# 1. 概述

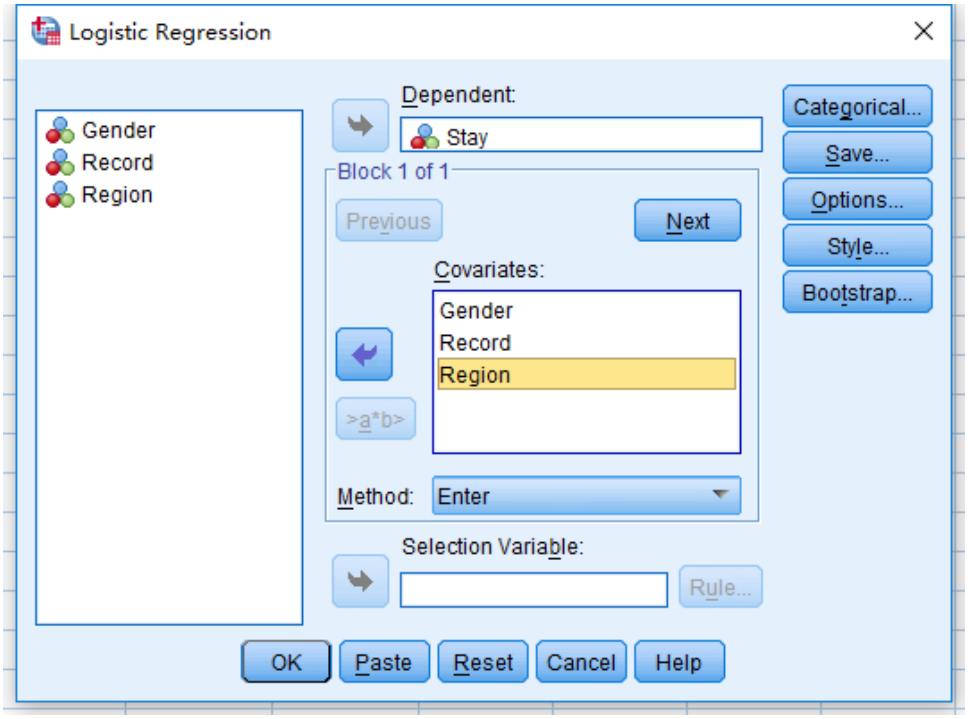
本节以 SPSS 软件演示 Logistic 回归模型的应用。二元 Logistic 回归模型使用大学生就业去向数据，文件为“graduate.sav”。多元 Logistic 回归模型使用工业用地更新数据，文件为“manufacture.sav”。除了 SPSS 的 sav 文件外，附件中还提供了相应的 csv 文件。

# 2. 二元 Logistic 回归

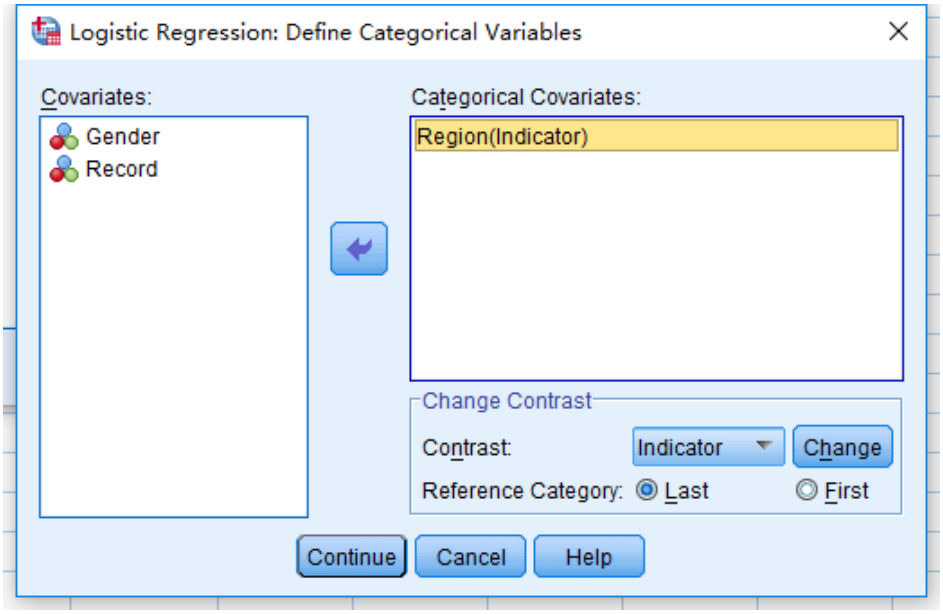
在 SPSS 中打开“graduate.sav”，如下所示。

| | Gender | Record | Region | Stay |
|---|---|---|---|---|
| 1 | 女 | 较好 | 东北 | 否 |
| 2 | 女 | 较好 | 东北 | 否 |
| 3 | 女 | 一般 | 西部 | 否 |
| 4 | 男 | 一般 | 中部 | 否 |
| 5 | 男 | 一般 | 西部 | 否 |
| 6 | 男 | 一般 | 西部 | 否 |
| 7 | 女 | 一般 | 西部 | 是 |

