用加計扣除等稅收優惠政策體系，借此激勵企業主動增加研發投入，有效降低創新成本負擔。

4. 回歸分析方法在教育收益方面的實證研究設計

本文研究的資料來源於北京大學中國社會科學調查中心發佈的中國家庭動態跟蹤調查（Chinese Family Panel Studies, CFPS）。CFPS 是一項系統全面的調查，涵蓋了社區、家庭和個體等多個層面的資料，旨在反映中國社會、經濟、人口、教育和健康的變遷。該調查通過科學的抽樣方法和嚴謹的資料收集流程，為研究中國社會提供了高品質的資料支援。選擇使用 2022 年的最新資料，保留 18-60 歲完成高中階段教育者共包含 32745 個樣本，涉

及個體的最高學歷、年齡、工作經歷、月收入、性別、教育類型、工作單位、戶口、政治面貌等資訊。足夠用來分析本文研究的兩個主要研究問題。確保分析結果的時效性和準確性。為本文的研究提供了堅實的資料基礎。

在資料分析中，變數值缺失是一個普遍存在的問題。缺失值的存在可能會影響資料的完整性和分析結果的準確性。為了保證資料的可用性，本文採取了直接刪除法對缺失值進行處理。具體而言，通過對缺失值的檢查，發現品質等級缺失值是 470，由於樣本總量是 32745，所以缺失率是 470/32745≈0.014%，也就是不到 0.2%。這個比例非常低，只有 0.2% 的缺失資料。通常來說，如果缺失率低於 5%，直接刪除不會對結果產生太大影響，尤其是在樣本量很大的情況下，既可以有效避免複雜填補操作引入額外偏差，又符合”奧卡姆剃刀”原則 -- 簡單有效。

因此本文採用直接刪除缺失值的方法。這種方法能夠保持原始資料的趨勢和變化，同時避免引入額外的誤差，從而符合線性模型的假設條件。雖然直接刪除法可能會導致部分樣本的丟失，但在確保資料品質和分析結果可靠性方面具有顯著優勢。

在完成資料處理後，本文對全樣本部分變數進行描述性統計分析。為探究“普職分流”政策下不同教育路徑的經濟回報差異，需明確區分普通教育與職業教育群體。本文將普通教育組定義為未接受職業教育的群體，包括高中、大專、本科。職業教育組則涵蓋接受中等職業教育的群體，如職高、中專、技校。教育類型變數設定為：普通教育組 =0，職業教育組 =1。

性別是影響教育回報的重要異質性因素。本文對關鍵引數進行數值化編碼，性別變數設定為：女性 =1，男性 =0。同時，考慮到家庭收入資料存在右偏性問題這可能對線性模型的正態性假設造成影響，因此我們對家庭收入進行了對數化轉換。這種處理方式不僅有助於滿足模型的正態性要求，還能有效避免因部分樣本收入為零而導致的對零值取對數時產生的缺失值問題。通過這一系列處理，我們為後續分析奠定了堅實的資料基礎。

工作經驗是決定收入的重要變數，通常呈現“倒 U 型”關係。根據明瑟方程，工作經驗的計算公式為：

$$工作經驗 = 年齡 - 受教育年限 -6$$

假設 6 歲為入學年齡，對於計算結果為負值的樣本（如未達工作年齡），將其替換為 0，並引入工作經驗的平方項以捕捉非線性效應。

此外，為控制省份層面的異質性干擾（如經濟水準、政策資源與文化差異），在回歸模型中納入省份固定效應。通過上述處理，本文能夠在保持資料合理性的同時，有效剝離混雜因素對教育收益率的潛在影響，從而提高研究結果的準確性和可靠性。

在“普職分流”政策背景下，探究教育收益率的性別差異化問題需要採用遞進式模型

設計，分層次解析其作用機制。首先，基於全樣本混合普通最小二乘法（OLS）回歸，估計教育類型、性別對收入的總體影響，並控制工作經驗、婚姻情況、年份和戶口固定效應。其次，通過分性別子樣本回歸檢驗教育收益率的異質性，分別計算男性和女性群體的教育邊際回報率，比較兩者差異。進一步構建教育年限與性別的交互項，通過交互項係數的顯著性及經濟含義，驗證性別因素是否顯著調節教育投入對收入的邊際效應，從而揭示教育收益率的性別差異形成機制。

