

glmnetUI: glmnet (Lasso 与弹性网络) 软件包的交互式界面

简体中文译本，由人工智能辅助翻译并经编辑部审校；英文版为权威版本。

William Bert Craytor

第 1 卷，第 1 期 · 2026 年夏季

发表日期：2026 年 7 月 1 日

DOI: [10.67330/ja8s2k39](https://doi.org/10.67330/ja8s2k39)

© 2026 William Bert Craytor. Published by Pacific Vista Net.

本开放获取文章依据 [知识共享署名 4.0 国际许可协议 \(CC BY 4.0\)](#) 的条款发布。

摘要。 glmnetUI 是 R 语言 glmnet 软件包的图形用户界面 (graphical user interface)，该软件包通过 Lasso、岭回归与弹性网络 (elastic net) 惩罚来拟合正则化的广义线性模型 (generalized linear model)。它提供三种用途模式——通用预测建模、房地产估价与市场区域分析——并引导用户完成数据导入、模型配置、交叉验证拟合、系数路径查看以及报告下载。本文记录了 glmnetUI 的数据格式要求、建模工作流程、输出显示界面以及完整的功能参考。

关键词： Lasso，岭回归，弹性网络，正则化回归，glmnet，R，图形用户界面，房地产估价

出版者按 (2026 年 7 月)：付印前不久，glmnetUI 已并入 earthUI (0.11.0 及以后版本)，本文所述的弹性网络工作流程现与 MARS 和 GAM 工作流程一同由 earthUI 提供。glmnetUI 仍以其最终发布形态 (0.5.0) 保留在 <https://github.com/wcraytor/glmnetUI>；后续开发将在 earthUI 中继续。本文适用于已发布的 glmnetUI。

目录

1 引言	3
1.1 历史与背景	3
1.1.1 Lasso	3
1.1.2 岭回归与弹性网络	4
1.1.3 glmnet 软件包	4
1.1.4 与 MARS (earth) 和 GAM (mgcv) 的比较	4
1.1.5 推荐的工作流程	5
1.2 什么是 glmnetUI?	5
1.3 三种用途模式	5
1.4 房地产专用功能	6
1.5 快速上手	6
2 MLS 输入数据要求	7
2.1 文件格式与结构	7
2.2 估价模式所需的列	7
2.3 特殊列命名约定	8
2.4 数据质量与完整性	8
2.5 标的行的位置	8
3 通用模式	9
3.1 概述	9
3.1.1 跳过第一行	9
3.1.2 按用途分别保存设置	9
3.2 侧边栏工作流程	9

3.3	项目	10
3.4	从 earthUI 导入 (可选)	10
3.5	主面板选项卡	10
3.6	设置的持久化	11
3.7	深色模式	11
3.8	区域与本地化设置	11
3.8.1	保存默认值	12
4	估价模式	13
4.1	标的行的处理	13
4.2	基准日与成交时距	13
4.3	特殊列指定	13
4.4	RCA 调整概览	13
5	市场区域分析模式	14
5.1	与估价模式的差异	14
5.2	何时使用市场模式	14
6	变量选择	15
6.1	目标变量	15
6.2	预测变量表	15
6.3	数据类型检测与手动覆盖	15
6.4	预期符号与约束的施加	15
6.5	特殊列类型参考	16
7	参数选择	17
7.1	Alpha (混合参数)	17
7.2	Lambda 的选择	17
7.3	Family (分布族)	17
7.4	Standardize (标准化)	17
7.5	符号约束	18
7.6	松弛 Lasso	18
7.7	随机种子	18
7.8	交互项矩阵	18
7.9	禁止进入主效应 (仅作交互项)	19
7.10	高级参数	19
7.11	设置默认值	19
8	拟合模型	20
8.1	拟合按钮	20
8.2	拟合过程中发生了什么	20
9	结果选项卡	21
9.1	Data Preview (数据预览)	21
9.2	Equation (方程)	21
9.3	Correlation (相关性)	21
9.4	Summary (摘要)	21
9.5	Coefficients (系数)	21
9.6	Variable Importance (变量重要性)	22
9.7	Contributions (贡献)	22
9.8	ANOVA (方差分析)	22
9.9	Diagnostics (诊断)	22

9.10	Glmnet Output (Glmnet 输出)	22
9.11	Report (报告)	23
10	下载数据	24
10.1	输出列	24
10.2	CQA 分数	24
11	RCA 计算与下载	25
11.1	打开 RCA 对话框	25
11.2	CQA 分数插值	25
11.3	输出列	25
12	销售比较表格	26
12.1	可比实例选择对话框	26
12.2	工作簿版式	26
12.3	工作表保护与剩余特征单元格	26
13	下载报告	27
13.1	两步式 Quarto 工作流程	27
13.2	报告格式	27
13.3	报告内容	27
14	与 earthUI 的比较	29
14.1	主要差异	29
14.2	共有的功能	29
14.3	何时该用哪一个	29
15	演示数据集: Appraisal_1.csv	30
15.1	说明	30
15.2	各列	30
15.3	建议的快速上手流程	30
16	系统要求与故障排查	32
16.1	受支持的平台	32
16.2	可选依赖	32
16.3	各平台说明	32
16.4	降级运行	32
16.5	故障排查	32

1 引言

1.1 历史与背景

1.1.1 Lasso

Lasso (Least Absolute Shrinkage and Selection Operator, 最小绝对收缩与选择算子) 由 Robert Tibshirani 于 1996 年提出。这个名称是对牛仔套索 (lasso) 的有意双关——套索是一种既能约束又能选择的绳索。Tibshirani 之所以选用这个名字, 是因为该方法真的会“套住”系数, 约束它们的绝对值, 并把其中一部分拉到恰好为零, 从而选出哪些变量留在模型之中。

Tibshirani 的洞见在于给最小二乘目标函数加上一个 L_1 惩罚 (penalty, 系数绝对值之和)。岭回归 (ridge regression, Hoerl 与 Kennard, 1970) 所用的 L_2 惩罚只会把系数向零收缩而永

远达不到零；与之不同， L_1 惩罚产生的是稀疏解，其中许多系数恰好等于零。这使得 Lasso 既是一种正则化（regularization）方法，也是一种变量选择方法。

1.1.2 岭回归与弹性网络

岭回归（ridge regression）由 Arthur Hoerl 与 Robert Kennard 于 1970 年提出。它添加的是 L_2 惩罚（系数平方之和），会按比例收缩所有系数，但把所有变量都保留在模型中。当预测变量之间存在相关性（多重共线性）时，岭回归很有效，但它不会简化模型。

弹性网络（elastic net）由 Hui Zou 与 Trevor Hastie 于 2005 年提出，它把两种惩罚结合起来。混合参数 α 控制两者的配比： $\alpha = 1$ 为纯 Lasso， $\alpha = 0$ 为纯岭回归，介于其间的取值则给出折中方案。弹性网络克服了 Lasso 的一个局限：当预测变量高度相关时，Lasso 往往只选中其中一个而忽略其余，而弹性网络则会把相关的预测变量成组保留。

1.1.3 glmnet 软件包

R 软件包 glmnet 由斯坦福大学的 Jerome Friedman、Trevor Hastie 与 Robert Tibshirani 开发。它使用极其高效的坐标下降算法为广义线性模型（generalized linear model, GLM）实现弹性网络正则化（Friedman, Hastie, and Tibshirani 2010），其后又扩展到 Cox 比例风险模型（Simon et al. 2011）以及全部广义线性模型族（Tay, Narasimhan, and Hastie 2023）。该软件包可处理 gaussian（线性）、binomial（逻辑）、poisson 及其他 GLM 族。

glmnet 的主要特性包括：

- 正则化路径：一次运算即可在整个 lambda 网格上拟合模型，比在每个 lambda 处分别拟合快得多。
- 交叉验证（cross-validation）：`cv.glmnet()` 通过 k 折交叉验证自动选出最优的正则化强度。
- 系数界限：支持对单个系数设定上限和下限（符号约束）。
- 松弛 lasso（relaxed lasso）：在选中的变量上以减弱的惩罚或完全不加惩罚重新拟合模型，以减小偏差。

