1. 概述

本节以 SPSS 软件演示多元线性回归模型的操作流程，一元线性回归的操作与之相同。案例数据是高铁站点周边的城镇化，SPSS 中的数据文件为“HST_urbanization.sav”，附件中还提供了相应的 csv 文件。

2. 多元线性回归的基本操作

在 SPSS 打开“HST_urbanization.sav”，如图 1 所示。

No	City	Area	D1	D2	Invest	Type	D3	GDP	Revenue	Population
1	北京	201.06	7.60	9.30	5493.50	1	5.00	14113.6	2353.93	1961.20
2	上海	194.65	10.30	2.10	5317.67	1	15.20	17166.0	2873.58	2302.66
3	济南	119.38	13.50	26.20	1987.44	2	20.60	3910.53	266.13	681.40
4	南京	188.24	11.70	26.80	3306.05	2	10.10	5130.65	518.80	800.76
5	泰安	72.63	5.90	78.20	1270.46	3	5.70	2051.68	116.95	549.42
6	苏州	56.80	16.00	21.90	3617.82	3	14.80	9228.91	900.55	1046.60
7	天津	133.96	14.80	16.60	6511.42	1	8.80	9224.46	1068.81	1293.82
8	徐州	71.38	9.40	35.60	2049.26	3	7.20	2942.14	222.16	858.05
9	蚌埠	32.30	8.00	125.70	528.73	3	5.30	636.89	42.90	316.45
10	廊坊	92.11	4.00	40.60	909.03	3	3.80	1351.10	105.86	435.88

图 1 京沪高铁站点周边城镇化数据

在菜单栏中选择“Analyze → Regression → Linear”，打开线性回归对话框。将“高铁站周边的建筑成区面积 [Area]”选入“Dependent”，作为模型的因变量；将“距离市中心 [D1]”、“距离最近机场 [D2]”、“固定资产投资 [Invest]”选入“Independent(s)”，作为模型的自变量；确认“Method”下拉菜单选中“Enter”，强制这些自变量进入模型，如图 2 所示。

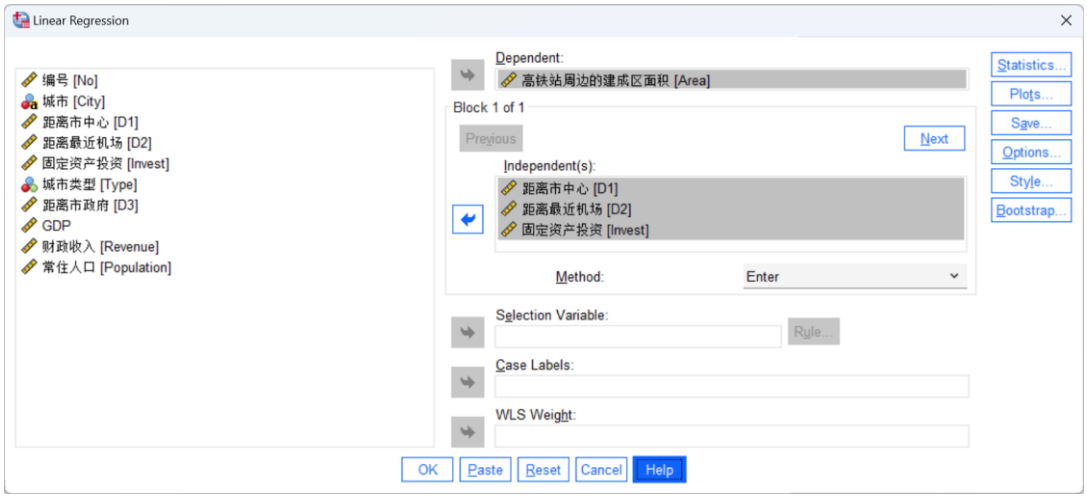


图 2 多元线性回归主对话框设置

点击“OK”运行分析，SPSS 的结果窗口中将依次报告以下结果。

首先是自变量的进入和剔除情况，如图 3。可以看到，SPSS 只估计了 1 个模型；“Variables Entered”中包括了“固定资产投资”、“距离市中心”、“距离最近机场”，即这 3 个变量均进入了这个模型；“Variables Removed”中为空，没有变量被剔除；“Method”为“Enter”，表明 3 个变量进入模型的方式是强制进入。

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	固定资产投资, 距离市中心, 距离最近机场[b]	.	Enter

a. Dependent Variable: 高铁站周边的建成区面积

b. All requested variables entered.

图 3 多元线性回归结果：自变量的进入和剔除情况

然后是模型概况，如图 4。SPSS 报告了模型的复相关系数（R）、样本决定系数（R Square）、修正后的样本决定系数（Adjusted R Square）、估计值的标准误（Std. Error of the Estimate）。我们主要关注样本决定系数和修正后样本决定系数。

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.871[a]	.759	.707	33.26394

a. Predictors: (Constant), 固定资产投资, 距离市中心, 距离最近机场

图 4 多元线性回归模型结果：模型概况

接下来报告的是线性回归的 F 检验结果，即方差分析表，如图 5 所示。

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	48735.282	3	16245.094	14.682	<.001[b]
	Residual	15490.855	14	1106.490		
	Total	64226.137	17			

a. Dependent Variable: 高铁站周边的建成区面积

b. Predictors: (Constant), 固定资产投资, 距离市中心, 距离最近机场

图 5 多元线性回归模型结果：方差分析表

然后，SPSS 报告了最为关键的系数估计和 T 检验结果，如图 6 所示。

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	173.626	32.881		5.280	<.001
	距离市中心	-6.569	2.087	-.441	-3.147	.007
	距离最近机场	-.896	.311	-.539	-2.885	.012
	固定资产投资	.014	.006	.433	2.232	.042

a. Dependent Variable: 高铁站周边的建成区面积

图 6 多元线性回归模型结果：系数估计和 T 检验

3. 虚拟变量

数据中的“Type”是一个分类变量，有“直辖市”、“省会城市”、“一般城市”三个分类，我们为其生成虚拟变量。由于 SPSS 只能为分类变量生成虚拟变量，因此需要首先在 SPSS 的变量视图中确认“Type”变量的测度方式（“Measures”）为无序分类变量（“Nominal”），如图 7 所示。

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
No	Numeric	8	0	编号	None	None	3	Right	Scale	Input
City	String	24	0	城市	None	None	4	Left	Nominal	Input
Area	Numeric	8	2	高铁站周边的建...	None	None	5	Right	Scale	Input
D1	Numeric	8	2	距离市中心	None	None	4	Right	Scale	Input
D2	Numeric	8	2	距离最近机场	None	None	4	Right	Scale	Input
Invest	Numeric	8	2	固定资产投资	None	None	6	Right	Scale	Input
Type	Numeric	8	0	城市类型	{1, 直辖市}...	None	7	Right	Nominal	Input
D3	Numeric	8	2	距离市政府	None	None	5	Right	Scale	Input
GDP	Numeric	8	2		None	None	5	Right	Ordinal	Input
Revenue	Numeric	8	2	财政收入	None	None	6	Right	Nominal	Input

图 7 设置和确认 Type 为无序分类变量

在菜单栏中选择“Transform → Create Dummy Variables”，进入创建虚拟变量对话框。在“Create Dummy Variables for”中选入“城市类型 [Type]”，确认下方的“Create main-effect dummies”处于选中状态，然后在“Root Names (One Per Selected Variable)”中输入虚拟变量前缀的名称，这里为“Type”，如图 8。