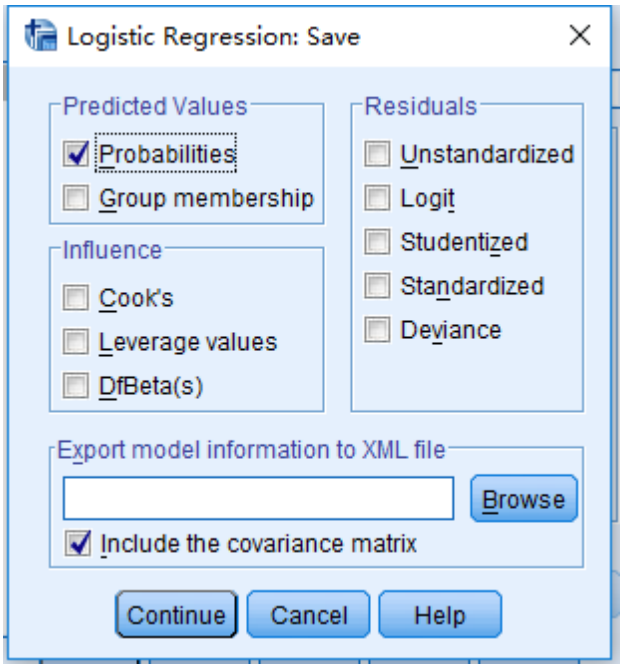
在菜单栏中选择“Analyze – Regression – Binary Logistic”，进入二元 Logistic 回归模型对话框。在“Dependent”中选择“Stay”，作为模型因变量，在“Covariates”中选入“Gender”、“Record”、“Region”，作为模型的自变量，如下图所示。



此时，“Covariates”中的三个变量均被默认视为连续变量。“Gender”和“Record”本身是二分类变量，这样是可以的，而“Region”是无序多分类变量，它的 4 个水平之间没有连续的数值关系，因此必须设置虚拟变量。SPSS 的 Logistic 回归模型可以方便地处理分类自变量，我们在上面的对话框中单击“Categorical”，在弹出的分类变量对话框中将“Region”从左侧的“Covariates”选入右侧的“Categorical Covariates”，将其声明为分类变量。同时，注意到最下方的“Reference Category:”中，“Last”处于被选中的状态，表明，Region 的最后一个水平（Region=4，中部）被设置为参照水平。




在模型估计完成后，SPSS 可以利用模型结果对数据集中的样本进行预测，并保存结果。点击“Continue”回到主对话框，点击“Save”进入保存对话框，在“Predicted Value”中勾选“Probabilities”，保存概率预测结果。如果勾选“Group membership”，则还将保存分类预测结果。



点击“Continue”回到主对话框，再点击“OK”运行分析。首先汇报的是数据汇总情况，本案例中共有 30 个样本，没有缺失样本，也没有样本被排除于分析之外。

**Case Processing Summary**

| Unweighted Cases[a] | | N | Percent |
|---|---|---|---|
| Selected Cases | Included in Analysis | 30 | 100.0 |
| | Missing Cases | 0 | .0 |
| | Total | 30 | 100.0 |
| Unselected Cases | | 0 | .0 |
| Total | | 30 | 100.0 |

a. If weight is in effect, see classification table for the total number of cases.

下面的因变量编码表提示了因变量的编码方式：Stay 的 2 个水平中，“否”为 0，“是”为 1。在结果解读时，应注意模型预测的是 Stay=1 的概率。

**Dependent Variable Encoding**

| Original Value | Internal Value |
|---|---|
| 否 | 0 |
| 是 | 1 |

下面的分类变量编码表提示了我们之前设置的分类自变量 Region 的编码方式。

**Categorical Variables Codings**

| | | Frequency | Parameter coding | | |
|---|---|---|---|---|---|
| | | | (1) | (2) | (3) |
| Region | 东北 | 6 | 1.000 | .000 | .000 |
| | 西部 | 8 | .000 | 1.000 | .000 |
| | 东部 | 7 | .000 | .000 | 1.000 |
| | 中部 | 9 | .000 | .000 | .000 |

接下来，结果窗口将呈现两个“Block”，对应于模型估计的两个阶段：SPSS 将首先估计一个仅有常数项、没有其他自变量的模型，称为零模型，其结果显示在“Block 0: Beginning Block”之中；然后，SPSS 将把我们指定的 3 个自变量纳入模型中，再进行估计，结果显示在“Block 1: Method=Enter”之中。这里的“Method”是自变量纳入的方式，“Enter”表示强制纳入，此外还有一系列逐步纳入、逐步剔除的自变量筛选模型。