1.1.4 与 MARS (earth) 和 GAM (mgcv) 的比较

earth 应用程序及其两个后处理模块（glmnet 与 mgcv）使用不同的建模引擎。下表总结了它们的主要差异：

特性	glmnet（弹性网络）	earth（MARS）	mgcv（GAM）
模型类型	线性（带正则化）	分段线性（自适应样条）	光滑非线性（惩罚样条）
非线性	仅能通过交互项或基函数展开实现	通过带数据驱动节点的铰链函数自动实现	通过光滑函数自动实现（薄板、三次样条等）
变量选择	内置，经由 L_1 惩罚（Lasso）	内置，经由前向 / 后向逐步法配合 GCV 剪枝	可选，经由双惩罚收缩（ <code>select=TRUE</code> ）
交互项	两两交叉乘积（不易解释）	基于铰链函数的乘积（可解释，最高 3 阶）	张量积光滑项（ <code>te()</code> 、 <code>ti()</code> ）
可解释性	系数即线性权重，逐个易于说明	g 函数按变量展示分段线性偏效应	光滑的偏效应曲线，非常直观
最适用于	高维数据、变量选择、小样本	自动识别非线性与交互作用	光滑的非线性关系、灵活建模
RCA 调整	逐变量的线性调整	经由 g 函数的分段线性调整	经由偏效应的光滑调整
过拟合控制	交叉验证选定的 lambda	基于 GCV 的剪枝	REML/GCV 光滑参数估计

MARS (Multivariate Adaptive Regression Splines, 多元自适应回归样条) 由 Jerome Friedman 于 1991 年提出。它通过从数据中自适应地选择铰链函数 $\max(0, x - k)$ 及其节点位置来构建分段线性模型。其 R 实现是 Stephen Milborrow 编写的 **earth** 软件包。MARS 的长处在于无需用户指定即可自动发现非线性关系与交互作用。它的 g 函数 (按变量归组的基项) 提供了高度可解释的偏效应曲线。

GAM (Generalized Additive Models, 广义可加模型) 由 Hastie 与 Tibshirani 于 1990 年正式确立。其 R 实现 **mgcv** 由 Simon Wood 编写, 提供带自动光滑度选择 (经由 REML 或 GCV) 的惩罚回归样条。GAM 产生的光滑偏效应曲线便于可视化和解释。当 **earthUI** 把节点位置导出到 **mgcvUI** 时, **earth** 得出的节点会作为 GAM 光滑项的起始点, 从而把 MARS 的自适应节点定位与 GAM 的光滑估计结合起来。

1.1.5 推荐的工作流程

对于房地产估价 (real estate appraisal) 以及其他需要可解释、可辩护模型的类似应用:

1. 从 **earthUI** 入手, 发现数据中的重要变量、非线性关系与交互作用。**earth** 的自动基函数选择与 g 函数能提供极佳的初步洞察。
2. 用 **mgcvUI** 加以细化, 如果您希望得到更光滑的偏效应。把 **earthUI** 的节点导入 **mgcvUI**, 即可从分段线性模型平滑过渡到光滑模型。
3. 使用 **glmnetUI**, 当您需要力度更大的变量选择 (Lasso)、预测变量数量接近或超过观测数量, 或者希望得到一个带正则化、便于辩护的纯线性模型时。

1.2 什么是 glmnetUI?

glmnetUI 是 R 语言 **glmnet** 软件包的图形用户界面 (graphical user interface)。它以本地 Shiny 应用程序的形式运行——无需登录、无需服务器、也没有账户。您从 R 中启动它, 导入数据集 (CSV 或 Excel), 配置模型, 然后交互式地进行拟合。

从当前版本起, **glmnetUI** 不再是一个单独安装的软件包: 它作为弹性网络后处理模块随 **earthUI** 软件包一起发布, 用 `earthUI::launch_glmnet()` 启动。应用程序本身——以及本文所述的一切——并无改变; 不同的只是安装方式和启动命令。三个例程 (**earth**、**glmnet**、**mgcv**) 共享同一个项目树、同一个设置数据库以及同一套支撑基础设施。

该应用程序提供了完整的工作流程: 数据导入、变量配置、模型拟合、诊断图、变量重要性、模型方程, 以及可下载的 Word、PDF 或 HTML 格式报告。

弹性网络回归结合了两种正则化技术:

- 岭回归 ($\alpha = 0$): 按比例把所有系数向零收缩。当您认为所有变量都相关时较为合适。
- **Lasso** 回归 ($\alpha = 1$): 可以把系数收缩到恰好为零, 从而实现自动变量选择。产生的模型更简洁、更易解释。
- 弹性网络 ($0 < \alpha < 1$): 两种做法的混合。当预测变量之间相关时很有用。

1.3 三种用途模式

每个 **glmnetUI** 项目都有一个 **Purpose** (用途), 在创建项目时选定 (第 1 节, *Project* (项目))。当前项目的用途决定哪些工具和界面元素会出现; 若要在另一种模式下工作, 请创建或打开一个具有该用途的项目。三种模式为:

- **General** (通用) ——适用于任何类型的总体或数据集的弹性网络回归。这是默认模式。它提供完整的正则化回归工作流程, 不含任何领域专用的附加功能。
- **For Appraisal** (用于估价) ——为房地产估价定制的弹性网络回归。增加了针对单一物业估值的功能, 包括标的物业处理、特殊列指定以及残差约束法 (Residual Constraint Approach, RCA)。
- **Market Area Analysis** (市场区域分析) ——为市场区域研究定制的弹性网络回归。增加了分析既定市场中成组物业的功能, 并可选择排除标的行。

在这三种模式下，核心建模引擎完全相同——您始终是在拟合一个弹性网络 (glmnet) 模型。用途设置只控制哪些附加工具和界面元素可用。

1.4 房地产专用功能

当选择 **For Appraisal** 或 **Market Area Analysis** 时，glmnetUI 会启用若干为房地产分析设计的功能：

- **特殊列指定**——每个预测变量都可以标记一个特殊角色，例如 `contract_date`、`dom`、`concessions`、`latitude`、`longitude`、`living_area`、`lot_size`、`actual_age`、`effective_age`、`area`、`site_dimensions` 或 `display_only`。这些指定控制该列在拟合与输出过程中的处理方式。
- **经纬度取整**——被指定为纬度或经度的列会自动四舍五入到小数点后 3 位，以防止过拟合。
- **Sale Age 列（成交时距）**——当某列被指定为 `contract_date` 且提供了 Effective Date（估价基准日）时，glmnetUI 会计算一个 `sale_age` 列（成交日期与基准日之间的天数），并用它替代原列作为预测变量。
- **RCA 计算（仅限估价模式）**——在估价模式下，拟合模型之后 glmnetUI 可以计算残差约束法（RCA）输出，得出逐个可比实例的调整值、净 / 毛调整汇总，以及标的物业的调整后成交价格。

提示：*General* 用途模式适用于任何数据集——金融、科学、工程或房地产。估价模式与市场模式只是为房地产专业人士增加了一些便利功能。

1.5 快速上手

使用 glmnetUI 的步骤：

1. 在 R 中安装 **earthUI**: `install.packages("earthUI")` ——glmnet 例程已包含在内。(关于开发版和可选组件，参见本期 earthUI 一文中的“Installing earthUI and vProlog”。)
2. 启动应用程序：在 R 中运行 `earthUI::launch_glmnet()`。应用程序会在您的网页浏览器中以端口 7879 打开。您也可以在浏览器中直接访问 <http://localhost:7879>。应用程序会记住您上次使用的用途模式并自动恢复。
3. 在侧边栏第 1 节创建或打开项目。项目设定用途和工作地点（国家、州 / 省、县、市）及其输入 / 输出文件夹。项目存放在共享的 `regProj` 根目录下，并与姊妹应用程序 earthUI 和 mgcvUI 共用。打开一个具有不同用途的项目，就是切换模式的方式。
4. 在第 2 节导入数据，从当前项目的输入文件夹中选择一个 CSV 或 Excel 文件。把文件放入该文件夹后，点击 **Refresh**（刷新）即可列出它们。
5. 从 **earthUI** 导入（可选）——第 3 节允许您导入一个 earthUI 结果 `.rds` 文件，以便把 earth 的铰链基函数与 glmnet 的正则化结合使用。若没有 earthUI 结果，可跳过此步。
6. 配置变量——选择目标变量和预测变量，设置数据类型、预期符号，并指定任何特殊列角色。
7. 设定 **glmnet** 参数——`alpha`、`lambda`、`family`、符号约束、交互项及其他选项。
8. 拟合模型——点击“Fit Glmnet Model”（拟合 Glmnet 模型），并在主面板中查看结果。
9. 导出——将预测结果下载为 Excel，生成并转换 Quarto 报告，或（在估价模式下）计算 RCA 调整值与销售比较表格（Sales Comparison Grid）。

设置会自动保存在浏览器的本地存储中，并在您重新载入同一个输入文件时恢复。

2 MLS 输入数据要求

对于房地产估价与市场分析工作流程，您的输入数据通常来自多重上市服务（Multiple Listing Service, MLS）导出文件。本章说明预期的文件结构以及 glmnetUI 可以使用的列。

2.1 文件格式与结构

glmnetUI 接受 CSV 与 Excel (.xlsx、.xls) 文件。导入时，列名会自动转换为 snake_case——例如，“Living SqFt”变为 living_sqft，“Contract Date”变为 contract_date，“Sale Price”变为 sale_price。这种规范化确保了整个工作流程中列引用的一致性。导入时使用的 CSV 分隔符和小数点标记由区域设置决定（参见第 3 章“Locale & Regional Settings”）。