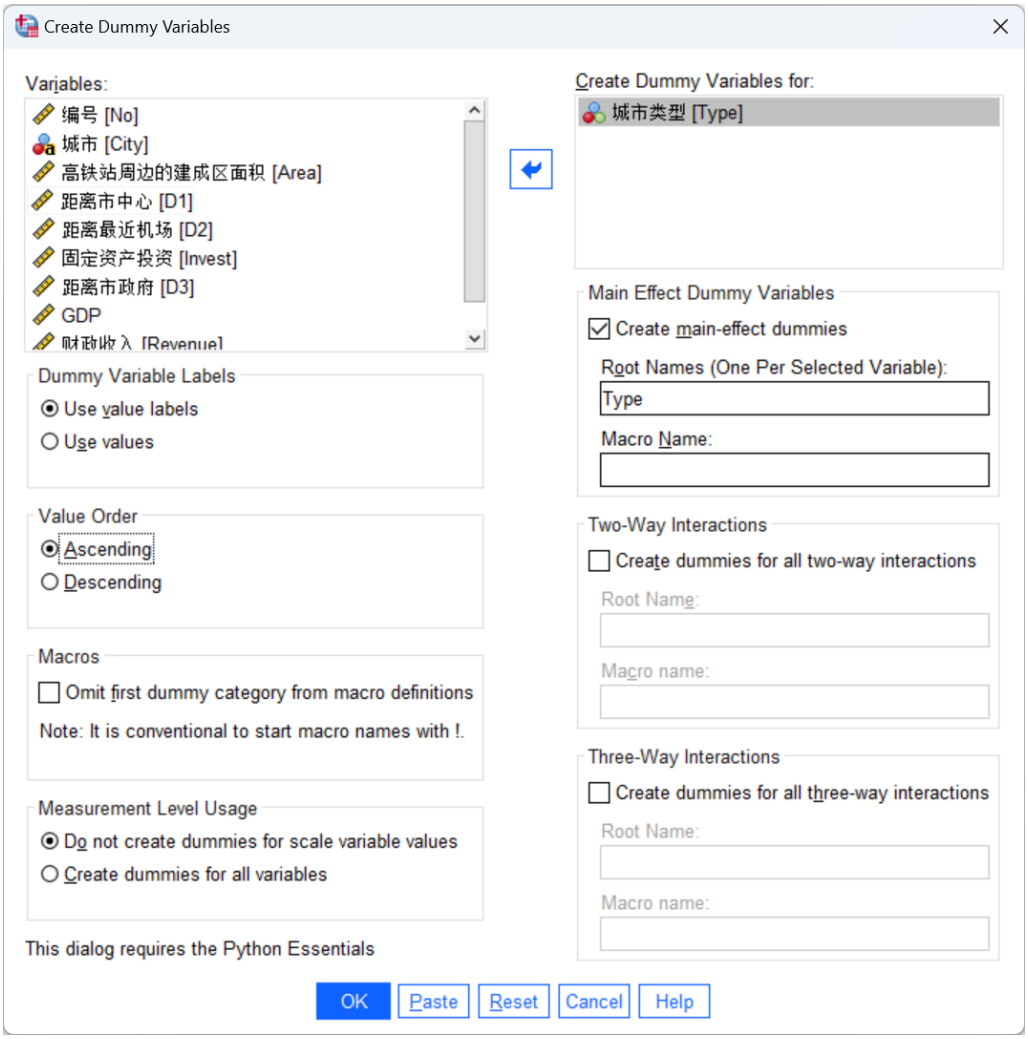


图 8 创建虚拟变量的设置

点击“OK”运行分析，SPSS 的结果窗口中显示如图 9 所示的信息，提示已创建了 3 个变量：“Type_1”对应于“Type=直辖市”，“Type_2”对应于“Type=省会城市”，“Type_3”对应于“Type=一般城市”。

Variable Creation

	Label
Type_1	Type=直辖市
Type_2	Type=省会城市
Type_3	Type=一般城市

图 9 虚拟变量创建结果

回到数据视图中，也可以看到新创建的 3 个虚拟变量已经被加入到数据的最后三列中，如图 10 所示。

No	City	Area	D1	D2	Invest	Type	D3	GDP	Revenue	Population	Type_1	Type_2	Type_3
1	北京	201.06	7.60	9.30	5493.50	直辖市	5.00	14113...	2353.93	1961.20	1.00	.00	.00
2	上海	194.65	10.30	2.10	5317.67	直辖市	15.20	17165...	2873.58	2302.66	1.00	.00	.00
3	济南	119.38	13.50	26.20	1987.44	省会城市	20.60	3910.53	266.13	681.40	.00	1.00	.00
4	南京	188.24	11.70	26.80	3306.05	省会城市	10.10	5130.65	518.80	800.76	.00	1.00	.00
5	泰安	72.63	5.90	78.20	1270.46	一般城市	5.70	2051.68	116.95	549.42	.00	.00	1.00
6	苏州	56.80	16.00	21.90	3617.82	一般城市	14.80	9228.91	900.55	1046.60	.00	.00	1.00
7	天津	133.96	14.80	16.60	6511.42	直辖市	8.80	9224.46	1068.81	1293.82	1.00	.00	.00
8	徐州	71.38	9.40	35.60	2049.26	一般城市	7.20	2942.14	222.16	858.05	.00	.00	1.00
9	蚌埠	32.30	8.00	125.70	528.73	一般城市	5.30	636.89	42.90	316.45	.00	.00	1.00
10	廊坊	92.11	4.00	40.60	909.03	一般城市	3.80	1351.10	105.86	435.88	.00	.00	1.00
11	沧州	32.42	8.20	104.10	1448.09	一般城市	6.30	2203.12	91.30	713.41	.00	.00	1.00
12	德州	12.06	12.70	91.50	1140.59	一般城市	9.50	1657.82	72.91	556.82	.00	.00	1.00
13	曲阜	16.96	8.90	71.60	107.24	一般城市	7.40	235.29	10.77	64.05	.00	.00	1.00

图 10 在数据视图中查看新创建的虚拟变量

为了避免完全共线性，对于拥有 N 个类别的分类变量，我们只需要使用 N-1 个虚拟变量。这里，我们只在线性回归中使用“Type_1”和“Type_2”两个虚拟变量。重新打开线性回归对话框，仍然将“高铁站周边的建筑成区面积 [Area]”选入“Dependent”，作为模型的因变量；在自变量方面，将“距离市中心 [D1]”、“距离最近机场 [D2]”、“Type=直辖市 [Type_1]”、“Type=省会城市 [Type_2]”选入“Independent(s)”；确认“Method”为“Enter”，如图 11 所示。

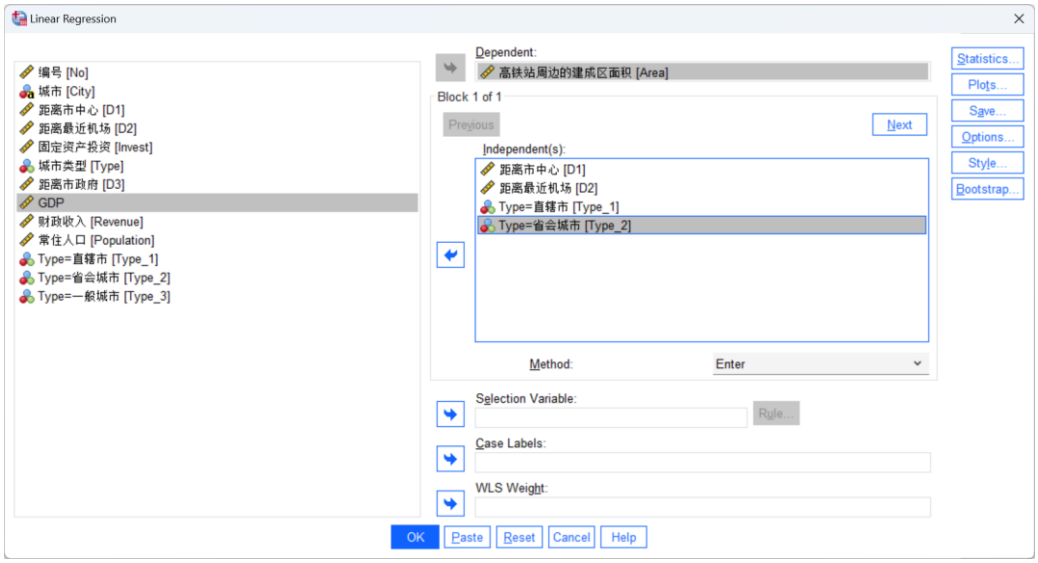


图 11 带虚拟变量的多元线性回归主对话框设置

点击“OK”运行分析，此时的模型拟合优度如图 12 所示。

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.937[a]	.878	.840	24.56719

a. Predictors: (Constant), Type=省会城市, Type=直辖市, 距离市中心, 距离最近机场

图 12 带虚拟变量的多元线性回归结果：模型概况

模型的参数估计与 T 检验结果如图 13 所示。

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	173.873	22.315		7.792	<.001
	距离市中心	-5.640	1.500	-.379	-3.759	.002
	距离最近机场	-.872	.201	-.524	-4.338	<.001
	Type=直辖市	72.299	18.805	.451	3.845	.002
	Type=省会城市	74.112	20.039	.390	3.698	.003

a. Dependent Variable: 高铁站周边的建成区面积

图 13 带虚拟变量的多元线性回归模型结果：参数估计和 T 检验

4. 多重共线性

重新打开线性回归对话框，仍然将“高铁站周边的建筑成区面积 [Area]”选入“Dependent”，作为模型的因变量；在自变量方面，将“距离市中心 [D1]”、“距离最近机场 [D2]”、“距离市政府 [D3]”、“固定资产投资 [Invest]”、“Type=直辖市 [Type_1]”、“Type=省会城市 [Type_2]”、“GDP”、“常住人口 [Population]”、“财政收入 [Revenue]”选入“Independent(s)”；确认“Method”为“Enter”，如图 14 所示。