由于“Block0”和“Block1”中结果的形式相近，这里我们仅介绍“Block1”的结果，即模型的最终结果。

首先，SPSS 汇报了模型系数的联合显著性，如下图所示。“Omnibus tests”将比较本模型和上一次估计的模型，检验两者的拟合优度是否存在显著差异。由于本模型中不涉及自变量筛选，因此上一步、上一个 Block、上一个模型都是一样的。“Chi-square”是前后两个模型在拟合优度上的差异，基于对数似然值计算；“df”是自由度，反映了前后两个模型在自变量数量的差异，这里为 5，包括“Gender”、“Record”，以及“Region”的 3 个虚拟变量；“Sig”是检验结果，0.097 小于 0.1，表明当前模型在 10%的水平下显著优于之前估计的零模型。

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 9.309 | 5 | .097 |
| | Block | 9.309 | 5 | .097 |
| | Model | 9.309 | 5 | .097 |

下面，SPSS 汇报了当前模型的“-2 倍对数似然值”以及 2 个 $r^2$ 指标。SPSS 并不汇报 McFadden $r^2$，该指标需要根据对数似然值自行计算。

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 32.147[a] | .267 | .356 |

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

接下来，SPSS 汇报了分类的实际值与预测值交叉表，行为实际分类，列为预测分类。模型的总体分类准确率为 76.7%。

**Classification Table[a]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | Stay | | Percentage Correct |
| | | | 否 | 是 | |
| Step 1 | Stay | 否 | 14 | 2 | 87.5 |
| | | 是 | 5 | 9 | 64.3 |
| | Overall Percentage | | | | 76.7 |

a. The cut value is .500

下面，SPSS 汇报了参数估计的结果。

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Gender | 1.466 | 1.017 | 2.075 | 1 | .150 | 4.330 |
| | Record | 2.067 | 1.123 | 3.387 | 1 | .066 | 7.902 |
| | Region | | | 3.533 | 3 | .317 | |
| | Region(1) | .393 | 1.382 | .081 | 1 | .776 | 1.481 |
| | Region(2) | 1.072 | 1.167 | .844 | 1 | .358 | 2.921 |
| | Region(3) | 2.323 | 1.278 | 3.304 | 1 | .069 | 10.208 |
| | Constant | -2.663 | 1.176 | 5.129 | 1 | .024 | .070 |

a. Variable(s) entered on step 1: Gender, Record, Region.

最后，回到数据视图中，可以看到在最右侧出现了新的一列“PRE_1”，这就是我们保存的“Stay=1”的概率预测结果。例如，对于第一个样本，一个成绩较好、户籍地位于东北的女生，预测其留在苏州的概率为 0.44925。

| | Gender | Record | Region | Stay | PRE_1 |
|---|---|---|---|---|---|
| 1 | 女 | 较好 | 东北 | 否 | .44925 |
| 2 | 女 | 较好 | 东北 | 否 | .44925 |
| 3 | 女 | 一般 | 西部 | 否 | .16916 |
| 4 | 男 | 一般 | 中部 | 否 | .23185 |
| 5 | 男 | 一般 | 西部 | 否 | .46854 |
| 6 | 男 | 一般 | 西部 | 否 | .46854 |
| 7 | 女 | 一般 | 西部 | 是 | .16916 |
| 8 | 女 | 较好 | 东部 | 是 | .84901 |