您的数据文件应当是一张扁平表格，每行一处物业，每列一个属性。文件的第一行必须包含列标题。

2.2 估价模式所需的列

虽然 glmnetUI 可以处理任意一组列，但完整的估价工作流程若具备以下各列会更为得力：

列	特殊类型	用途
Sale Price（成交价格）	（目标变量）	模型的响应变量
Contract Date（合同日期）	contract_date	用于计算 sale_age（距基准日的天数）
Listing Date（挂牌日期）	listing_date	在没有 DOM 列时，与合同日期一起计算 DOM
Days on Market（在市天数）	dom	在市天数；显示于导出结果中
Concessions（成交让利）	concessions	成交让利；净成交价 = 成交价格 - 让利
Living Area (SF)（居住面积）	living_area	启用每平方英尺残差（residual_sf、cqa_sf）
Lot Size（地块面积）	lot_size	宗地面积列
Site Dimensions（宗地尺寸）	site_dimensions	与地块面积归为一组
Latitude（纬度）	latitude	四舍五入到小数点后 3 位
Longitude（经度）	longitude	四舍五入到小数点后 3 位
Area ID（区域编号）	area	市场区域 / 街区标识
Actual Age（实际年龄）	actual_age	物业年龄
Effective Age（有效年龄）	effective_age	物业的有效年龄
Address（地址）	display_only	显示于导出结果中；不进入模型

电子表格的列名可以使用外语——“特殊”名称之所以用英文，是为了让 R 程序能对它们作特殊处理。除此之外，您所给定的列名会出现在回归模型、图形中，以及（若在做估价）输出报告中。

并非所有列都是必需的。glmnetUI 会自行适应——若缺少某一列，相应的功能就直接省略。不过，对于房地产定价模型，为达到可接受的拟合效果，强烈建议具备以下几列：

1. **Sale Age** (成交时距) ——合同成交日期与估价或分析基准日之间的天数。若使用跨年度的成交历史,尤其是超过 5 年的期间, `sale_age` 在估计成交价格时往往起核心作用。
2. **Living Area** (居住面积) ——也常被称为“Living Sqft”、“GLA”(gross living area, 总居住面积)等。这是成交价格的又一主要决定因素。
3. **Total Bath Count** (卫生间总数) ——全卫、四分之一卫、半卫和四分之三卫的总数。例如,两个全卫加一个半卫记为 2.5。
4. **Garage Bays** (车位数) 或 **Garage Area** (车库面积) ——车库停车位数量或车库面积。
5. **Lot Size** (地块面积) ——物业的土地面积,通常以平方英尺或英亩计。
6. **Longitude** (经度)、**Latitude** (纬度),以及(若有) **Area ID** (区域编号)。这些位置变量有助于模型解释地理上的价格差异。

2.3 特殊列命名约定

glmnetUI 通过特殊类型指定而非列名来识别列。在 MLS 导出文件中,您可以随意命名各列——关键在于您在变量配置表(第 6 章)中指定了正确的特殊类型。

例如,您的 MLS 可能把居住面积导出为“GLA”、“Living SqFt”、“liv_area”或“gross_living_area”。导入之后(列名转为 snake_case),您只需在 Special(特殊)下拉框中把它指定为 living_area。此后无论其原始名称如何,glmnetUI 都会用它来计算每平方英尺的残差。

2.4 数据质量与完整性

- **缺失值 (NA)**: 任何预测变量列或目标列中含 NA 的行,会在拟合前自动剔除。
- **日期列**: 符合常见日期格式的字符列在导入时会自动解析。当检测不到 4 位年份时,优先按 2 位年份格式解析。
- **数值列**: 成交价格、居住面积、地块面积、让利等字段必须为数值。若您的 MLS 导出的价格带有货币符号或千位分隔符(例如“\$350,000”或“350.000,00”),可能需要在导入前先行清洗。
- **因子列**: 文本列、TRUE/FALSE 列和 R 因子列会自动按分类变量处理——回归只能把它们当作因子。而数值列(例如数字形式的区域码或风格码)只有在您勾选其 **Factor** (因子)复选框时才会变成因子;glmnetUI 绝不会根据数值的取值范围推断其为分类变量,因此像 bath_count 这样的离散数值预测变量仍保持连续。Special 角色与是否作为因子无关。

提示: 导入后请查看预测变量表中的 NA 一列。缺失值很多的列可能导致大量行被剔除。可以考虑排除高 NA 列,或在导入前清洗数据。

2.5 标的行的位置

在 **Appraisal** (估价) 模式下,第 1 行必须是标的物业。其余各行均为可比成交实例。标的行不参与模型拟合。拟合完成后,模型仍会为标的行生成预测值。

在 **Market Area Analysis** (市场区域分析) 模式下,把标的放在第 1 行是可选的。

在 **General** (通用) 模式下没有特殊的行处理——所有行一视同仁。

3 通用模式

3.1 概述

启动 glmnetUI 时的默认模式即为通用 (General Purpose) 模式。它为任何数据集——而不仅仅是房地产数据——提供完整的弹性网络回归工作流程。您可以把 glmnetUI 用于科学数据、金融分析、工程研究, 或任何回归问题。

在通用模式下, 界面会略去房地产专用功能 (特殊列、成交时距、坐标取整、RCA)。侧边栏经过精简, 专注于变量选择、参数配置、模型拟合与导出。

3.1.1 跳过第一行

打开通用或市场项目后, 侧边栏顶部会出现一个 **Skip first row** (跳过第一行) 复选框。勾选后, 第 1 行将不参与模型拟合。当第 1 行包含一个不应影响模型的标的或参照观测值时, 这一选项很有用。在估价模式下, 第 1 行 (标的物业) 始终自动排除。

3.1.2 按用途分别保存设置

设置 (预测变量选择、模型参数、交互项) 会按输入文件与用途模式的每种组合分别保存。在通用、估价和市场模式之间切换时, 各模式的设置互不影响、各自保留。

3.2 侧边栏工作流程

侧边栏按编号、可折叠的小节组织, 引导您从项目选择一路走到报告导出。多数小节只在打开活动项目后才出现, 其中几个还与项目的用途有关。输出文件夹由活动项目推导得出——没有单独的输出文件夹字段。

1. Project (项目) ——创建或打开项目。项目设定用途和工作地点 (国家、州 / 省、县、市) 及其输入 (`<os>_in/`) 和输出文件夹。项目存放在共享的 `regProj` 根目录下 (默认 `~/regProj`, 可用 `REGPROJ_ROOT` 环境变量覆盖), 并与姊妹应用程序 `earthUI` 和 `mgcvUI` 共用, 因此三者的模型可以并存。点击 **Close Project** (关闭项目) 即可切换。

2. Import Data (导入数据) ——从活动项目的输入文件夹中选择一个 CSV 或 Excel 文件。通过访达 / 资源管理器把文件放入该文件夹, 然后点击 **Refresh** (刷新) 列出它们。对于含多个工作表的 Excel 文件, 会出现工作表选择器。列名会自动转换为 `snake_case`。

3. Import from earthUI (从 earthUI 导入, 可选) ——导入一个 `earthUI` 结果 `.rds` 文件, 以便把 `earth` 的铰链基函数与 `glmnet` 的弹性网络正则化结合使用。浏览按钮的样式与第 2 节的文件选择器一致。此步骤为可选——若没有可导入的 `earthUI` 结果, 可以跳过。

4. Variable Configuration (变量配置) ——目标变量选择器, 带 Include (纳入)、Factor (因子) 和 Force (强制保留) 复选框的预测变量表, 以及 Sign (符号) 下拉框。Special (特殊) 列——用于指定合同日期、坐标等——只在估价与市场模式下出现。详见第 6 章。

5. glmnet Call Parameters (glmnet 调用参数) ——全部模型配置: `alpha`、`lambda`、`family`、`standardize`、符号约束、松弛 `lasso`、随机种子、交互项矩阵, 以及高级参数 (`lambda.min.ratio`、`nlambda`、CV 损失度量、收敛阈值、最大迭代次数、截距)。完整的参数参考见第 7 章。

6. Fit Glmnet Model (拟合 Glmnet 模型) ——运行模型的按钮。结果会显示在右侧的各个选项卡中。

7. Download Output (下载输出) ——把预测值、残差、CQA 分数和逐变量贡献导出为 Excel 文件, 存入项目的输出文件夹。模型拟合后, 在所有用途模式下均可用。

8. Calculate RCA Adjustments & Download (计算 RCA 调整值并下载) ——为每个可比实例计算逐变量的 RCA 调整值, 并通过 CQA 插值得出标的残差。仅在估价与市场模式下、拟合之后出现。详见第 11 章。