首先採用全樣本混合 OLS 回歸方法，其核心目的在於全面估計教育年限、工作經驗等關鍵變數對收入所產生的總體影響。在這一過程中，將性別、教育類型、年份以及戶口固定效應等諸多可能影響收入的因素一併納入控制範圍。關鍵運算式如下：

$$\ln(Y_i)=\beta_0+\beta_1 Edu+\beta_2 Exp+\beta_3 Exp^2++\gamma_1 Female+\delta(EduType_i\times Gender_i)+\epsilon_i$$

其中，=1 表示職業高中，0 為普通高中；Gender_i=1 表示女性，0 表示男性 ;

表 4-1 OLS 回歸結果

變數名	係數	穩健標準誤	t 值	P 值	95% 置信區間
職業教育	0.2640182	0.052464	5.03	0.000	[0.1611546,0.3668817]
性別	0.0657454	0.0329209	2.00	0.046	[0.001199，0.1302918]

注：性別變數係數為正表示男性收入更高

從表 4-1 資料可以看出，在混合普通最小二乘法模型中，職業教育變數的係數為 0.2640。這表明在控制其他變數的情況下，接受職業教育的個體對數收入顯著高於普通教育群體。在保持其他所有變數不變的前提下，接受職業教育的個體相較於接受普通教育的個體，其對數收入平均邊際效應增加約 2.64%。這一結果驗證了“普職分流”政策下職業教育的經濟價值。

再看性別變數的回歸係數，發現教育收益率存在顯著分化，其值為 0.0657，表明男性對數收入顯著高於女性，邊際效應約為 6.57%。男性通過職業教育獲得的收入溢價顯著高於女性，且女性在職業教育群體中的占比不足 7%，反映出教育路徑選擇中的性別偏好差異。這一發現具有重要的研究價值，它反映出樣本中性別收入差異的局部特徵。在勞動力市場中，性別歧視、職業隔離以及社會分工差異等因素，都可能導致這種性別收入差異的出現。然而，僅通過當前全樣本混合 OLS 回歸結果還不能完全確定這一差異的穩健性，因此需要進一步結合分性別回歸進行驗證。

分性別回歸是通過將樣本按性別分組，分別建立獨立回歸模型，以檢驗引數（教育類型、性別）對因變數（收入）的影響是否存在性別異質性。本研究採用分性別回歸方法，為進

一步驗證性別差異的穩定性，分別建立男性和女性獨立模型，以檢驗“普職分流”政策下職業教育對收入影響的性別異質性。關鍵運算式如下：

$$\ln(Y_i)=\beta_0+\beta_1 edu+\beta_2 exp+\beta_3 exp^2++\gamma_1 female+\gamma_2 exp^2+\delta voc\_edu+\epsilon_i$$

表 4-2 分性別回歸結果

變數名	係數	穩健標準誤	t 值	P 值	95% 置信區間
職業教育 - 男性	0.2535086	0.0730611	3.47	0.001	[0.110225,0.3967921]
職業教育 - 女性	0.2829519	0.0739536	3.83	0.001	[0.1378845,0.4280193]

在“普職分流”政策背景下，教育收益率的性別差異呈現顯著分異性特徵。職業教育路徑對男女性均具有正向收益。在男性樣本中，職業教育係數約為 0.2535，略低於全樣本均值，但 t 值為 3.47，P 值為 0.001，表明職業教育對男性收入提升作用穩定。在女性樣本中，職業教育係數為 0.2829519，相對較高，說明女性通過職業教育獲得的收入回報更顯著，且存在職業發展路徑的結構性差異男性職業教育群體收入增長更依賴工作經驗積。是因為男性工作經驗係數 -0.021 顯著，女性不顯著，而女性職業教育者面臨更明顯的職業天花板效應，工作經驗平方項係數 -0.0004 顯著，反映收入增速隨工齡增長加速下降。但 t 值為 4.97 性別收入差異在男性內部可能因行業分佈（如製造業、建築業）不同而存在異質性。