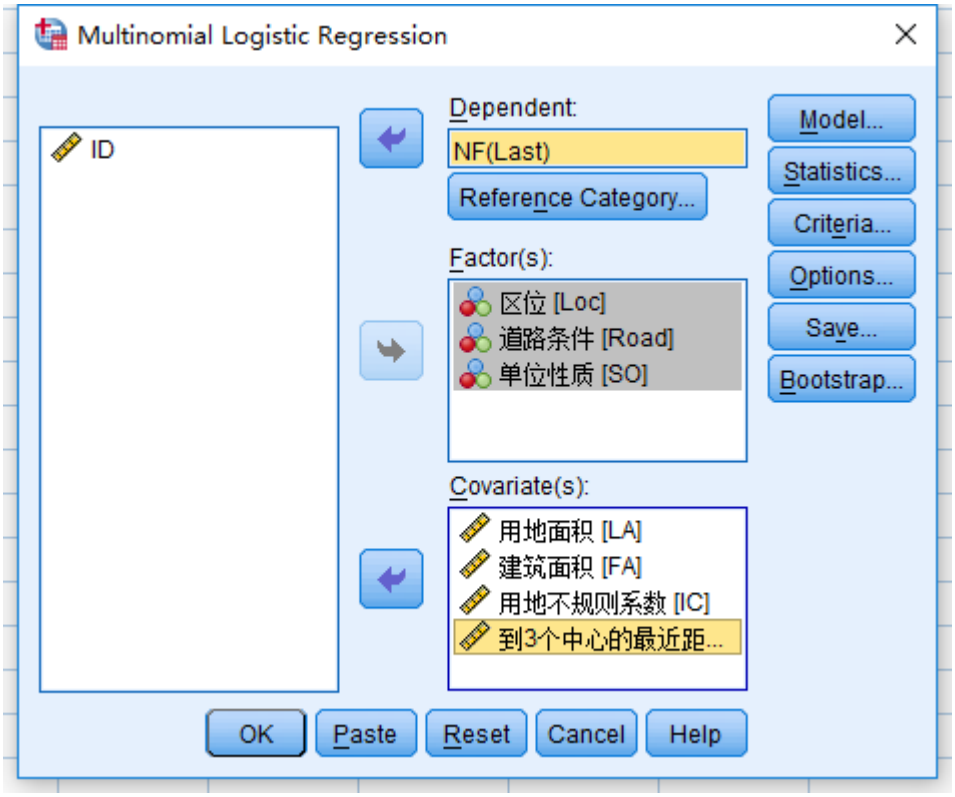
## 3. 多元 Logistic 回归

在 SPSS 中打开“manufacture.sav”，如下所示。

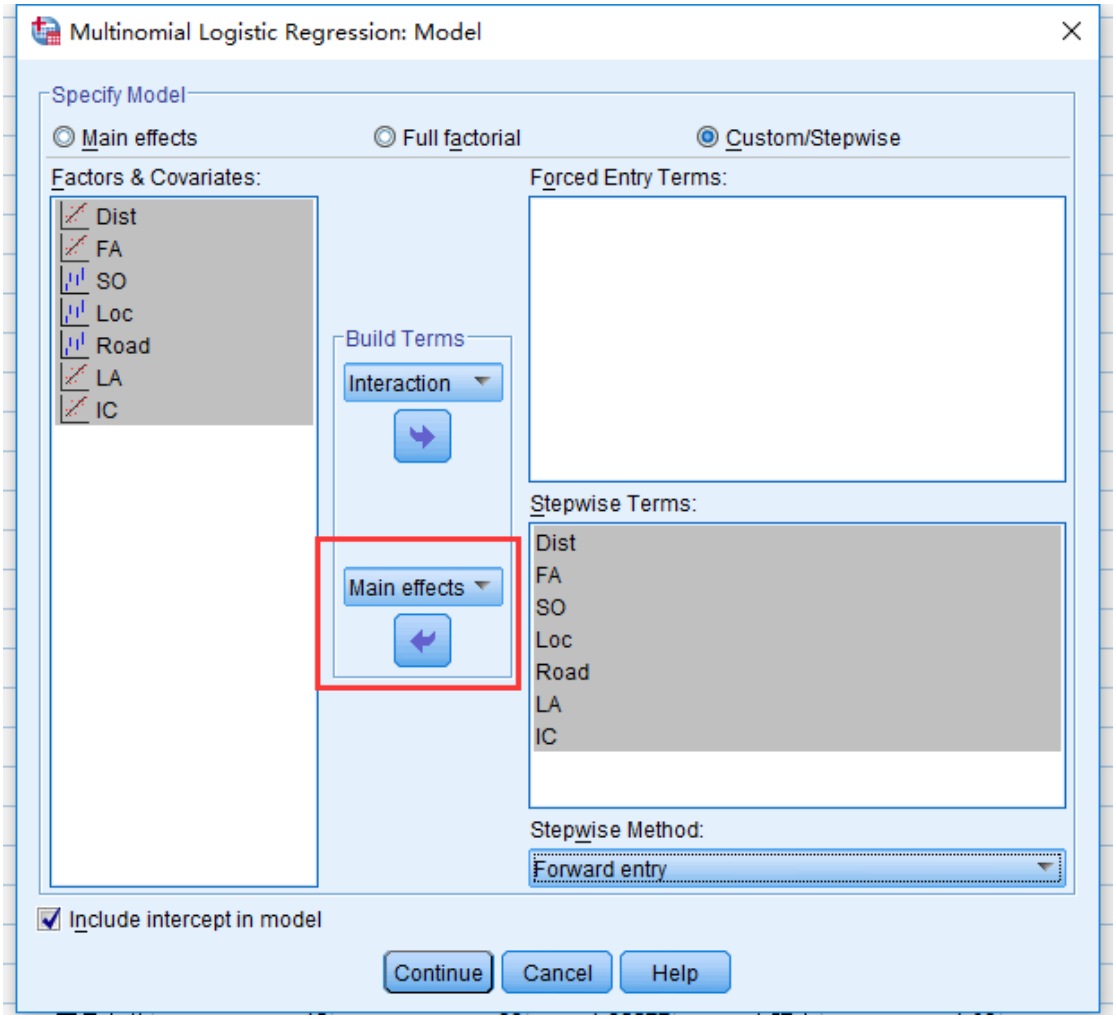
| | ID | NF | LA | FA | IC | Loc | Dist | Road | SO |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 创意产业园 | .47 | .57 | 5.26096 | 内环内 | 2.12 | 2 | 国有企业 |
| 2 | 2 | 创意产业园 | 1.77 | 3.34 | 5.01841 | 内环内 | .46 | 2 | 国有企业 |
| 3 | 3 | 创意产业园 | 3.34 | 3.96 | 4.22295 | 内中环之间 | .47 | 干路 | 非国有企业 |
| 4 | 4 | 创意产业园 | .83 | 1.87 | 3.89304 | 内环内 | .79 | 2 | 国有企业 |
| 5 | 5 | 创意产业园 | .56 | .82 | 4.01145 | 内环内 | 1.21 | 2 | 非国有企业 |
| 6 | 6 | 创意产业园 | 1.22 | 1.83 | 4.08635 | 中外环之间 | .64 | 2 | 国有企业 |
| 7 | 7 | 创意产业园 | .10 | .35 | 4.50294 | 中外环之间 | .55 | 干路 | 国有企业 |
| 8 | 8 | 创意产业园 | .39 | 1.42 | 4.89926 | 内环内 | .48 | 干路 | 国有企业 |
| 9 | 9 | 创意产业园 | .66 | 2.98 | 4.18706 | 内环内 | .24 | 干路 | 国有企业 |
| 10 | 10 | 创意产业园 | .45 | 1.44 | 4.48896 | 内环内 | .26 | 干路 | 国有企业 |