9. Generate Sales Grid & Download (生成销售比较表格并下载) ——生成中间的销售比较表格, 按毛调整百分比对可比实例排序。仅在估价模式下、计算完 RCA 调整值之后出现。详见第 12 章。

10. Generate Quarto Report (生成 Quarto 报告)——把一个自包含的 Quarto (.qmd) 报告包——含全部图形与表格——写入项目的输出文件夹。模型拟合后即可使用。

11. Convert Quarto Report (转换 Quarto 报告)——把生成的 .qmd 报告包渲染为 HTML、Word 或 PDF。PDF 需要安装 LaTeX；若未检测到，则隐藏 PDF 选项。详见第 13 章。

在市场模式与通用模式下，仅限估价的步骤会被跳过，两个报告小节的编号也相应调整。

3.3 项目

glmnetUI 把工作组织为共享 regProj 根文件夹下的项目 (earthUI 与 mgcvUI 使用的是同一棵目录树)。每个项目拥有一个输入文件夹 (<os>_in) 和按引擎划分的输出文件夹 (例如 <os>_out_glmnet)。

要创建项目，请点击第 1 节中的 + **New...** (新建)，并提供：

- **Purpose (用途)**——General、Appraisal 或 Market Area。
- **Location (地点)**——项目的政治 / 行政层级。需要填写哪些层级取决于您选择的国家，因此各国的命名惯例都能自动支持——美国项目使用州 / 县 / 市，法国项目使用 region/département/commune，日本项目使用都道府县 / 市或区 / 町丁，依此类推。
- **Project Name (项目名称)**——一个简短标签，限 8 个字符 (字母、数字、_ 和 -)。

应用程序不会原样使用您输入的名称：它会在前面加上地点代码和创建日期时间，构造出一个唯一的带时间戳的文件夹名，形式为 <country>_<admin level codes>_<YYYYMMDD-HHMMSS>_<your name> ——例如，一个名为 burl 的美国项目会变成 us_ca_081_burlin_20260604-131500_burl。8 个字符的限制只适用于您键入的那一部分。只有新建的项目才会带上时间戳前缀；磁盘上已有的项目保留原名，并且照常完全可用。

3.4 从 earthUI 导入 (可选)

侧边栏的第 3 节——**Import from earthUI** (从 earthUI 导入)——允许您导入一个 earthUI 结果 .rds 文件。这使得 glmnetUI 能够把 earth 的铰链基函数 (分段线性项) 与 glmnet 的弹性网络正则化结合起来，把 earth 的自动非线性识别与 glmnet 的变量选择和系数约束合二为一。

导入 earthUI 结果后，驱动 glmnet 拟合的是 earth 模型的预测变量，而不是 Include 复选框，因为 glmnet 是在 earth 的基函数展开上拟合的。变量配置表会明确体现这一点：**Include** 与 **Type** 被锁定，earth 模型未使用的预测变量所在行被禁用，并有一条提示说明预测变量集来自 earth 模型。**Force** 与 **Sign** 对 earth 的预测变量仍可编辑，因为它们在拟合时仍会作用于 earth 基函数。

浏览按钮的样式与第 2 节的文件选择器一致。此步骤完全可选——若没有 earthUI 结果文件，直接跳过第 3 节进入第 4 节即可。

3.5 主面板选项卡

拟合完成后，主面板提供以下选项卡：

选项卡	内容
Data Preview (数据预览)	导入数据的预览。在估价模式下，分为标的物业表与可比成交实例表。
Equation (方程)	以 LaTeX/MathJax 显示的拟合模型方程，展示每个非零系数与交互项。
Correlation (相关性)	数值型预测变量相关系数的热力图（拟合前即可查看）。
Summary (摘要)	关键指标 (R^2 、调整 R^2 、 GR^2 、 $CV R^2$ 、RMSE、MAE) 与估价指标 (COD、PRD、中位数比率)。
Coefficients (系数)	非零系数表，附符号违例警示。
Variable Importance (变量重要性)	预测变量重要性 ($ \beta \times sd(x)$) 的条形图与排序表。可用时支持 plotly 交互式悬停。
Contributions (贡献)	带斜率标注的逐变量偏效应图。数值型预测变量为散点图 + 拟合线，因子为箱线图，交互项为三维曲面 / 散点 / 热力图。
ANOVA (方差分析)	方差分解表：逐变量的平方和、占模型平方和的百分比、系数。
Diagnostics (诊断)	四幅图：系数路径、CV 误差、实际值对预测值、残差对拟合值。
Glmnet Output (Glmnet 输出)	原始模型输出：所用随机种子、模型打印结果、选定的 λ 、 $\lambda_{min}/1se$ 、 γ (若使用松弛 lasso)，以及完整的系数向量。
Report (报告)	报告配置与导出字段。
RCA Adjustments (RCA 调整)	RCA 分析结果（估价模式，计算完成后）。

3.6 设置的持久化

glmnetUI 会自动把您的配置保存到浏览器的本地存储中，以输入文件名作为索引键。当您重新载入同一个文件时，所有设置都会恢复：目标变量选择、预测变量复选框、数据类型、预期符号、glmnet 参数以及交互项矩阵。上次使用的用途模式也会全局保存，并在重新启动应用程序时恢复。有三个选项可供选择：

- **Use last settings for input file** (使用该输入文件的上次设置) ——恢复为此文件保存的设置
- **Use default settings** (使用默认设置) ——应用您保存的全局默认值
- **glmnet defaults** (glmnet 默认值) ——重置为出厂默认值 ($\alpha = 1$ 、 $CV \lambda_{1se}$ 、10 折、gaussian、开启标准化)

Save current as default (将当前设置存为默认) 按钮可把当前的全部设置保存为全局默认值。

3.7 深色模式

点击右上角的月亮 / 太阳图标，即可在浅色与深色主题之间切换。主题偏好保存在本地存储中，跨会话保持。所有界面元素（表格、图形、卡片、按钮）都会通过 CSS 变量适配所选主题。

3.8 区域与本地化设置

glmnetUI 通过基于国家的区域设置系统支持各国的数字与 CSV 格式惯例。标题栏中的 **Settings** (设置) 下拉菜单提供面向 31 个受支持国家的 **Country** (国家) 与 **Paper** (纸张) 选择器。每个预设会配置：

- **CSV 分隔符**——美国 / 英国 / 日本用逗号 (,)，欧洲多数国家用分号 (;)，因为那里逗号被用作小数点。
- **小数点标记**——句点 (.) 或逗号 (,)。
- **千位分隔符**——逗号 (美国 / 英国 / 日本)、句点 (德国 / 意大利 / 西班牙)、空格 (芬兰 / 法国 / 波兰 / 波罗的海国家 / 乌克兰 / 俄罗斯)，或撇号 (瑞士)。
- **纸张尺寸**——Letter (美国 / 加拿大 / 墨西哥) 或 A4 (其他所有地区)。

3.8.1 保存默认值

点击 **Save as my default** (存为我的默认值)，即可全局保存您的区域设置偏好。这些默认值将适用于今后所有会话，无论您载入哪个数据文件。

提示：图形坐标轴和斜率标注上的数字格式会自动适配您的区域设置。德语区域用句点作千位分隔符 (200.000)，芬兰用空格 (200 000)，瑞士用撇号 (200'000)。不显示任何货币符号——*glmnetUI* 与币种无关。

4 估价模式

当您把 Purpose 选为 **For Appraisal**（用于估价）时，glmnetUI 会针对单一物业估值进行自我配置。第 3 章所述的全部功能依然可用；本章只讲估价模式特有的附加内容。

4.1 标的行的处理

在估价模式下，数据集的第 1 行即标的物业，其余各行均为可比成交实例。您的输入文件必须据此组织（参见第 2 章）。标的的成交价格可以留空或填任意值——glmnetUI 在拟合时会自动把它当作 NA 处理。

导入之后，Data Preview 选项卡会分为两部分：**Subject Property**（标的物业，第 1 行）与 **Comparable Sales**（可比成交实例，第 2 行及以后）。第 1 行始终不参与模型拟合。拟合完成后，模型仍会为标的行生成预测值。

4.2 基准日与成交时距

在估价与市场模式下，变量配置小节中会出现一个 **Effective Date**（估价基准日）字段（默认为当天日期）。如果您在 Special 列下拉框中把某列指定为 `contract_date`，glmnetUI 就会计算一个 `sale_age` 列——即每笔成交的合同日期与基准日之间的整数天数。该列会作为预测变量加入模型。

当基准日发生变化时，`sale_age` 会自动重新计算。

4.3 特殊列指定

在估价与市场模式下，变量配置表中每个预测变量都会出现一个 **Special**（特殊）下拉框。特殊类型及其作用的完整参考见第 6 章。

4.4 RCA 调整概览

Calculate RCA Adjustments & Download（计算 RCA 调整值并下载）按钮（侧边栏第 8 节，在估价与市场模式下拟合之后可见）会为每个可比实例计算相对于标的的市场推导调整值。完整的 RCA 工作流程见第 11 章。