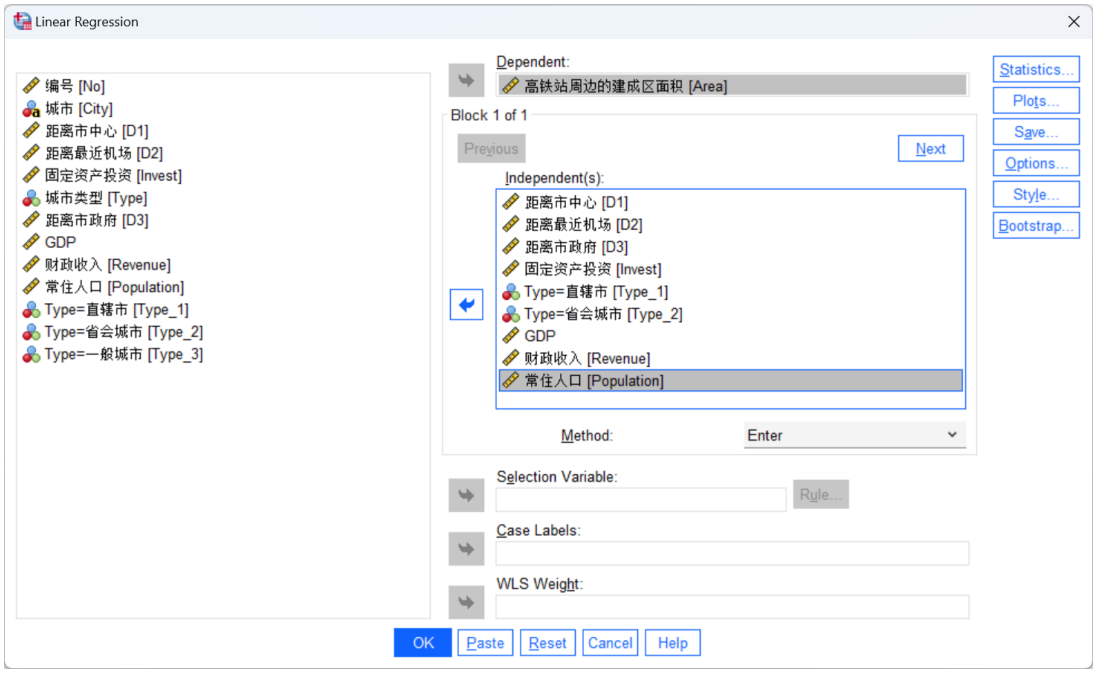


图 14 可能存在多重共线性问题的多元线性回归模型设置

点击“Statistics”，进入统计量子对话框，勾选“Collinearity diagnostics”，要求 SPSS 报告模型的多重共线性诊断结果，如图 15 所示。

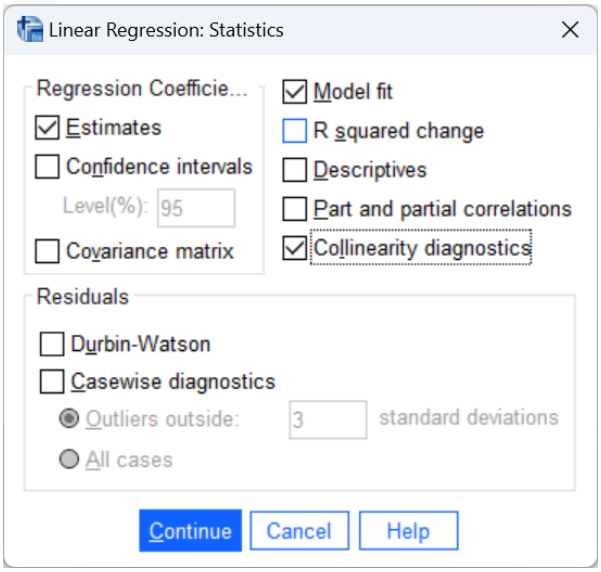


图 15 多重共线性诊断设置

点击“Continue”回到主对话框，点击“OK”运行分析，这里我们只关注结果窗口中新增加的内容。注意到此时的参数估计结果如图 16 所示，在最右侧增加了两列——容忍度“Tolerance”和方差膨胀因子“VIF”。

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	176.900	24.382		7.255	<.001		
	距离市中心	-3.283	2.184	-.221	-1.503	.171	.347	2.878
	距离最近机场	-.729	.280	-.438	-2.608	.031	.265	3.778
	距离市政府	-5.383	2.516	-.420	-2.139	.065	.194	5.154
	固定资产投资	-.006	.020	-.197	-.320	.757	.020	50.680
	Type=直辖市	42.610	61.415	.266	.694	.507	.051	19.632
	Type=省会城市	109.450	25.148	.576	4.352	.002	.427	2.341
	GDP	.013	.016	1.000	.819	.436	.005	199.404
	财政收入	-.008	.085	-.102	-.091	.930	.006	169.679
	常住人口	-.036	.044	-.346	-.827	.432	.043	23.437

a. Dependent Variable: 高铁站周边的建成区面积

图 16 多重共线性诊断结果：容忍度和 VIF

除了这两个指标以外，SPSS 还提供了基于特征值与变异构成的多重共线性诊断指标，如图 17 所示。与 VIF 相比，该指标不够直观，因此实务中并不常用的指标，我们不要求掌握。以下说明供感兴趣的同学了解。

图 17 所示的结果中有 10 个维度（“Dimension”），每个维度对应一个特征值（“Eigenvalue”）和一个条件指数（“Condition Index”）。特征值从上到下降序排列，条件指数是最大特征值与当前维度特征值的比值的算术平方根。例如，最大特征值为 6.787，第 5 个维度的特征值为 0.129，则第 5 个维度的条件指数为

$\sqrt{6.787/0.129} = 7.267$。我们应当主要关注最后几个维度的特征值和条件指数，如果特征值很小，相应的条件指数很大（>30），则提示存在多重共线性。该表中，最后一个维度的条件指数达到了 71.391，提示多重共线性。同时，图 17 的结果中还报告了各自变量（包括常数项）的变异构成（“Variance Proportions”），其意义是每个解释变量能够被各维度解释的比例。如果某个维度在两个或多个解释变量上都具有较高的变异构成（>0.5），则提示这些解释变量之间存在多重共线性。

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions									
				(Constant)	距离市中心	距离最近机场	距离市政府	固定资产投资	Type=直辖市	Type=省会城市	GDP	财政收入	常住人口
1	1	6.787	1.000	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	2	1.769	1.959	.00	.00	.01	.00	.00	.01	.02	.00	.00	.00
	3	.935	2.694	.00	.00	.02	.00	.00	.00	.29	.00	.00	.00
	4	.244	5.275	.00	.02	.12	.03	.00	.03	.24	.00	.00	.00
	5	.129	7.267	.00	.05	.02	.01	.01	.07	.00	.00	.01	.01
	6	.062	10.475	.10	.00	.03	.27	.02	.15	.02	.00	.00	.01
	7	.035	13.925	.58	.36	.17	.13	.00	.04	.06	.00	.00	.00
	8	.027	15.757	.27	.28	.14	.15	.06	.03	.10	.00	.02	.09
	9	.011	24.560	.01	.18	.42	.01	.09	.04	.00	.05	.00	.76
	10	.001	71.391	.03	.09	.07	.39	.82	.63	.28	.95	.96	.13