在菜单栏中选择“Analyze – Regression – Multinomial Logistic”，进入多元 Logistic 回归模型对话框。在“Dependent”中选择“更新后的使用功能 [NF]”，作为模型因变量，注意到此时在“Dependent”框中显示的是“NF(Last)”，“Last”表示该变量的最后一个水平（NF=4，生产物流）被自动设置为参照水平，我们可以通过“Reference Category”设置不同的因变量参照水平。然后，将“区位 [Loc]”、

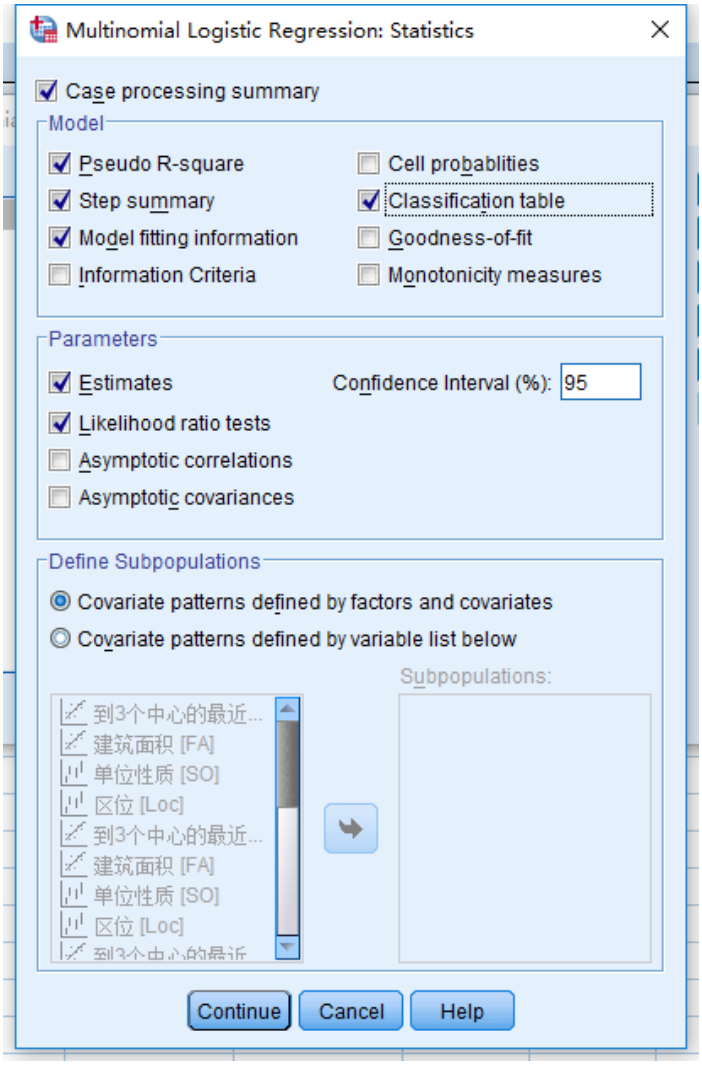
“道路条件 [Road]”、“单位性质 [SO]”这三个变量选入“Factor(s)”中，作为分类自变量，SPSS 将自动将它们设置虚拟变量；同时，将“用地面积 [LA]”、“建筑面积 [FA]”、“用地不规则系数 [IC]”、“到 3 个中心的最近距离 [Dist]”这四个变量选入“Covariate(s)”中，作为连续自变量，如下图所示。