提示：您赋予标的的 *CQA* 分数决定了残差分布中有多大比例归属于标的。分数为 5.00 时，标的位于各可比实例的中位数处。

5 市场区域分析模式

当您把 Purpose 选为 **Market Area Analysis**（市场区域分析）时，glmnetUI 提供与估价模式相同的房地产专用功能（特殊列、成交时距、坐标取整），但其取向是分析一组物业，而非为单一标的估值。

5.1 与估价模式的差异

- 标的行处理更为灵活——在市场模式下，第 1 行不会被自动当作标的物业。
- 没有销售比较表格——“Calculate RCA Adjustments & Download”（第 8 节）步骤在市场模式下可用，但销售比较表格步骤（第 9 节）仅限估价模式。

5.2 何时使用市场模式

在以下情形下适合采用市场区域分析模式：

- 为某个街区或市场区域建立回归模型，以了解价值驱动因素
- 分析面积、年龄、地块面积和位置等变量如何影响一组物业的成交价格
- 为市场状况分析或街区界定准备支撑材料
- 处理某个数据集时，您需要特殊列功能（成交时距、坐标取整），但不需要 RCA 调整工作流程

提示：当您需要特殊列功能但不需要 RCA 调整工作流程时，市场模式同样适用于一般性的房地产回归分析。

6 变量选择

侧边栏的第 4 节——**Variable Configuration**（变量配置）——是您选择哪些列参与模型以及如何处理它们的地方。

6.1 目标变量

第 4 节顶部的 **Target (response) variable**（目标 / 响应变量）下拉框会列出数据集中的每一列。请选择一列作为响应变量（例如成交价格）。目标列会自动从预测变量列表中排除。

6.2 预测变量表

在目标变量选择器下方，有一张表列出其余每一列，包含以下字段：

列	说明
Variable （变量）	列名
Type （类型）	数据类型下拉框：numeric、integer、character、Date、POSIXct
Include （纳入）	复选框——把该列作为预测变量纳入模型
Factor （因子）	复选框——把该变量当作分类因子处理（在模型矩阵中生成哑变量列）
Force （强制保留）	复选框——强制进入模型（惩罚因子 = 0，绝不会被 lasso 剔除）
Special （特殊）	下拉框（仅估价 / 市场模式）——参见下文的特殊列类型参考
Sign （符号）	预期的系数符号：正、负或不限
NAs	缺失值计数

6.3 数据类型检测与手动覆盖

glmnetUI 在导入时会自动检测数据类型，能够识别数值型、整型和日期型的列。看起来像日期的字符列（符合常见日期格式模式）会被归类为 Date。

您可以通过修改 **Type** 下拉框来覆盖任何检测结果。更改类型会影响该列在模型矩阵中的编码方式。

文本列、TRUE/FALSE 列和 R 因子列会自动按分类变量处理。数值列只有在您勾选其 **Factor** 复选框时才会变成因子——glmnetUI 绝不会根据数值的取值范围推断其为分类变量，因此像卫生间数量这样的离散数值预测变量除非您另行指定，否则保持连续。Factor 设置通过 localStorage 持久保存。

6.4 预期符号与约束的施加

对每个预测变量，**Sign** 下拉框用于指定您预期其系数为正、为负还是不限方向。拟合之后：

- 符号与预期不符的系数会在系数表中被标记出来
- 符号警示会显示在 Summary 选项卡中

当 glmnet 参数（第 7 章）中启用了 **Enforce Sign Constraints**（施加符号约束）复选框时，模型会被强制产生与预期符号一致的系数。这是通过 glmnet 中的 upper.limits 与 lower.limits 实现的。

6.5 特殊列类型参考

在估价与市场模式下，**Special** 下拉框提供以下选项：

Weighting (加权):

- **weight** ——观测权重列（所有模式下均可用；只允许指定一列；权重为 0 的行不参与拟合）

Date & Time Types (日期与时间类型):

- **contract_date** ——触发依据基准日自动计算 **sale_age**
- **listing_date** ——在计算在市天数时作为备用来源
- **dom** ——标识在市天数 (Days on Market) 列
- **sale_age** ——指定一个已有的成交时距列，这样任何一列都可以充当成交时距，而不必让 glmnetUI 从 **contract_date** 计算

Monetary Types (货币类型):

- **concessions** ——标识成交让利（卖方补贴、买方激励等）

Transaction Types (交易类型):

- **sale_type** ——标识成交 / 交易类型列

Size & Location Types (面积与位置类型):

- **latitude** ——数值自动四舍五入到小数点后 3 位
- **longitude** ——与纬度作相同的取整处理
- **area** ——市场区域或街区标识（例如 **area_id**、**neighborhood**、**market_area**）
- **living_area** ——启用按平方英尺计的残差计算 (**residual_sf** 与 **cqa_sf**)
- **lot_size** ——宗地面积列
- **site_dimensions** ——与地块面积归为一组（例如“75x120”）

Age Types (年龄类型):

- **actual_age** ——物业年龄列
- **effective_age** ——物业的有效年龄

Display Types (仅显示类型):

- **display_only** ——该列会包含在 Excel 导出结果中，但完全不参与模型拟合。可用于地址字段、MLS 编号、地块编号或其他参考数据。可以有多列采用此指定。

提示：被指定为 *display_only* 的列仍保留在数据集中并出现在 Excel 导出结果里，但完全不进入预测变量表。可用于 ID 列、地址或其他参考字段。

7 参数选择

侧边栏的第 5 节——**glmnet Call Parameters** (glmnet 调用参数)——提供弹性网络模型的全部配置选项。每个参数旁都有一个蓝色帮助图标 (?), 悬停可看到说明。

7.1 Alpha (混合参数)

alpha 参数控制正则化的类型:

Alpha 取值	类型	行为
0	岭回归	按比例收缩所有系数; 绝不会把任何系数置为零
1	Lasso	可以把系数置为恰好为零 (变量选择)
0 至 1	弹性网络	岭回归与 lasso 的混合

有两种 alpha 选择方法:

- **Fixed** (固定)——通过滑块设定单一 alpha 值 (默认: 1.0)
- **Grid Search** (网格搜索)——测试多个 alpha 值, 选出 CV 误差最小的那个。可配置取值范围与取值个数。

7.2 Lambda 的选择

lambda 控制正则化的强度:

方法	说明
交叉验证	推荐。使用 k 折 CV 测试多个 lambda 值。
手动	若您有充分理由, 可输入指定的 lambda 值。

使用交叉验证时, 有两种 lambda 取法:

- **lambda.1se** (默认)——在最小值一个标准误范围内的最大 lambda。产生更简洁、正则化更强的模型。估价工作推荐使用。
- **lambda.min**——使交叉验证误差最小的 lambda。产生拟合最好但可能更复杂的模型。

Number of CV Folds (CV 折数) 参数控制交叉验证时数据的划分方式 (默认: 10 折)。

7.3 Family (分布族)

为您的响应变量选择分布族:

分布族	适用情形
gaussian	连续型响应 (例如成交价格)。最常用。
binomial	二值结果 (例如已售 / 未售)。
poisson	计数数据 (例如成交笔数)。

7.4 Standardize (标准化)

勾选时 (默认), 所有预测变量在拟合前都会被缩放为均值 0、标准差 1。这确保惩罚对所有预测变量一视同仁, 而不受其原始量纲影响。系数会按原始量纲返回。通常应保持开启。

7.5 符号约束

启用 **Enforce Sign Constraints** 后，系数会被约束为与变量表（第 6 章）中设定的预期符号一致：

- “positive”（正）符号强制系数 ≥ 0
- “negative”（负）符号强制系数 ≤ 0
- 设为“either”（不限）的变量不受约束

这是通过 `glmnet()` 中的 `upper.limits` 与 `lower.limits` 参数实现的。

7.6 松弛 Lasso

常规 lasso 对变量选择和系数估计使用同一个惩罚，这可能把重要系数过度收缩到零。松弛 lasso (relaxed lasso) 把这两步分开：

1. 首先，通过 lasso 选择变量（由惩罚决定哪些系数得以保留）
2. 然后，对保留下来的变量以更小的惩罚或不加惩罚重新拟合，得到偏差更小的估计

Gamma 滑块控制松弛的程度：

Gamma	效果
0	完全松弛（OLS 重新拟合，完全不作收缩）
1	不作松弛（与常规 glmnet 相同）
介于其间	在松弛拟合与带惩罚拟合之间取折中

在使用交叉验证时，gamma 会被自动选取以获得最优表现。

7.7 随机种子

Random seed（随机种子）文本框预填了一个随机整数，用于使交叉验证的折分配可复现——拟合前会调用 `set.seed()`，所用的种子会显示在 **Glmnet Output** 选项卡以及拟合状态消息中。每次拟合之后会自动生成一个新种子（使用系统时间，而非 R 的随机数发生器）。当前文件最近使用的 5 个种子会以可点击链接的形式列出，便于快速调回；载入新文件时种子历史会重置。