a. Dependent Variable: 高铁站周边的建成区面积

图 17 多重共线性诊断结果：特征值和变异构成（了解）

5. 逐步回归

重新打开线性回归对话框，不改变自变量与因变量的设定，只是将“Method”下拉菜单由“Enter”更改为“Stepwise”，使用逐步回归法，如图 18 所示。

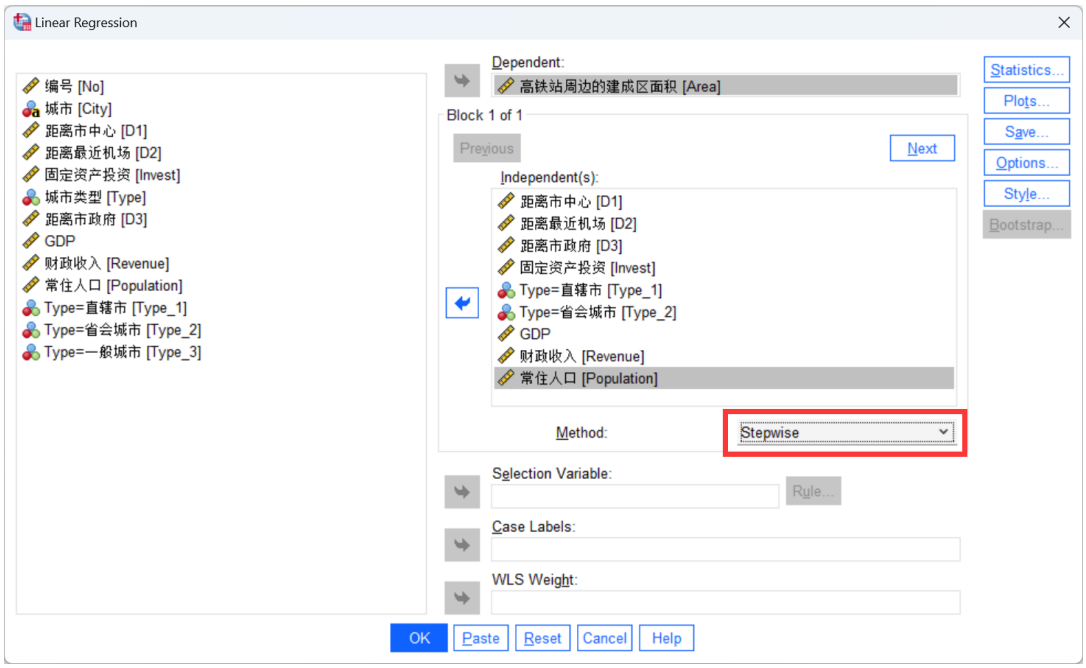


图 18 逐步回归设置

点击“Statistics”，确认统计量子对话框中的“Collinearity diagnostics”仍处

于选中状态，点击“Continue”回到主对话框，然后点击“OK”运行分析。此时，SPSS 结果窗口中汇报的内容虽然相似，但每一个表格中不再只有 1 个模型，而是包括了多个模型，体现了逐步筛选自变量的过程。

首先是变量进出的情况（“Variables Entered/Removed”），如图 19 所示。从“Model”列中可以看到，SPSS 共估计了 4 个模型。模型 1 中，进入的变量（“Variables Entered”）为“距离最近机场”，被剔除的变量（“Variables Removed”）为空，变量进出的方法是“Stepwise”，即逐步回归法。类似地，模型 2 中进一步纳入了“距离市中心”，模型 3 中进一步纳入了“固定资产投资”，模型 4 中进一步纳入了“Type=省会城市”，每一步都没有之前进入的变量被剔除，变量进出的方法均为“Stepwise”。

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	距离最近机场	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	距离市中心	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	固定资产投资	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
4	Type=省会城市	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: 高铁站周边的建成区面积

图 19 逐步回归结果：自变量进入和剔除

接下来是反映每一步模型拟合优度的模型概况，如图 20 所示。下方的注释提示了当前模型中的自变量。

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.743[a]	.553	.525	42.37084
2	.820[b]	.673	.629	37.41996
3	.871[c]	.759	.707	33.26394
4	.924[d]	.855	.810	26.79329

a. Predictors: (Constant), 距离最近机场

b. Predictors: (Constant), 距离最近机场, 距离市中心

c. Predictors: (Constant), 距离最近机场, 距离市中心, 固定资产投资

d. Predictors: (Constant), 距离最近机场, 距离市中心, 固定资产投资, Type=省会城市

图 20 逐步回归模型结果：模型概况

接下来是每一步模型的 F 检验，即方差分析表，如图 21 所示。

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35501.525	1	35501.525	19.775	<.001[b]
	Residual	28724.611	16	1795.288		
	Total	64226.137	17			
2	Regression	43222.337	2	21611.168	15.434	<.001[c]
	Residual	21003.800	15	1400.253		
	Total	64226.137	17			
3	Regression	48735.282	3	16245.094	14.682	<.001[d]
	Residual	15490.855	14	1106.490		
	Total	64226.137	17			
4	Regression	54893.694	4	13723.423	19.117	<.001[e]
	Residual	9332.443	13	717.880		
	Total	64226.137	17			

a. Dependent Variable: 高铁站周边的建成区面积

b. Predictors: (Constant), 距离最近机场

c. Predictors: (Constant), 距离最近机场, 距离市中心

d. Predictors: (Constant), 距离最近机场, 距离市中心, 固定资产投资

e. Predictors: (Constant), 距离最近机场, 距离市中心, 固定资产投资, Type=省会城市

图 21 逐步回归模型结果：方差分析表

接下来是最重要的结果：每一步模型的参数估计、T 检验、多重共线性诊断指标，如图 22 所示。

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	152.773	16.855		9.064	<.001		
	距离最近机场	-1.237	.278	-.743	-4.447	<.001	1.000	1.000
2	(Constant)	215.219	30.476		7.062	<.001		
	距离最近机场	-1.375	.253	-.827	-5.444	<.001	.946	1.057
	距离市中心	-5.308	2.260	-.357	-2.348	.033	.946	1.057
3	(Constant)	173.626	32.881		5.280	<.001		
	距离最近机场	-.896	.311	-.539	-2.885	.012	.494	2.023
	距离市中心	-6.569	2.087	-.441	-3.147	.007	.876	1.141
	固定资产投资	.014	.006	.433	2.232	.042	.458	2.182
4	(Constant)	161.662	26.798		6.033	<.001		
	距离最近机场	-.710	.258	-.427	-2.751	.017	.464	2.154
	距离市中心	-7.480	1.710	-.502	-4.375	<.001	.847	1.180
	固定资产投资	.016	.005	.510	3.218	.007	.446	2.244
	Type=省会城市	61.799	21.100	.325	2.929	.012	.907	1.102

a. Dependent Variable: 高铁站周边的建成区面积

图 22 逐步回归模型结果：参数估计、T 检验、多重共线性

除了纳入的自变量之外，SPSS 还报告了每一步估计中未被纳入当前模型的自变量情况，如图 23 所示，不必特别关注。

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	距离市中心	-.357[b]	-2.348	.033	-.518	.946	1.057	.946
	距离市政府	-.225[b]	-1.214	.244	-.299	.792	1.262	.792
	固定资产投资	.268[b]	1.136	.274	.281	.495	2.022	.495
	Type=直辖市	.349[b]	1.975	.067	.454	.758	1.319	.758
	Type=省会城市	.210[b]	1.248	.231	.307	.952	1.051	.952
	GDP	.342[b]	1.532	.146	.368	.516	1.938	.516
	财政收入	.391[b]	2.000	.064	.459	.616	1.624	.616
	常住人口	.321[b]	1.638	.122	.390	.661	1.513	.661
2	距离市政府	.002[c]	.011	.992	.003	.525	1.905	.525
	固定资产投资	.433[c]	2.232	.042	.512	.458	2.182	.458
	Type=直辖市	.318[c]	2.065	.058	.483	.753	1.328	.714
	Type=省会城市	.266[c]	1.881	.081	.449	.933	1.072	.913
	GDP	.386[c]	2.062	.058	.483	.512	1.952	.509
	财政收入	.360[c]	2.118	.053	.493	.612	1.633	.584
	常住人口	.327[c]	1.967	.069	.465	.661	1.514	.639
3	距离市政府	.043[d]	.228	.823	.063	.520	1.924	.428
	Type=直辖市	.101[d]	.318	.756	.088	.182	5.488	.111
	Type=省会城市	.325[d]	2.929	.012	.631	.907	1.102	.446
	GDP	.155[d]	.474	.643	.130	.172	5.830	.153
	财政收入	.180[d]	.693	.501	.189	.265	3.780	.198
	常住人口	.084[d]	.272	.790	.075	.194	5.143	.135
4	距离市政府	-.211[e]	-1.295	.220	-.350	.401	2.496	.401
	Type=直辖市	.385[e]	1.543	.149	.407	.162	6.173	.106
	GDP	.306[e]	1.195	.255	.326	.166	6.042	.153
	财政收入	.316[e]	1.591	.138	.417	.254	3.944	.198
	常住人口	.134[e]	.543	.597	.155	.193	5.169	.135

a. Dependent Variable: 高铁站周边的建成区面积

b. Predictors in the Model: (Constant), 距离最近机场

c. Predictors in the Model: (Constant), 距离最近机场, 距离市中心

d. Predictors in the Model: (Constant), 距离最近机场, 距离市中心, 固定资产投资

e. Predictors in the Model: (Constant), 距离最近机场, 距离市中心, 固定资产投资, Type=省会城市

图 23 逐步回归模型结果：被剔除的自变量

最后，SPSS 报告了每一步模型基于特征值和变异构成的多重共线性诊断结果，如图 24 所示，解释方法同上。这种方法不如 VIF 指标常用，感兴趣的同学了解即可。

Collinearity Diagnostics[a]