下面通過構建“性別 × 教育類型”交互效應模型，系統檢驗了性別對“普職分流”政策下教育收益率的調節作用。模型在分性別回歸的基礎上引入交互項（性別 × 教育類型），以量化性別與教育類型的聯合效應。回歸結果顯示如下：

表 4-3 拓展模型結果

變數名	係數	穩健標準誤	t 值	P 值	95% 置信區間
性別 * 職業教育	0.20411	0.029745	6.8734	0.000	[0.14667, 0.26257]
男 普通教育	10.82961	0.0266148	506.90	0.000	[10.77743-10.8818]
男 職業教育	11.05295	0.0704461	156.90	0.000	[10.91483-11.19107]
女 普通教育	10.8891	0.0220683	493.43	0.000	[10.84583-10.93237]
女 職業教育	11.18965	0.0667153	167.72	0.000	[11.05885-11.32046]

交互項係數顯著為正這表明職業教育的收入回報在女性群體中更高，性別差異通過教育類型進一步被放大。通過計算邊際效應發現，男性職業教育群體的對數收入比普通教育群體高 0.22，約為 22%；而女性職業教育群體的對數收入比普通教育群體高 0.30，約為 30%。由此可見，女性職業教育回報率顯著高於男性，充分說明教育類型對性別收入差距

具有明顯的調節作用。

為了更直觀地展示性別與教育類型對收入的影響，繪製了邊際效應圖，如圖 1 所示。從圖中可以清晰看到不同性別和教育類型組合下的對數收入情況。這一圖表為理解性別與教育類型在收入回報上的差異提供了直觀依據，有助於更深入地分析相關現象。結果如下：

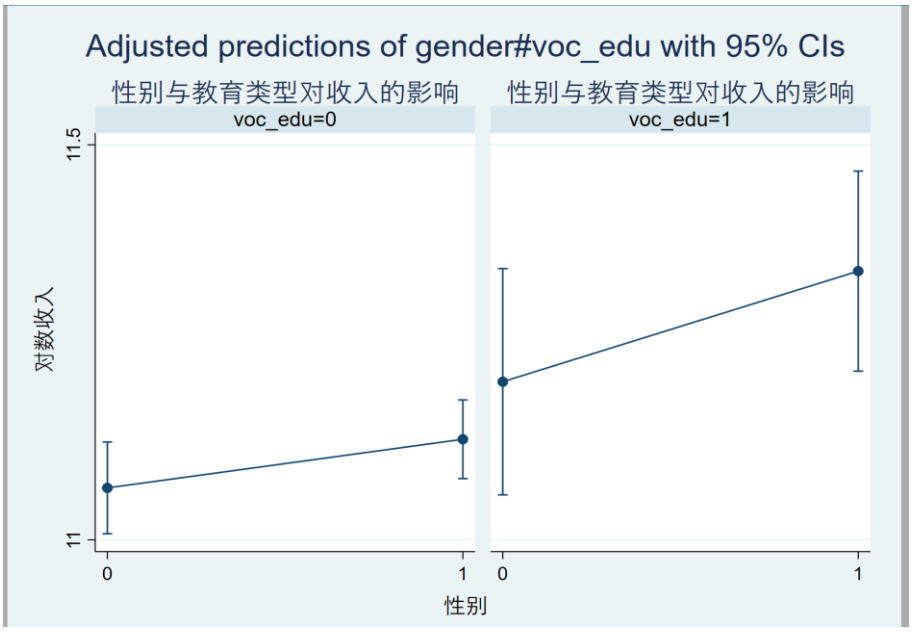


圖 4-1 邊際效應圖

職業教育對應的柱狀圖高於普通教育，就直觀地表明職業教育的收入回報更高；同理，女性對應的柱狀圖高於男性，說明女性的收入回報更高。綜合來看，性別對收入有影響，男性對數收入相對較高，但女性在職業教育中的表現突出，在一定程度上縮小了性別收入差距。性別與教育年限的交互項對收入的影響在統計上微顯著，表明兩者交互效應對收入影響為顯著，體現了職業教育在提升個體收入方面的重要作用。