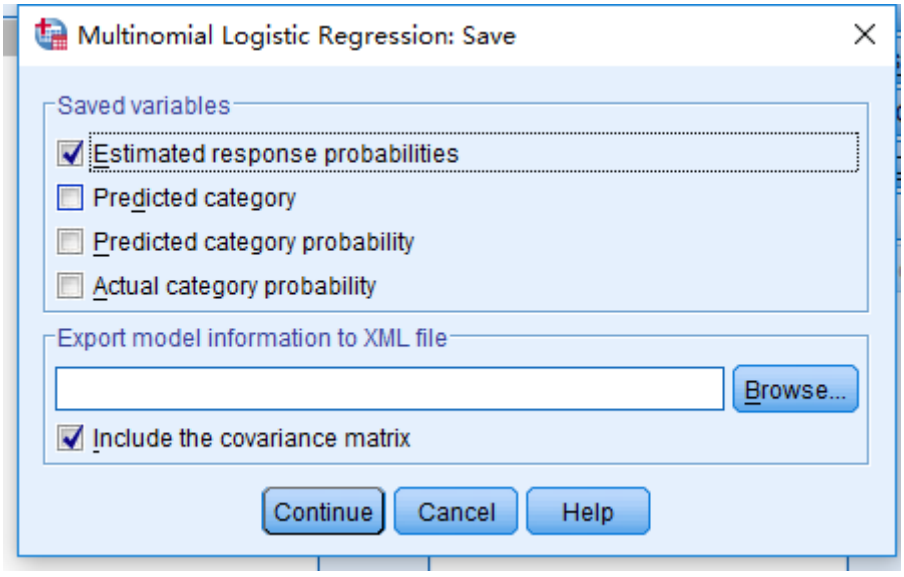

本模型中，我们使用了筛选自变量的逐步回归技术。设置方式如下：点击“Model”，在弹出的模型设定对话框中，将“Specify Model”的选中项由“Main effects”改为“Custom/Stepwise”；然后，将左侧“Factors & Covariates”中的全部自变量选入右下方的“Stepwise Terms”中，由模型进行筛选。需要注意的是，“Stepwise Terms”左侧的“Build Terms”下拉列表中默认选中“Interaction”，我们需要首先将其改选为“Main effects”，即使用各自变量的主效应（而非彼此的交互效应），然后再将自变量选入逐步回归项中。此外，下方的“Stepwise Method”可用于选择具体的逐步回归方法，这里使用默认的“Forward entry”即可。如果希望有一些自变量不受逐步回归的影响，而强制纳入模型，则可以将其选入右上方的“Forced Entry Terms”中。




点击“Continue”回到主对话框，然后点击“Statistics”进入统计量对话框，在“Model”中选中“Classification table”，要求 SPSS 报告分类交叉表。




点击“Continue”回到主对话框，然后点击“Save”进入保存对话框，在“Saved variables”中勾选“Estimated response probabilities”，保存分类预测概率。



点击“Continue”回到主对话框，点击“OK”运行模型。在 SPSS 的结果窗口中，首先汇报的是样本汇总情况，包括因变量和三个分类自变量的水平设定，每个水平的样本量和占比。

**Case Processing Summary**

| | | N | Marginal Percentage |
|---|---|---|---|
| 更新后的使用功能 | 创意产业园 | 13 | 11.5% |
| | 商务办公 | 33 | 29.2% |
| | 商业服务 | 30 | 26.5% |
| | 生产物流 | 37 | 32.7% |
| 区位 | 内环内 | 63 | 55.8% |
| | 内中环之间 | 22 | 19.5% |
| | 中外环之间 | 28 | 24.8% |
| 道路条件 | 干路 | 26 | 23.0% |
| | 支路 | 87 | 77.0% |
| 单位性质 | 国有企业 | 82 | 72.6% |
| | 非国有企业 | 31 | 27.4% |
| Valid | | 113 | 100.0% |
| Missing | | 0 | |
| Total | | 113 | |
| Subpopulation | | 113[a] | |

a. The dependent variable has only one value observed in 113 (100.0%) subpopulations.