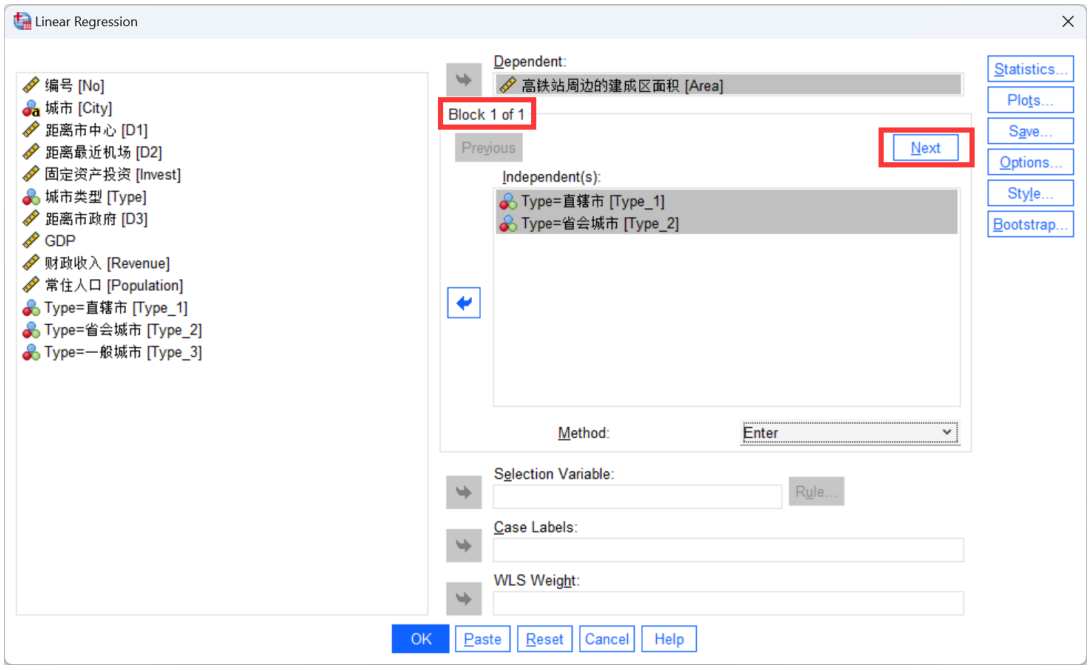
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	距离最近机场	距离市中心	固定资产投资	Type=省会城市
1	1	1.806	1.000	.10	.10			
	2	.194	3.047	.90	.90			
2	1	2.632	1.000	.01	.04	.02		
	2	.317	2.881	.01	.69	.11		
	3	.051	7.202	.98	.27	.87		
3	1	3.213	1.000	.01	.01	.01	.01	
	2	.653	2.218	.00	.14	.00	.13	
	3	.094	5.843	.00	.25	.72	.51	
	4	.040	9.005	.99	.60	.27	.35	
4	1	3.360	1.000	.00	.01	.01	.01	.01
	2	.906	1.926	.00	.03	.00	.00	.70
	3	.608	2.351	.00	.10	.00	.14	.19
	4	.088	6.180	.00	.23	.78	.46	.07
	5	.039	9.307	.99	.63	.21	.39	.02

a. Dependent Variable: 高铁站周边的建成区面积

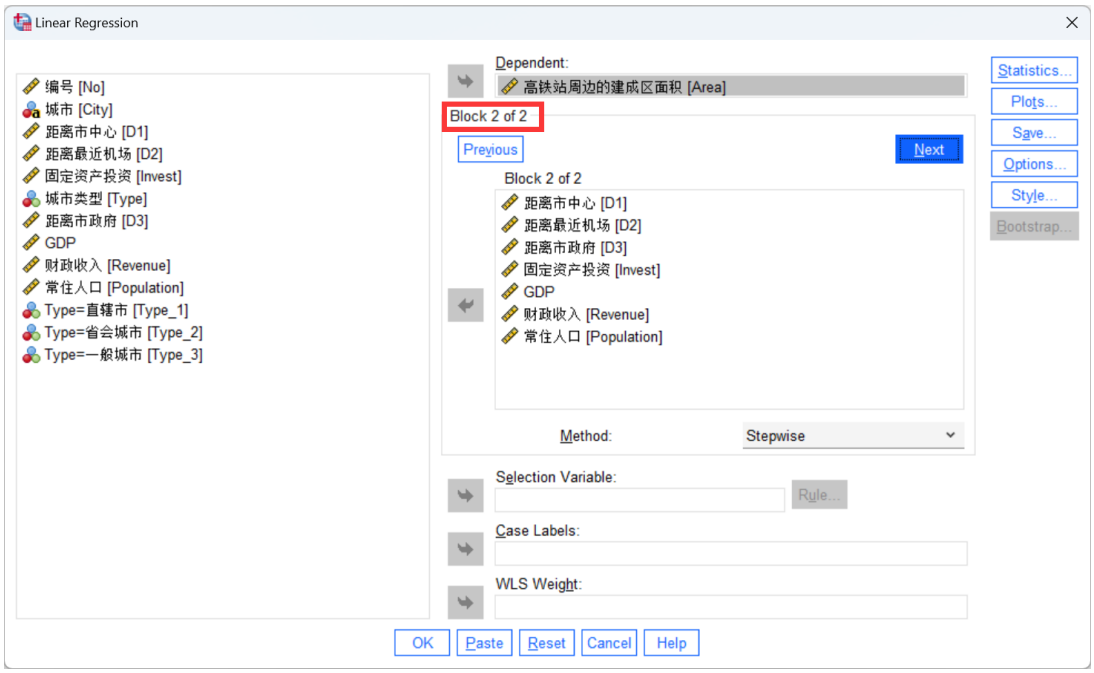
图 24 逐步回归模型结果：基于特征值和变异构成的多重共线性诊断

我们在课程中还提到，有时候需要混合使用不同的自变量进入方式。例如，如果要在逐步回归中保证虚拟变量同进同出，可以先把虚拟变量以“Enter”方式强制进入，然后把其他自变量以“Stepwise”方式逐步筛选进入。下面对这种方式进行演示。

重新打开线性回归对话框，清除“Independent(s)”框中的所有自变量。首先将“Type=直辖市 [Type_1]”、“Type=省会城市 [Type_2]”这两个属于同一分类变量的虚拟变量选入“Independent(s)”框中，并在“Methods”下拉菜单中选择“Enter”，如图 25(a)所示。这两个变量是第一组自变量，因为我们目前只有这一组，所以显示为“Block 1 of 1”。然后，点击“Independent(s)”框右上角的“Next”按钮，进入下一组自变量的设置，把“距离市中心 [D1]”、“距离最近机场 [D2]”等其他自变量选入，并在“Methods”下拉菜单中选择“Stepwise”，如图 25(b)所示。此时，模型中有了两组自变量，而这些要被筛选的自变量是第二组，因此提示为“Block 2 of 2”。如果我们点击“Previous”按钮去回看第一组的 2 个虚拟变量，它们将显示为“Block 1 of 2”。由此，即完成了以不同的自变量进入方式处理多组自变量的设置。



(a) 第一组自变量的进入方式：“Enter”



(b) 第二组自变量的进入方式：“Stepwise”

图 25 混合使用不同的自变量进入方式的线性回归模型设置

点击“OK”运行模型。从图 26 所示的自变量进入/剔除结果可以看到，SPSS 共估计了 3 个模型：模型 1 中，两个虚拟变量被强制进入，然后开始对“Block 2”的自变量进行筛选；模型 2 中，“距离最近机场”被选入模型；模型 3 中，“距离市中心”被选入模型，至此模型筛选结果，其余自变量未被纳入模型。

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Type=省会城市, Type=直辖市[b]	.	Enter
2	距离最近机场	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	距离市中心	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: 高铁站周边的建成区面积

b. All requested variables entered.