本章通過全樣本混合 OLS 回歸、分性別回歸以及構建拓展模型，深入分析了“普職分流”教育收益率的性別差異。發現職業教育在提升收入方面具有明顯優勢，且女性在職業教育中的收入回報相對男性更為突出。實證結果顯示如下：

其一，職業教育回報率的性別差異。明瑟收入拓展模型顯示，交互項係數顯著為正，表明職業教育的收入回報在女性群體中更高。這一結果打破傳統認知，反映出當下社會經濟環境裡，女性在教育和就業領域競爭力逐步提升，職業教育為女性實現更高收入提供了有效路徑。

其二，教育年限增收效應的性別差異。男女性教育年限係數比較，發現女性每增加 1 年教育，收入增長率比男性高出約 1.23 個百分點。這既符合人力資本理論中“邊際回報遞增”假說，也暗示女性可借助教育突破職業限制，獲取更高收入，同時可能存在勞動力市

場對高學歷女性的補償溢價。

最後，教育類型與性別交互效應。性別與教育類型交互項對收入影響顯著，表明教育類型對性別收入差距有明顯調節作用。在職業教育中，女性的收入提升幅度大於男性，而在普通教育中，這種性別間的收入差距並不如此明顯。這意味著不同教育路徑對男女收入的影響存在差異，職業教育在縮小性別收入差距上有獨特作用。

在本研究中，通過構建一系列模型逐步驗證研究假設，並進行穩健性交叉驗證，以此深入剖析“普職分流”背景下教育收益率的性別差異。全樣本混合 OLS 回歸從整體層面估計教育收益率，回答“教育是否顯著提升收入”這一關鍵問題；分性別回歸聚焦於檢驗性別異質性，判斷“教育回報是否存在性別差距”；拓展模型則進一步揭示性別與教育之間的交互機制，探究“性別如何調節教育對收入的影響”。同時，通過比較不同模型的係數穩定性，以及判斷拓展模型中交互項的顯著性，進行穩健性交叉驗證，以確保研究結果的可靠性。

實證結果分析可知，目前社會勞動力市場中性別差異化仍然存在，接下來，主要探討性別差異機制成因，字勞動力市場上性別差異如何通過教育選擇影響人力資本積累的差異，為後續政策建議提供理論基礎，通過政策引導資源優化、消除制度性障礙，可進一步釋放職業教育的平等化潛力，為實現社會公平與經濟效率的雙重目標提供支撐。

第一，隱性勞動力市場歧視持續存在。勞動力市場存在隱性性別歧視，即在男女教育水準相同條件的前提下，女性薪酬增幅也可能受限。企業對於女性的預判往往示基於群體特徵而非個體能力，認為生育、家庭責任等因素會降低工作投入，從而在招聘條件、薪酬福利和晉升中設置隱性限制。比如，社會普遍認為男性更適合技術性崗位，這種偏見導致女性在同等條件下收入比男性低 10%-15%。在職業發展過程中，女性晉升機會也相對較少，像製造業領域，男性職校生晉升速度比女性高出約 30%，使得女性教育投資的回報難以充分實現，影響教育收益率。

第二，職業選擇性別隔離是教育收益率差異的另一個核心成因。職業院校的課程設置與實訓資源配置是一把傾斜的天平，存在明顯性別傾向。技術類職業的實訓設備投入占比高達 65%，而護理、文秘等服務類專業長期依賴理論教學。男性多集中於機械類等技術專業，女性則偏向服務類專業。這種差異延伸至勞動力市場，造成行業隔離。服務類行業普遍薪資水準低於技術類行業，導致不同性別職業發展路徑和收入水準不同。例如，機電、數控等技術類專業女生不足 20%，而幼教、護理等服務類專業女生超 90%，直接導致職業教育收益率的性別差異。