7.8 交互项矩阵

一个可交互的上三角矩阵让您控制哪些变量对之间允许产生交互作用。每个单元格内有一个复选框——勾选表示交互项 $x1:x2$ 会被纳入模型矩阵。

- **Allow All**（全部允许）——勾选所有交互变量对
- **Clear All**（全部清除）——取消勾选所有变量对（新文件的默认状态是全部不勾选）
- 点击变量名——切换该变量的所有交互项

交互项的选择会按输入文件分别保存到浏览器 `localStorage`，并在您重新载入同一文件时恢复。

只要启用了任何交互项，导出的报告中就会包含一张 **Enabled Interactions**（已启用交互项）矩阵（两个坐标轴都是预测变量，纳入 $x1:x2$ 乘积项之处打勾），以横向页面呈现，与 earthUI 的 **Allowed Interactions** 矩阵相对应。请注意，lasso 仍可能把某个已启用项的系数收缩为零。

提示：当预测变量很多时，lasso 会自动把不需要的交互项归零。先用 “Allow All” 起步，再让模型自行选择，是一种合理的做法。

警告

启用 glmnet 交互项会产生在法庭或审计场合难以解释和说明的调整值。请谨慎使用交互项，并核实所得调整值是否合理。

7.9 禁止进入主效应（仅作交互项）

在交互项矩阵中右键点击某个变量名，即可禁止它以主效应的身份进入模型。变量名后会出现一个粗体的“1”标记。被禁止时：

- 该变量会从主公式项中移除
- 该变量仍可出现在交互项（x1:x2）中
- 再次右键点击即可解除禁止
- 该状态按输入文件通过 localStorage 持久保存

一个典型用例：在为应随物业面积按比例变化的时间调整建模时，禁止 sale_age 进入主效应，只允许它与 living_area 产生交互。这样得到的是 sale_age:living_area 交互项，而不是一个平坦的 sale_age 系数——即时间增值率随房屋面积而异。

7.10 高级参数

glmnet 调用参数底部有一个可折叠的“Advanced”（高级）小节，其中是一些附加设置。它们采用合理的默认值，但仍然可见，以便所有模型设置都能为法庭或审计留下书面记录。

- **Lambda Min Ratio** ——正则化路径中最小 lambda 与最大 lambda 之比。默认值：0.00001。对于小数据集（观测数 <50），取较大的值（例如 0.01）可能表现更好。
- **Number of Lambda Values** ——路径中 lambda 值的个数。默认值：100。
- **CV Loss Metric** ——交叉验证所用的损失函数：MSE（gaussian 的默认值）、MAE（对离群值稳健）或 Deviance（偏差）。
- **Convergence Threshold** ——坐标下降的收敛阈值。默认值：1e-07。
- **Max Iterations** ——坐标下降的最大迭代次数。默认值：100,000。若出现收敛警告，可调大此值。
- **Fit Intercept** ——是否包含截距（常数）项。默认：开启。在估价 RCA 中，截距即“基准值”。

7.11 设置默认值

有三个选项控制载入文件时参数如何初始化：

- **Use last settings for input file** ——恢复上次处理此文件时所用的确切设置
- **Use default settings** ——应用您保存的全局默认值
- **glmnet defaults** ——重置为出厂设置（alpha = 1、CV lambda.1se、10 折、gaussian、开启标准化）

点击 **Save current as default**，即可把当前全部设置存为今后文件所用的全局默认值。

8 拟合模型

8.1 拟合按钮

侧边栏的第 6 节包含 **Fit Glmnet Model** (拟合 Glmnet 模型) 按钮。点击它即以您当前的配置运行模型。

8.2 拟合过程中发生了什么

当您点击 Fit 时, glmnetUI 会:

1. 准备数据——剔除含 NA 的行, 施加权重 (若已指定), 排除标的行 (估价模式下), 并把因子变量编码为哑变量列。
2. 构建模型矩阵——生成设计矩阵, 含交互项矩阵中所有被允许的交互项。
3. 施加约束——若启用了符号约束, 则为系数设定 `upper.limits` 与 `lower.limits`。
4. 运行交叉验证——若选择了 CV, 则调用 `cv.glmnet()` 寻找最优 `lambda`。若启用了松弛 lasso, 则调用 `cv.glmnet(..., relax = TRUE)`。
5. 拟合最终模型——保存拟合好的模型、系数、预测值与残差。
6. 更新所有结果选项卡——填充 Summary、Coefficients、Equation、Variable Importance、Contributions、ANOVA 与 Diagnostics。

Fit 按钮下方的状态消息会显示观测数量、被排除的行数以及选定的 `lambda`。

9 结果选项卡

9.1 Data Preview (数据预览)

以交互式 DataTable 显示导入的数据。在估价 / 市场模式下, 预览分为两张表: **Subject Property** (标的物业, 第 1 行) 与 **Comparable Sales** (可比成交实例, 第 2 行及以后)。

每个单元格都保持单行显示, 并按列宽截断, 因此较宽的自由文本字段 (如物业备注) 不会把行撑得很高。双击任一单元格可弹出窗口, 显示其完整、未截断的内容。

9.2 Equation (方程)

显示通过 MathJax 以 LaTeX 渲染的拟合模型方程。展示所有非零系数, 并采用规范的数学排版:

- 正系数以 + 号显示
- 负系数以 - 号显示
- 交互项显示为 $x_1 \times x_2$
- 数字带千位逗号, 并保留适当的小数位

9.3 Correlation (相关性)

一张热力图矩阵, 展示所有数值型预测变量与响应变量之间的 Pearson 相关系数。数据导入后立即可用——无需拟合模型。

- 颜色梯度: 蓝色 (-1) → 白色 (0) → 红色 (+1)
- 每个单元格中打印相关系数值
- 文字与坐标轴尺寸随变量数量自适应
- 有助于在拟合前识别多重共线性

9.4 Summary (摘要)

以卡片形式显示的模型拟合统计量:

- R^2 ——被解释方差的比例
- $\text{Adj } R^2$ ——按预测变量个数调整后的 R^2
- GR^2 ——广义 R^2 , 由 glmnet 在选定 lambda 处的偏差比得出
- $\text{CV } R^2$ ——交叉验证 R^2 (使用 CV 时)
- **RMSE** ——均方根误差
- **MAE** ——平均绝对误差

在估价 / 市场模式下, 还会显示以下卡片:

- **Median Ratio** —— (预测值 / 实际值) 的中位数
- **COD** ——离散系数 (Coefficient of Dispersion, 越低越好)
- **PRD** ——价格相关差异 (Price-Related Differential)

当训练 R^2 超出 $\text{CV } R^2$ 达 0.1 以上时, 会出现过拟合警示。

卡片下方有一张带符号警示的系数表, 为方便查阅而重复了 Coefficients 选项卡的内容。

9.5 Coefficients (系数)

一张列出所有非零系数的表, 显示:

- 变量名
- 系数值
- 预期符号 (来自变量配置)
- 若系数方向与预期不符, 则给出符号警示标记

9.6 Variable Importance (变量重要性)

标准化的系数量级：对每个预测变量计算 $|\beta| \times \text{sd}(x)$ ，因子变量则跨其哑变量列汇总。有两种显示模式：

- 交互式 **plotly** (安装了 **plotly** 软件包时)：水平条形图，悬停提示显示精确的重要性、系数与相对百分比。
- 静态 **ggplot2** (备用方案)：水平条形图，坐标轴标签带千位逗号。

图表下方有一张 **DataTable**，显示 **Variable**、**Importance**、**Coefficient** 与 **Relative %**。

9.7 Contributions (贡献)

逐变量的偏效应图，展示每个预测变量对模型预测值的贡献。请从下拉框中选择一个变量。

对数值型预测变量：一张该变量的 x 值对其贡献 ($\beta \times x$) 的散点图，附一条红色线段和一个斜率标注。斜率标注显示每单位的边际效应 (例如 $+1,234.56/\text{unit}$)，对于纬度 / 经度这类取值范围很小的变量，单位会自适应调整 (例如 $+0.12/0.001$)。副标题显示截距 (基准) 值。

对因子变量：一张箱线图，展示各因子水平上贡献值的分布。

对双变量交互项：三幅上下叠放的可视化图——联合贡献的三维曲面图 (可用时为交互式 **plotly**，否则为静态 **persp()**)、按贡献着色的散点图，以及跨两个变量取值范围的平均贡献热力图。