图 26 混合使用不同的自变量进入方式的线性回归模型结果：自变量的进入/剔除

图 27 是该模型的参数估计、T 检验、多重共线性诊断结果。可以看到，最终结果（模型 3）保证了虚拟变量的同进同出，并相应调整了其他待筛选自变量的纳入情况。

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	63.528	11.232		5.656	<.001		
	Type=直辖市	113.030	25.938	.705	4.358	<.001	.975	1.026
	Type=省会城市	90.285	30.759	.475	2.935	.010	.975	1.026
2	(Constant)	108.237	19.346		5.595	<.001		
	Type=直辖市	75.121	26.158	.469	2.872	.012	.684	1.463
	Type=省会城市	64.884	27.687	.341	2.343	.034	.858	1.165
	距离最近机场	-.729	.275	-.438	-2.652	.019	.667	1.498
3	(Constant)	173.873	22.315		7.792	<.001		
	Type=直辖市	72.299	18.805	.451	3.845	.002	.683	1.465
	Type=省会城市	74.112	20.039	.390	3.698	.003	.845	1.183
	距离最近机场	-.872	.201	-.524	-4.338	<.001	.643	1.554
	距离市中心	-5.640	1.500	-.379	-3.759	.002	.925	1.081

a. Dependent Variable: 高铁站周边的建成区面积

图 27 混合使用不同的自变量进入方式的模型结果：参数估计、T 检验、多重共线性诊断

6. 残差分析与回归诊断

重新打开线性回归对话框，将“高铁站周边的建筑成区面积 [Area]”选入“Dependent”，作为模型的因变量；在自变量方面，将“距离市中心 [D1]”、“距离最近机场 [D2]”、“Type=直辖市 [Dummy_1]”、“Type=省会城市 [Dummy_2]”选入“Independent(s)”；将“Method”选择为“Enter”，要求这 4 个自变量强制进入模型，如图 28 所示。

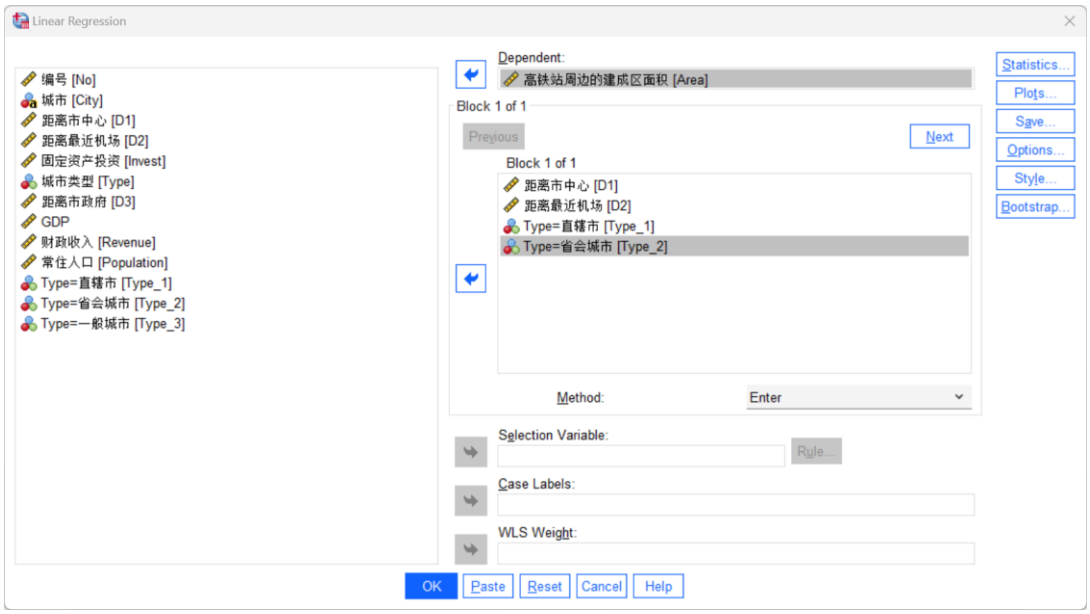


图 28 残差分析的主对话框设置：重新估计一个简单的线性回归模型

点击“Statistics”进入统计量对话框，勾选 “Durbin-Watson”（如图 29 所示），要求报告 Durbin-Watson 统计量，以检验时间序列相关。但是需要注意的是，本例并非时间序列数据，因此这样的检验是无意义的，仅演示操作方法。

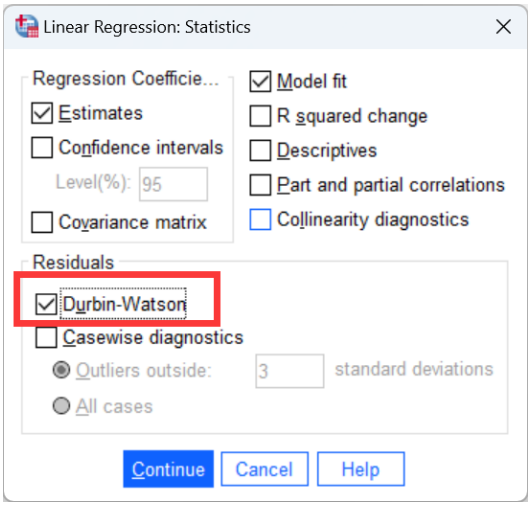


图 29 残差分析设置：Durbin-Watson 统计量

点击“Continue”回到主对话框，点击“Plot”进入绘图子对话框。首先，勾选下方“Standardized Residual Plots”中的“Histogram”和“Normal probability plot”，要求绘制标准化残差的直方图与 P-P 图，以检验非正态问题。然后，在右上方的“Scatter 1 of 1”中，将左框内的“*ZRESID”（标准化残差）选入“Y”，将“*ZPRED”（标准化预测值）选入“X”，绘制预测值~残差散点图（如图 30 所示），以检验异方差问题。

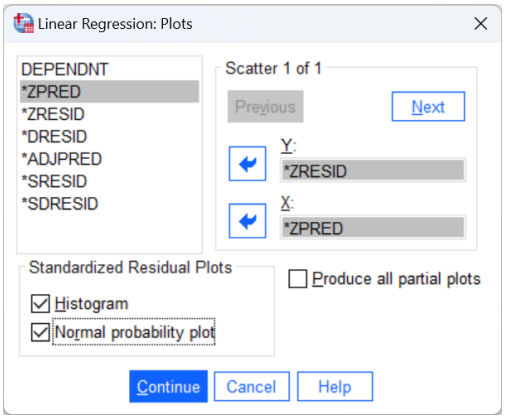


图 30 残差分析设置：绘制残差图

点击“Continue”回到主对话框，点击“Save”进入保存子对话框，该对话框用于将一些结果以变量的形式保存到数据中。我们在右上方的“Residuals”中勾选“Unstandardized”，要求保存非标准化残差，该残差将用于后续分析；然后在“Distances”中勾选“Cook's”，要求保存库克距离，以检查强影响点，如图 31 所示。如果希望保存预测结果，可以勾选左上方“Predicted Values”中的“Unstandardized”，即未标准化预测值。

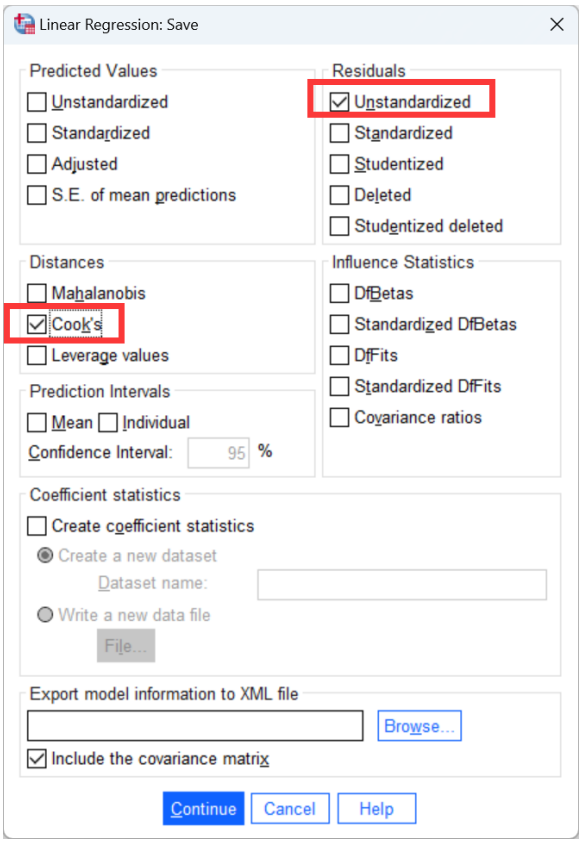


图 31 残差分析设置：保存残差和库克距离

点击“Continue”回到主对话框，点击“OK”运行分析。对于前面出现过的结果，这里就不再赘述，仅关注新增加的结果。

首先，在“Model Summary”中出现了 Durbin-Watson 统计量，如图 32 所示。这里的取值为 2.171，提示没有时间序列相关。但是，本案例不是时间序列数据，因此该统计量没有意义，仅用于演示操作和解读方式。另外，本案例是空间数据，