然后，SPSS 报告了逐步回归法筛选自变量的过程。“Model0”是只有常数项（Intercept）的零模型，SPSS 将一步一步地纳入自变量，每次纳入时计算此时模型的“-2 倍对数似然值”，该值的下降意味模型拟合优度的提高。每纳入一个自变量，SPSS 将比较当前模型与未纳入该自变量的上一个模型的拟合优度差异，反映在表中的“Chi-Square”统计量上，该统计量就是两个模型在“-2 倍对数似然值”的差。基于该统计量，SPSS 将检验两个模型的差异是否具有显著意义，如果差异显著，则说明新纳入的自变量能够显著地改善模型。从表中可以看到，SPSS 最终纳入了 5 个自变量。

**Step Summary**

| Model | Action | Effect(s) | Model Fitting Criteria | Effect Selection Tests | | |
|---|---|---|---|---|---|---|
| | | | -2 Log Likelihood | Chi-Square[a] | df | Sig. |
| 0 | Entered | Intercept | 299.652 | . | | |
| 1 | Entered | Dist | 266.687 | 32.965 | 3 | .000 |
| 2 | Entered | FA | 246.009 | 20.677 | 3 | .000 |
| 3 | Entered | Road | 232.530 | 13.480 | 3 | .004 |
| 4 | Entered | LA | 220.488 | 12.042 | 3 | .007 |
| 5 | Entered | Loc | 205.477 | 15.010 | 6 | .020 |

Stepwise Method: Forward Entry

a. The chi-square for entry is based on the likelihood ratio test.

上面报告的每一步模型的改善，接下来，SPSS 又报告了最终模型相比于只有常数项的零模型的改善。从表中可以看到，最终模型的“-2 倍对数似然值”下降了 94.174，在 1%的水平下显著。

**Model Fitting Information**

| Model | Model Fitting Criteria | Likelihood Ratio Tests | | |
|---|---|---|---|---|
| | -2 Log Likelihood | Chi-Square | df | Sig. |
| Intercept Only | 299.652 | | | |
| Final | 205.477 | 94.174 | 18 | .000 |

然后，SPSS 报告了 3 个 $r^2$ 指标，包括了在二元 Logistic 回归中未报告的 McFadden $r^2$，这些指标都是基于对数似然值计算的。

**Pseudo R-Square**

| | |
|---|---|
| Cox and Snell | .565 |
| Nagelkerke | .608 |
| McFadden | .314 |

接下来，SPSS 通过似然比检验再一次确认所纳入的各个自变量是否具有显著意义。该检验的原理是，对于任何一个自变量，从最终的完整模型中将该自变量剔除，此时模型的拟合优度将下降，反映为“-2 倍对数似然值”的上升，SPSS 将检验该变化是否具有统计显著意义。可以看到，下表中的 5 个变量均至少在 5%的水平下显著，表明任何一个变量的剔除都会使模型的拟合优度显著降低。

**Likelihood Ratio Tests**

| Effect | Model Fitting Criteria | Likelihood Ratio Tests | | |
|---|---|---|---|---|
| | -2 Log Likelihood of Reduced Model | Chi-Square | df | Sig. |
| Intercept | 205.477[a] | .000 | 0 | . |
| Dist | 225.638 | 20.161 | 3 | .000 |
| FA | 218.842 | 13.365 | 3 | .004 |
| Loc | 220.488 | 15.010 | 6 | .020 |
| Road | 220.918 | 15.440 | 3 | .001 |
| LA | 216.747 | 11.270 | 3 | .010 |