9.8 ANOVA (方差分析)

方差分解表，为每个预测变量显示：

- **SS** ——平方和： $\sum (\text{contribution}_j - \overline{\text{contribution}_j})^2$
- **% of Model SS** ——占模型总平方和的百分比
- **Coefficient** ——该预测变量的主导系数

表中含一行截距和一行 **TOTAL MODEL** (模型合计)。交互项 (名称中含“:”者) 作为单独的组显示。

9.9 Diagnostics (诊断)

四幅诊断图，采用大字号 (基准字号 16pt)、15 个坐标轴刻度，标签带千位逗号 (不使用科学计数法)：

1. 系数路径 (Coefficient Path) ——展示随着 $\log(\lambda)$ 增大，所有系数如何变化。曲线按变量着色。有助于直观理解正则化路径与变量选择过程。
2. **CV 误差** (CV Error) ——交叉验证误差随 $\log(\lambda)$ 的变化。散点为 CV 误差均值；误差棒为 ± 1 个标准误。虚线竖线标出 $\lambda_{\min}$ (蓝色) 与 λ_{1se} (绿色)。
3. 实际值对预测值 (Actual vs Predicted) ——观测值与预测值的散点图。散点应聚集在 45 度虚线参照线附近。散布越宽，说明预测误差越大。
4. 残差对拟合值 (Residuals vs Fitted) ——残差与拟合值的散点图。应呈现围绕零线 (虚线) 的随机散布。若出现某种模式，则提示模型设定有误 (例如非线性、异方差)。

每幅图都有一个 **Download PNG** 按钮，可按 150 DPI 保存。

9.10 Glmnet Output (Glmnet 输出)

原始模型输出，显示：所用随机种子、模型打印结果、选定的 λ 、 $\lambda_{\min}$ / λ_{1se} 、 γ (若启用松弛 **lasso**)，以及完整的系数向量。

9.11 Report (报告)

配置报告字段（估价师姓名、物业地址、报告日期、档案编号），并可直接从该选项卡导出为 Word (.docx) 或 PDF。完整的基于 Quarto 的报告工作流程（HTML、Word 或 PDF）见第 13 章。

10 下载数据

拟合之后，**Download Output (Excel)**（下载输出，Excel）按钮（侧边栏第 7 节，在所有用途模式下均可用）会把含预测值与诊断信息的 Excel 文件导出到项目的输出文件夹。该小节标题在估价与市场模式下为“Download Estimated Sale Prices & Residuals”（下载估计成交价格与残差），在通用模式下为“Download Estimated Target Variable(s) & Residuals”（下载估计目标变量与残差）。

10.1 输出列

列	说明
est_<target>	模型预测值（例如 est_sale_price）
residual	实际值 - 预测值
cqa	状况-品质-吸引力分数（Condition-Quality-Appeal，0-10 分制）
residual_sf	残差 / 居住面积（若已指定 living_area）
cqa_sf	依据 residual_sf 排名计算出的 CQA
<var>_contribution	各变量对预测值的贡献
basis	截距值（所有行相同）

成功完成后，下载按钮上会出现一个白色对勾。

10.2 CQA 分数

CQA 把每一行的残差与其余各行相比较，按 0-10 分制排名：

- **CQA 高（~9-10）**：成交价远高于预测值——很可能状况 / 品质 / 吸引力更优
- **CQA 低（~0-1）**：成交价远低于预测值——很可能状况较差或属不良资产处置
- **CQA ~5**：接近中位数

在估价 / 市场模式下，若指定了 living_area 列，各行会按 residual_sf 降序排列。

11 RCA 计算与下载

RCA (Residual Constraint Approach, 残差约束法) 工作流程在 **Appraisal** (估价) 与 **Market Area Analysis** (市场区域分析) 模式下、模型拟合之后可用。后续的销售比较表格 (第 12 章) 仅限估价模式。

11.1 打开 RCA 对话框

点击侧边栏第 8 节的 **Calculate RCA Adjustments & Download** 按钮。然后选择：

- 分数类型：**CQA** 或 **CQA per SF** (每平方英尺 CQA)。若选择“CQA per SF”，标的的残差会按居住面积缩放。
- 输入标的的 CQA 分数 (0.00–9.99，默认 5.00)

11.2 CQA 分数插值

1. 对各可比实例的 CQA 分数与残差进行排序
2. 用线性插值把您给定的 CQA 值映射到一个残差
3. 标的价值 = 模型预测值 + 插值所得残差

11.3 输出列

列	说明
subject_value	模型预测值 + 插值所得残差
<var>_adjustment	标的贡献 – 可比实例贡献
residual_adjustment	标的残差 – 可比实例残差
net_adjustments	所有调整值之和
gross_adjustments	各调整值绝对值之和
adjusted_sale_price	可比实例成交价格 + 净调整额

计算成功后，按钮上会出现一个白色对勾。

12 销售比较表格

销售比较表格 (Sales Comparison Grid) 仅在 **Appraisal** (估价) 模式下、计算完 RCA 调整值 (第 8 步) 之后可用。它会生成一份格式化的 Excel 工作簿, 可直接编入估价报告。

12.1 可比实例选择对话框

点击侧边栏第 9 节的 **Generate Sales Grid & Download** 按钮, 会弹出一个模态对话框, 列出 RCA 输出中的所有可比成交实例。可比实例分为两组:

- **Recommended comps** (推荐的可比实例, 预先勾选): 毛调整额 < 成交价格的 25%, 按成交时距排序 (最新在前)。
- **Other comps** (其他可比实例, 默认不勾选): 其余的可比实例, 按成交时距排序。

最多可选 30 个可比实例, 然后点击 **Generate Sales Grid** 生成工作簿。

12.2 工作簿版式

生成的工作簿最多含 10 个工作表, 每个工作表并排放置 3 个可比实例:

- 表头行: 标的与可比实例的地址、成交价格、成交日期以及在市天数
- 分组行: 位置 (纬度、经度、区域)、宗地面积 (地块面积、宗地尺寸) 与年龄 (实际年龄、有效年龄), 带组标题与子项目
- 模型变量行: 每个非零模型系数一行, 显示标的贡献、各可比实例贡献以及调整百分比
- 剩余特征行: 留空且未锁定的单元格, 供估价师填写模型未涵盖的特征 (例如状况、品质、景观)
- 汇总行: 净调整百分比、毛调整百分比与调整后成交价格, 均带有实时的 Excel 公式

12.3 工作表保护与剩余特征单元格

每个工作表都设有密码保护以防误改, 但剩余特征的取值单元格是明确解锁的。这样估价师既能填入模型之外的特征取值, 又能保留由公式驱动的调整计算。

13 下载报告

13.1 两步式 Quarto 工作流程

从侧边栏导出报告采用两步式的 Quarto 工作流程：

1. **Generate Quarto Report** (生成 Quarto 报告, 侧边栏第 10 节) —— 把一个自包含的 Quarto 报告包 (.qmd 源文件加上图形、report_data.rds 与 reference.docx) 写入项目的 glmnet 输出文件夹。
2. **Convert Quarto Report** (转换 Quarto 报告, 侧边栏第 11 节) —— 把 .qmd 文件渲染为您选择的输出格式。路径默认指向最近生成的报告包, 但您也可以浏览选择任何一个 .qmd。

13.2 报告格式

转换步骤中有三种格式可选：

- **HTML** —— 自包含的 HTML 文档。在所有平台上都能使用, 除 pandoc 外无需额外软件。推荐的默认选项。
- **Word (.docx)** —— 格式化文档, 便于编辑与分发。在所有平台上均可使用。
- **PDF** —— 通过 Quarto 配合 LaTeX 生成。纸张尺寸遵循区域设置 (Letter 或 A4)。需要安装 LaTeX (参见系统要求附录)。若未检测到 LaTeX, 格式复选框中会自动隐藏 PDF 选项。

有 Quarto 可用时使用 Quarto 生成报告, 否则回退到 rmarkdown。Report 选项卡另外还提供直接导出为 Word 和导出为 PDF 的按钮。

13.3 报告内容

报告涵盖所有选项卡的内容：

- 数据集说明 (文件名、观测数、响应变量、预测变量、用途)
- 模型设定 (alpha、lambda、family、standardize、是否松弛、非零系数)
- 结果摘要 (R^2 、调整 R^2 、广义 R^2 、CV R^2 、RMSE、MAE)
- 模型方程 (以 LaTeX 渲染)
- 系数表
- 变量重要性 (条形图与表格)
- 与 Contributions 选项卡一致的逐变量贡献图：
 - 数值型预测变量：带线性拟合线的散点图
 - 因子变量：按水平分组的箱线图
 - 交互项：按贡献着色的散点图、平均贡献热力图, 以及静态三维曲面快照 (persp)
- Enabled Interactions (已启用交互项) 矩阵 (启用了任何交互项时) —— 两个坐标轴都是预测变量, 纳入 $x_1:x_2$ 乘积项之处打勾, 以横向页面呈现
- 相关系数矩阵热力图
- ANOVA 分解表
- 诊断图：
 - 系数路径
 - 交叉验证误差
 - 实际值对预测值
 - 残差对拟合值
 - 正态 Q-Q 图
- 模型输出