第三，教育資源配置不均。在教育階段，男女獲取的教育資源存在差異。職業教育中，

與高薪技術行業相關的教育資源可能更多向男性傾斜，女性難以獲得同等優質資源。在一些技術培訓課程中，男性參與比例更高，這使得男性在進入高收入行業和崗位時更具競爭力，加劇教育收益率的性別差異。

“普職分流”背景下性別差異成因主要是勞動力市場歧視存在隱形歧視、職業選擇性別隔離、教育資源配置不均三大主要原因，為了進一步縮小性別差異影響，將從個人、學校、社會三個方法量化普職教育對個人收入的具體影響，從而蓋上性別差異性。

在職業教育階段，加強對學生的職業指導，引導學生根據自身興趣和能力選擇專業，而非受性別刻板印象的影響。學校和教育機構可以開設職業探索課程，組織學生參觀不同行業的企業，瞭解各種職業的實際工作內容和發展前景。此外，政府可以出臺相關政策，鼓勵企業為女性提供進入傳統男性主導行業的培訓和就業機會，對積極推動性別平等就業的企業給予稅收優惠等獎勵。加大對職業教育的投入，特別是對女性占比較低的技術類專業的投入，改善教學設施和師資力量，為女性提供更多優質的教育資源。設立專項獎學金和助學金，鼓勵女性報考技術類專業，提高女性在這些領域的參與度。同時，加強對職業教育課程的改革，使其更加貼近市場需求，提高學生的就業競爭力。

縮小教育收益率的性別差異需個人、學校、社會三維的共同行動，個人需突破傳統認知桎梏，學校要重新構建教育資源配置規則，社會則須以制度剛性約束歧視行為。唯有通過政策強制力打破“歧視 - 隔離 - 資源剝奪”的惡性循環，才能真正釋放職業教育的平等化潛力，實現人力資本優化配置與社會公平的雙贏。

本研究聚焦於“普職分流”教育收益率性別差異，運用明瑟收入模型進行教育收益率估算。在研究過程中，採用了多種資料預處理方法對關鍵變數進行合理定義與編碼，確保資料的品質和分析的準確性。同時，構建了明瑟收入方程及其拓展模型，將性別、教育類型作為引數，收入作為因變數，以此預測不同教育類型的教育收益率。

通過一系列的實證分析，研究發現職業教育對個體收入存在正向邊際效應，不過其經濟回報的顯著性有待進一步提升。在分性別回歸分析中，職業教育回報尚未呈現出明顯的性別結構性分化，但在教育年限增收效應和工作經驗對收入的影響方面，存在顯著的性別異質性。並且，性別與教育年限的交互作用也對教育的增收效應產生影響。

參考文獻

1) 邢蓓蓓 , 楊現民 , 李勤生 . 教育大資料的來源與採集技術 [J]. 現代教育技術 , 2016, 26(08): 14–21.

2) 謝賢君 , 鬱俊莉 . 大資料如何影響企業全要素生產率——來自《促進大資料發展行動綱要》實施的准自然試驗 [J]. 當代經濟管理 , 2023, 45(08): 22–32.

3) 黃偉 . 大資料技術在商業銀行金融風險管理的應用研究 [D]. 廈門大學 , 2022.

4) 王欣 . 大資料背景下銀行業風險管理與內部控制研究 [D]. 中南財經政法大學 , 2021.

5) 錢才銀 . 互聯網金融風險評估及其防範研究 [D]. 湖北民族學院 , 2017.

6) 於泓飛 . 金融風險管理中大資料的運用 [J]. 老字型大小品牌行銷 , 2023(01): 74–76.

7) 趙佳欣 , 姚舒炅 , 王俊銘 . 大資料視角下企業金融風險管理研究 [J]. 商展經濟 , 2024(24): 101–104.

8) 唐國豪 , 朱琳 , 廖存非 , 等 . 基於自編碼機器學習的資產定價研究——中國股票市場的金融大資料分析視角 [J]. 管理科學學報 , 2024, 27(09): 82–97.

9) 黃金波 , 尤亦玲 , 李仲飛 . 基於前瞻資訊的廣義風險與收益率預測 [J]. 管理科學學報 , 2024, 27(03): 91–111.