因此可以通过测度残差的 Moran's I 来检验空间序列相关。SPSS 中没提供相应功能，同学们在以后学习到 ArcGIS 或 Geoda 的相关内容时可以再进行操作。

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.937[a]	.878	.840	24.56719	2.171

a. Predictors: (Constant), Type=省会城市, Type=直辖市, 距离市中心, 距离最近机场

b. Dependent Variable: 高铁站周边的建成区面积

图 32 残差分析结果：模型概况中的 Durbin-Watson 统计量

在参数估计结果的后面，SPSS 报告了一些与残差相关的汇总统计量——最小值、最大值、平均值、标准差、样本量，如图 33 所示。其中，“Predicted Value”是预测值，“Std. Predicted Value”是标准化预测值，“Residual”是非标准化残差，“Std. Residual”是标准化残差，“Cook's Distance”是库克距离。我们重点关注库克距离，可以看到，其最大值为 0.592，因此不存在强影响点。

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	13.5894	195.1982	92.3977	57.58882	18
Std. Predicted Value	-1.368	1.785	.000	1.000	18
Standard Error of Predicted Value	7.239	17.423	12.581	3.152	18
Adjusted Predicted Value	3.6420	191.6955	91.3287	59.23690	18
Residual	-44.27223	35.41373	.00000	21.48339	18
Std. Residual	-1.802	1.442	.000	.874	18
Stud. Residual	-1.886	1.710	.019	1.050	18
Deleted Residual	-59.58979	59.58979	1.06905	31.83038	18
Stud. Deleted Residual	-2.126	1.866	.017	1.124	18
Mahal. Distance	.532	7.606	3.778	2.220	18
Cook's Distance	.000	.592	.109	.185	18
Centered Leverage Value	.031	.447	.222	.131	18

a. Dependent Variable: 高铁站周边的建成区面积

图 33 残差分析结果：残差的汇总统计

我们暂时回到数据视图，如图 34 所示。可以看到，SPSS 已经在最右侧生成了两个新变量：“RES_1”为残差，“COO_1”为库克距离。如果上表中提示库克距离的最大值超过阈值（如 1.0），则我们可以在数据视图中对库克距离进行排序，找到强影响点，再进行进一步的分析。

No	City	Area	D1	D2	Invest	Type	D3	GDP	Revenue	Population	Type_1	Type_2	Type_3	RES_1	COO_1
1	北京	201.06	7.60	9.30	5493.50	直辖市	5.00	14113...	2353.93	1961.20	1.00	.00	.00	5.86372	.01087
2	上海	194.65	10.30	2.10	5317.67	直辖市	15.20	17165...	2873.58	2302.66	1.00	.00	.00	8.40362	.01816
3	济南	119.38	13.50	26.20	1987.44	省会城市	20.60	3910.53	266.13	681.40	.00	1.00	.00	-29.61758	.59185
4	南京	188.24	11.70	26.80	3306.05	省会城市	10.10	5130.65	518.80	800.76	.00	1.00	.00	29.61758	.59185
5	秦安	72.63	5.90	78.20	1270.46	一般城市	5.70	2051.68	116.95	549.42	.00	.00	1.00	.23283	.00000
6	苏州	56.80	16.00	21.90	3617.82	一般城市	14.80	9228.91	900.55	1046.60	.00	.00	1.00	-7.73584	.00990
7	天津	133.96	14.80	16.60	6511.42	直辖市	8.80	9224.46	1068.81	1293.82	1.00	.00	.00	-14.26733	.07450
8	徐州	71.38	9.40	35.60	2049.26	一般城市	7.20	2942.14	222.16	858.05	.00	.00	1.00	-18.43467	.01865
9	蚌埠	32.30	8.00	125.70	528.73	一般城市	5.30	636.89	42.90	316.45	.00	.00	1.00	13.16221	.04613
10	廊坊	92.11	4.00	40.60	909.03	一般城市	3.80	1351.10	105.86	435.88	.00	.00	1.00	-23.79518	.09351
11	沧州	32.42	8.20	104.10	1448.09	一般城市	6.30	2203.12	91.30	713.41	.00	.00	1.00	-4.42101	.00198
12	德州	12.06	12.70	91.50	1140.59	一般城市	9.50	1657.82	72.91	556.82	.00	.00	1.00	-10.38720	.00945
13	曲阜	16.96	8.90	71.60	107.24	一般城市	7.40	235.29	10.77	64.05	.00	.00	1.00	-44.27223	.06762
14	枣庄	44.99	15.40	84.20	720.18	一般城市	3.70	1362.04	76.71	372.93	.00	.00	1.00	31.39824	.13629

图 34 残差分析结果：数据视图中保存每个样本的残差和库克距离

回到输出窗口中。在残差统计指标的下面，SPSS 呈现了标准化残差的直方图与 P-P 图，如图 35 所示，这两张图用于正态性检验。

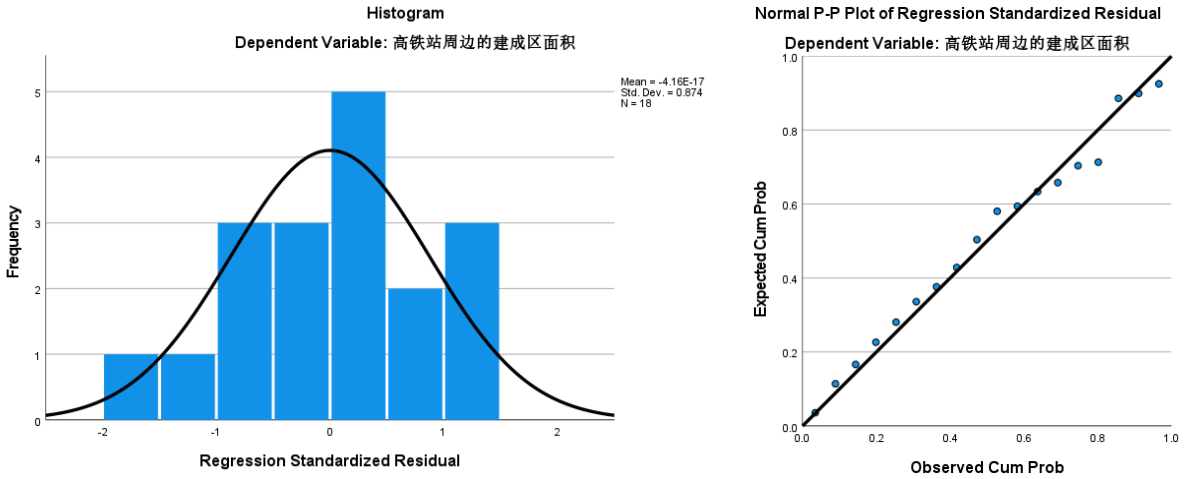


图 35 残差分析结果：正态性检验图

下面，SPSS 呈现了预测值~残差散点图。我们双击该图进入“Chart Editor”窗口，在菜单栏中选择“Options → Hide Grid Lines”，隐藏网格线；然后单击鼠标右键弹出快捷菜单，在菜单中选择“Add Y Axis Reference Line”，在图中增加一条 Y=0 的水平参考线，如图 36 所示，然后关闭“Chart Editor”窗口。