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

接下来，SPSS 报告的是最重要的参数估计结果。

**Parameter Estimates**

| 更新后的使用功能[a] | | B | Std. Error | Wald | df | Sig. | Exp(B) | 95% Confidence Interval for Exp (B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Lower Bound | Upper Bound |
| 创意产业园 | Intercept | -1.216 | 1.536 | .627 | 1 | .429 | | | |
| | Dist | -2.954 | .929 | 10.107 | 1 | .001 | .052 | .008 | .322 |
| | FA | 2.931 | 1.148 | 6.515 | 1 | .011 | 18.743 | 1.974 | 177.936 |
| | [Loc=1] | .757 | 1.353 | .313 | 1 | .576 | 2.133 | .150 | 30.239 |
| | [Loc=2] | -.012 | 1.356 | .000 | 1 | .993 | .988 | .069 | 14.087 |
| | [Loc=3] | 0[b] | . | . | 0 | . | . | . | . |
| | [Road=1] | 3.787 | 1.197 | 10.019 | 1 | .002 | 44.146 | 4.230 | 460.715 |
| | [Road=2] | 0[b] | . | . | 0 | . | . | . | . |
| | LA | -.964 | 1.575 | .375 | 1 | .540 | .381 | .017 | 8.351 |
| 商务办公 | Intercept | 1.155 | 1.331 | .753 | 1 | .385 | | | |
| | Dist | -1.759 | .622 | 8.004 | 1 | .005 | .172 | .051 | .582 |
| | FA | 2.977 | 1.322 | 5.067 | 1 | .024 | 19.620 | 1.469 | 262.012 |
| | [Loc=1] | .983 | 1.030 | .910 | 1 | .340 | 2.671 | .355 | 20.107 |
| | [Loc=2] | .216 | 1.082 | .040 | 1 | .842 | 1.241 | .149 | 10.340 |
| | [Loc=3] | 0[b] | . | . | 0 | . | . | . | . |
| | [Road=1] | 2.703 | 1.034 | 6.838 | 1 | .009 | 14.930 | 1.968 | 113.246 |
| | [Road=2] | 0[b] | . | . | 0 | . | . | . | . |
| | LA | -5.220 | 2.250 | 5.382 | 1 | .020 | .005 | 6.577E-5 | .445 |
| 商业服务 | Intercept | -.861 | 1.224 | .495 | 1 | .482 | | | |
| | Dist | -.651 | .503 | 1.679 | 1 | .195 | .521 | .195 | 1.397 |
| | FA | -.208 | 1.141 | .033 | 1 | .855 | .812 | .087 | 7.601 |
| | [Loc=1] | 1.985 | .937 | 4.486 | 1 | .034 | 7.282 | 1.160 | 45.733 |
| | [Loc=2] | -1.446 | 1.276 | 1.284 | 1 | .257 | .236 | .019 | 2.872 |
| | [Loc=3] | 0[b] | . | . | 0 | . | . | . | . |
| | [Road=1] | 2.621 | .957 | 7.496 | 1 | .006 | 13.743 | 2.106 | 89.701 |
| | [Road=2] | 0[b] | . | . | 0 | . | . | . | . |
| | LA | .835 | 1.464 | .326 | 1 | .568 | 2.306 | .131 | 40.612 |

a. The reference category is: 生产物流.

b. This parameter is set to zero because it is redundant.

然后，SPSS 报告了各样本实际类别和预测类别的交叉表，如下所示。表中

的各行代表分类的实际结果，各列代表预测结果。可以看到，本模型的平均分类准确率为 64.6%；对更新为创意产业园的预测最差，准确率为 53.8%；对更新为生产物流的预测最佳，准确率为 70.3%。

最后，我们回到数据视图。可以看到，SPSS 已经计算了每个样本更新为 4 种用地类型的概率值，分别保存在“EST1_1”、“EST2_1”、“EST3_1”、“EST4_1”之中。例如，样本 1 更新为创意产业园、商务办公、商业服务、生产物流的概率预测值分别为 0.00、0.05、0.48、0.47。

| | ID | NF | LA | FA | IC | Loc | Dist | Road | SO | EST1_1 | EST2_1 | EST3_1 | EST4_1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 创意产业园 | .47 | .57 | 5.26096 | 内环内 | 2.12 | 支路 | 国有企业 | .00 | .05 | .48 | .47 |
| 2 | 2 | 创意产业园 | 1.77 | 3.34 | 5.01841 | 内环内 | .46 | 支路 | 国有企业 | .97 | .01 | .01 | .00 |
| 3 | 3 | 创意产业园 | 3.34 | 3.96 | 4.22295 | 内中环之间 | .47 | 干路 | 非国有企业 | 1.00 | .00 | .00 | .00 |
| 4 | 4 | 创意产业园 | .83 | 1.87 | 3.89304 | 内环内 | .79 | 支路 | 国有企业 | .38 | .42 | .14 | .06 |
| 5 | 5 | 创意产业园 | .56 | .82 | 4.01145 | 内环内 | 1.21 | 支路 | 非国有企业 | .03 | .17 | .52 | .28 |
| 6 | 6 | 创意产业园 | 1.22 | 1.83 | 4.08635 | 中外环之间 | .64 | 支路 | 国有企业 | .60 | .08 | .11 | .21 |
| 7 | 7 | 创意产业园 | .10 | .35 | 4.50294 | 中外环之间 | .55 | 干路 | 国有企业 | .16 | .72 | .10 | .02 |
| 8 | 8 | 创意产业园 | .39 | 1.42 | 4.89926 | 内环内 | .48 | 干路 | 国有企业 | .36 | .60 | .04 | .00 |
| 9 | 9 | 创意产业园 | .66 | 2.98 | 4.18706 | 内环内 | .24 | 干路 | 国有企业 | .70 | .30 | .00 | .00 |
| 10 | 10 | 创意产业园 | .45 | 1.44 | 4.48896 | 内环内 | .26 | 干路 | 国有企业 | .49 | .48 | .03 | .00 |
| 11 | 11 | 创意产业园 | 2.11 | 1.80 | 4.92679 | 内中环之间 | .92 | 支路 | 国有企业 | .29 | .00 | .13 | .58 |