10) 謝平 , 鄒傳偉 , 劉海二 . 互聯網金融的基礎理論 [J]. 金融研究 , 2015(08): 1–12.

11) 魏宇 , 李霞飛 , 梁超 . 公共衛生事件下我國風險與避險資產溢出效應——基於收益與風險分析的視角 [J]. 管理科學學報 , 2024, 27(06): 127–148.

12) 蒯婷婷 . 大資料背景下國家審計在防範金融風險中的作用研究 [D]. 哈爾濱商業大學 , 2022.

13) 苗子清 . 基於大資料方法的中國系統性金融風險監測和預警研究 [D]. 中國社會科學院大學 , 2022.

14) Yue H, Liao H J, Li D, et al. Enterprise Financial Risk Management Using Information Fusion Technology and Big Data Mining[J]. Wireless Communications and Mobile Computing, 2021, Article ID 9980163.

15) Cerchiello P, Giudici P. Big Data Analysis for Financial Risk Management[J]. Journal of Big Data, 2016, 3(1): 1–18.

16) Wei L, Miao X, Jing H, et al. Bank Risk Aggregation Based on the Triple Perspectives of Bank Managers, Credit Raters, and Financial Analysts[J]. Finance Research Letters, 2022, 46: 102–114.

17) Arrow K J. The Economic Implications of Learning by Doing[J]. Review of Economic Studies, 1962, 29(3): 155–173.
18) Barro R J. Economic Growth in a Cross Section of Countries[J]. Quarterly Journal of Economics, 1991, 106(2): 407–443.
19) Sachs J D, Warner A. Economic Reform and the Process of Global Integration[J]. Brookings Papers on Economic Activity, 1995(1): 1–118.
20) Krugman P. Scale Economies, Product Differentiation, and the Pattern of Trade[J]. American Economic Review, 1980, 70(5): 950–959.
21) Chenery H B, Syrquin M. Patterns of Development of the World Economy[R]. World Bank Staff Working Paper, No. 412, 1975.
22) 柳鍵 , 李勝勝 , 周雲蕾 . 我國 CPI 與經濟增長、工農業產品價格相關性研究——基於因數分析和多元線性回歸模型 [J]. 價格理論與實踐 , 2017(01): 91–94. DOI:10.19851/j.cnki.cn11-1010/f.2017.01.022.
23) 王璐瑤 . 基於動態空間面板模型的廣東省經濟影響因素的研究 [D]. 暨南大學 , 2021. DOI:10.27167/d.cnki.gjinu.2021.001309.
24) 陳雷 , 邢宏珍 . 基於多元線性回歸的十堰市經濟發展影響因素分析 [J]. 湖北工業職業技術學院學報 , 2019, 32(03): 41–44.
25) 張禹 , 朱勇振 , 劉仁貴 , 等 . 區域經濟增長影響因素分析——以河南省為例 [J]. 內蒙古科技與經濟 , 2022(14): 48–51.
26) 鄧愛民 , 李鵬 . 中國旅遊經濟影響因素分析與實證研究 [J]. 宏觀經濟研究 , 2022(03): 106–115+137. DOI:10.16304/j.cnki.11-3952/f.2022.03.003.
27) 馮美君 . 中國數字產業的經濟增長動力分析及影響因素研究 [D]. 東北財經大學 , 2022. DOI:10.27006/d.cnki.gdbcu.2022.001019.
28) 郭子誠 , 楊燦 , 張潔 , 等 . 人工智慧驅動下的鄉村經濟發展視覺化探索 [J]. 福建電腦 , 2025, 41(03): 40–44. DOI:10.16707/j.cnki.fjpc.2025.03.009.
29) 劉旋 , 區路騏 , 操傑睿 . 淺析運用多元線性回歸模型分析影響稅收收入的經濟因素 [J]. 商訊 , 2021(36): 143–145.
30) 丁姝 , 張霞偉 . 對中國 GDP 增長的主要因素分析 [J]. 現代經濟資訊 , 2015(15): 15.