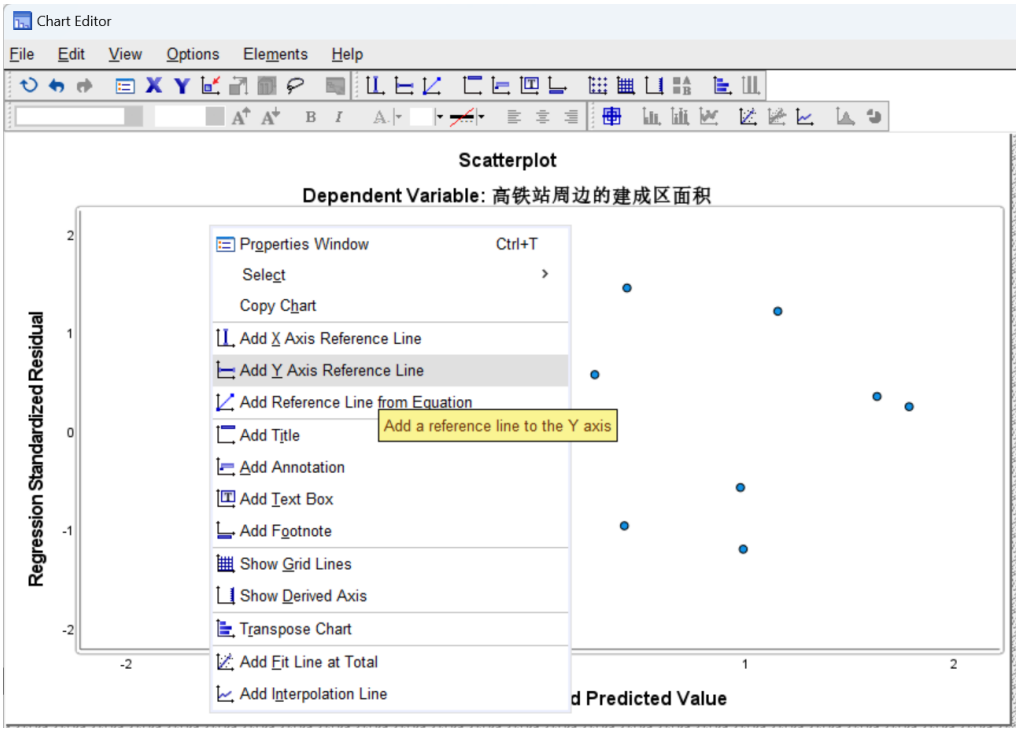


图 36 残差分析：编辑残差图

由此就得到了如图 37 所示的预测值~残差散点图，可用于检查异方差问题。

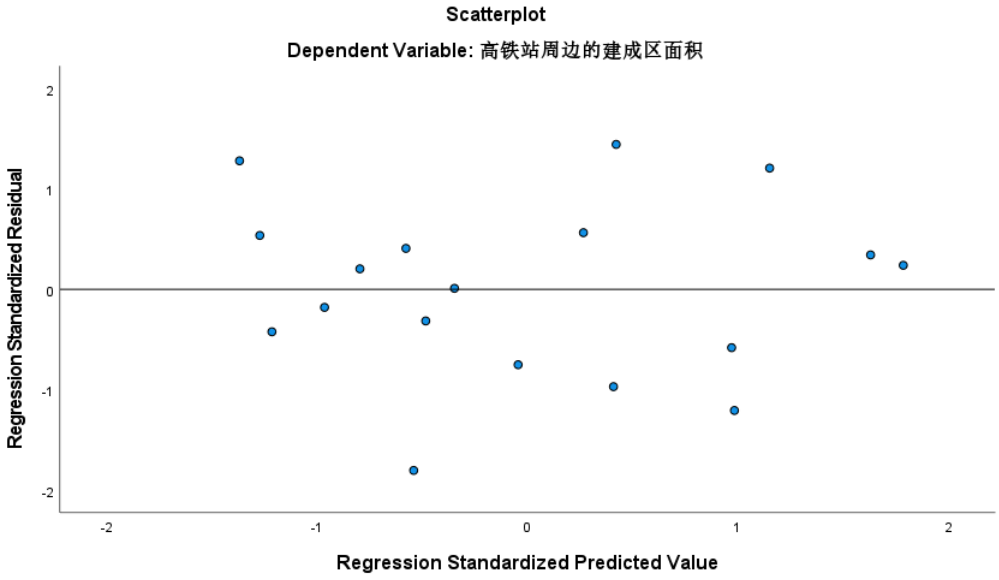


图 37 残差分析结果：预测值~残差散点图

最后，我们使 SPSS 的一般绘图工具绘制另一个残差图。在菜单栏中选择“Graphs → Legacy Dialogs → Scatter/Dot”，在弹出的对话框中选择“Simple Scatter”，然后点击“Define”进入散点图定义对话框。将“Unstandardized Residual [RES_1]”选入“Y Axis”，将“距离最近机场 [D2]”选入“X Axis”，如图 38 所示。

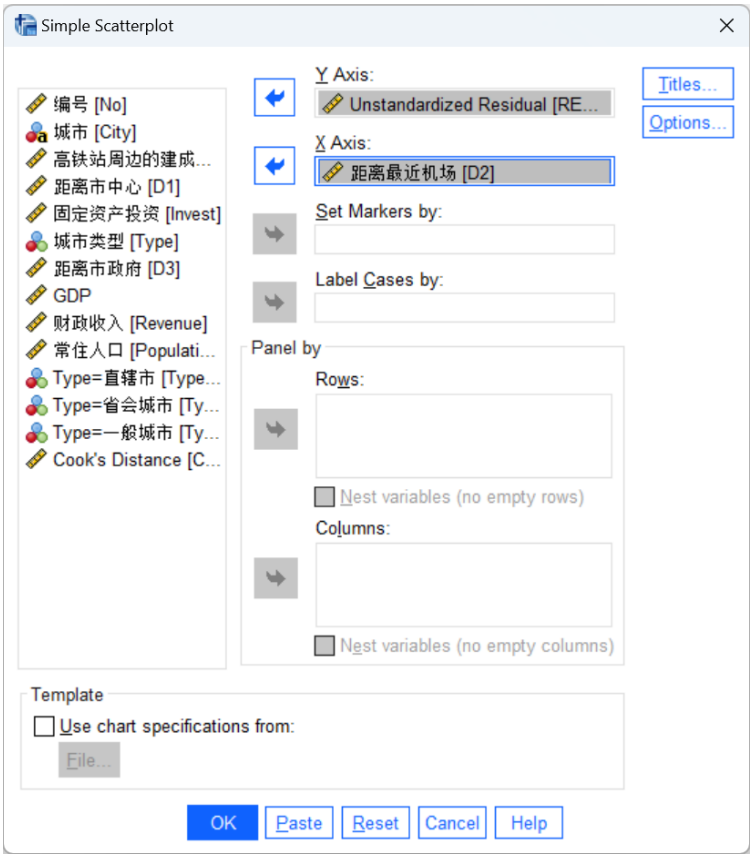


图 38 残差分析设置：绘制自变量~残差散点图

点击“OK”，在生成的散点图中以相同的方式添加水平参考线，即得到如图 39 所示的自变量~残差散点图，用以判断散点图中 x 轴的自变量与因变量是否具有线性关系。

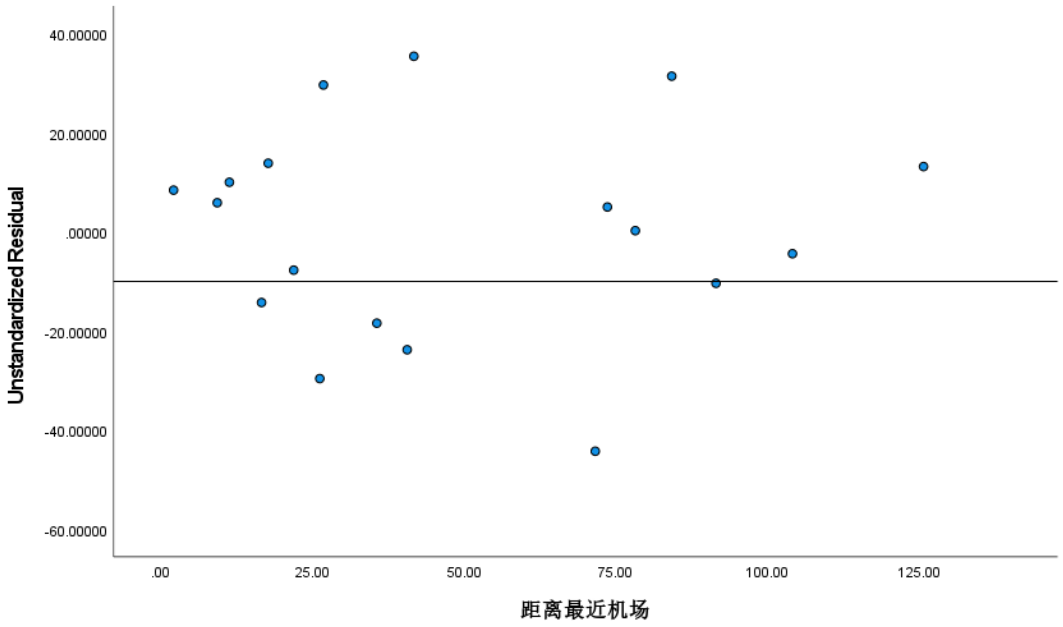


图 39 残差分析结果：自变量~残差散点图