所有坐标轴标签与颜色图例中的数字均带千位逗号（不使用科学计数法）。
报告生成成功后，报告按钮上会出现一个白色对勾；渲染过程中会显示已用时间计时器。

14 与 earthUI 的比较

earth 应用程序及其 glmnet 后处理模块同属一套回归建模工具。它们共用相同的数据格式、特殊列类型、RCA 工作流程和演示数据集，但使用不同的建模引擎。

14.1 主要差异

特性	glmnetUI	earthUI
方法	弹性网络 (glmnet)	MARS/Earth (earth)
变量选择	经由 lasso 惩罚自动进行	经由前向 / 后向剪枝自动进行
系数关系	仅限线性效应	分段线性 (铰链函数)
交互项	经由交互项矩阵两两组合	最高 3 阶
系数符号	可通过约束强制施加	无法直接强制施加
松弛拟合	可使用松弛 lasso	不适用
Alpha 调优	可用网格搜索	不适用 (无混合参数)
销售比较表格	可用 (第 9 步)	可用 (第 8 步)
g 函数图	不适用	归组项可视化
报告格式	HTML、Word、PDF	HTML、Word、PDF

14.2 共有的功能

两个工具都提供：

- CSV 与 Excel 导入，并把列名转换为 snake_case
- 支持 31 个国家的区域设置，含 CSV 分隔符、小数点标记与纸张尺寸
- 特殊列指定 (contract_date、living_area、area、latitude、longitude 等)
- 依据基准日自动计算 sale_age
- 深色 / 浅色模式切换
- 通过 localStorage 持久保存设置
- 相关性热力图、变量重要性、贡献图、ANOVA、诊断图
- CQA 评分与 RCA 调整工作流程
- 带可比实例选择和格式化 Excel 输出的销售比较表格
- 兼容同一批演示数据集 (Appraisal_1.csv)

14.3 何时该用哪一个

- 使用 **glmnetUI**，当您想要一个简洁、可解释、并能自动选择变量的线性模型时。当预测变量很多而您希望由模型挑出最重要的那些时，弹性网络尤其合适。
- 使用 **earthUI**，当您预期存在非线性关系（例如居住面积对价格的影响在不同面积区间有所不同），或需要用铰链函数刻画阈值效应时。earth 模型还能刻画更高阶的交互作用（2 阶和 3 阶）。此外，earthUI 还提供一项基于 Prolog/DCG 的实验性备注抽取功能，这在 glmnetUI 中没有对应项。
- 两者并用，以在同一份数据上比较线性模型与非线性模型。若两者给出的 R^2 相近，出于可解释性考虑，可能更宜采用较简洁的 glmnet 模型。

15 演示数据集: Appraisal_1.csv

15.1 说明

演示数据集随 glmnetUI 一同发布（并与 earthUI 共用）。可用以下方式定位：

```
demo_file <- system.file("extdata", "Appraisal_1.csv", package = "glmnetUI")
```

该文件包含 1,502 笔住宅成交记录（外加第 1 行的 1 处标的物业），取自一份模拟的 MLS 导出文件。数据代表一个多区域市场中的独栋住宅成交，涵盖不同的物业面积、房龄与位置。

这并非真实数据，但以北加州一个真实街区为原型。所有身份识别信息均已更改或移除。

15.2 各列

列	类型	特殊类型	说明
weight	numeric	weight	观测权重（0 = 不参与拟合）
id	numeric	display_only	内部记录 ID
property_id	numeric	display_only	MLS 物业标识
listing_id	character	display_only	MLS 挂牌编号
parcel_number	character	display_only	县估税官地块编号（APN）
street_address	character	display_only	物业地址
city_name	character	display_only	城市
postal_code	character	display_only	邮政编码
county_name	character	display_only	县
contract_date	Date	contract_date	成交合同日期（用于计算 sale_age）
sale_age	numeric	—	从合同日期到基准日的天数
sale_price	numeric	（目标变量）	成交价格——响应变量
living_sqft	numeric	living_area	总居住面积（平方英尺）
beds_total	integer	—	卧室数量
baths_total	numeric	—	卫生间总数（例如 2.5 = 2 全卫 + 1 半卫）
lot_size	numeric	lot_size	地块面积（平方英尺）
area_id	integer	area	MLS 区域标识
age	numeric	actual_age	物业年龄（年）
latitude	numeric	latitude	纬度（入模前四舍五入到小数点后 3 位）
longitude	numeric	longitude	经度（入模前四舍五入到小数点后 3 位）
garage_spaces	integer	—	车库车位数
days_on_market	integer	dom	在市天数
listing_date	Date	listing_date	挂牌日期
sale_concessions	numeric	concessions	卖方让利

15.3 建议的快速上手流程

1. 启动 glmnet 例程: `earthUI::launch_glmnet()`
2. 新建一个用途为 **For Appraisal** 的项目（第 1 节）
3. 把 Appraisal_1.csv 放入项目的输入文件夹，并在第 2 节中选中它（必要时点击 **Re-fresh**）
4. 选择 `sale_price` 作为目标变量
5. 按上表所示指定各列的特殊类型
6. 纳入以下预测变量: `sale_age`、`living_sqft`、`baths_total`、`lot_size`、`area_id`（作为因子）、`age`、`latitude`、`longitude`、`garage_spaces`
7. 保持 $\alpha = 1$ (Lasso)，使用 CV 配 `lambda.1se`

8. 点击 **Fit Glmnet Model**
9. 查看 Summary、Coefficients 与 Diagnostics 选项卡
10. 下载输出 (第 7 步), 查看 CQA 排名
11. 以约 ~5.00 的 CQA 分数计算 RCA 调整值 (第 8 步)
12. 用推荐的可比实例生成销售比较表格 (第 9 步)

16 系统要求与故障排查

16.1 受支持的平台

glmnetUI 可在 **macOS**、**Windows** 与 **Linux** 上运行。强烈建议使用 RStudio Desktop (2023.06 及以上版本) ——它自带生成报告所需的 **pandoc**。

16.2 可选依赖

- **LaTeX** (TinyTeX、MiKTeX 或 MacTeX): 仅用于生成 PDF 报告。若未检测到, PDF 选项会自动隐藏。安装方式: `tinytex::install_tinytex()`
- **sysfonts + showtext**: 为图形提供 Roboto Condensed 字体。若不可用或处于离线状态, 会自动回退到系统无衬线字体。
- **earth** ($\geq 5.3.0$): 仅在使用 earthUI 导入通道时需要。

16.3 各平台说明

macOS: 问题最少。安装 TinyTeX 以生成 PDF: `tinytex::install_tinytex()`

Windows: 与 RStudio 配合良好。公司配发或权限受限的机器可能对临时目录有限制——应用程序会在控制台给出警告, 但仍会继续运行, 只是设置不能持久保存。

Linux: 可能需要安装系统库: `sudo apt install libcurl4-openssl-dev libssl-dev libxml2-dev libsqlite3-dev libfontconfig1-dev`

16.4 降级运行

缺失的组件	表现
没有 LaTeX	隐藏 PDF 选项。HTML 与 Word 仍可使用。
没有网络 / 字体加载失败	使用系统无衬线字体。控制台记录相应消息。
没有 RSQLite / 文件系统只读	设置不能持久保存。应用程序照常运行。
临时目录不可写	报告生成失败并给出明确错误。请把 TMPDIR 设为可写位置。

16.5 故障排查

“**PDF option not available**”(PDF 选项不可用): 在 R 中运行 `tinytex::install_tinytex()`, 然后重启应用程序。

“**Settings will not persist**”(设置无法保存): 检查用户数据目录的权限 (macOS: `~/Library/Application Support/R/glmnetUI/`, Linux: `~/.local/share/R/glmnetUI/`, Windows: `%APPDATA%/R/data/glmnetUI/`)。

端口 **7879** 已被占用: 运行 `lsof -ti:7879 | xargs kill(macOS/Linux)` 或 `Stop-Process -Id (Get-NetTCPConnection -LocalPort 7879).OwningProcess` (Windows PowerShell)。

参考文献

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software* 33.1, pp. 1–22. DOI: [10.18637/jss.v033.i01](https://doi.org/10.18637/jss.v033.i01).
- Simon, Noah et al. (2011). “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent”. In: *Journal of Statistical Software* 39.5, pp. 1–13. DOI: [10.18637/jss.v039.i05](https://doi.org/10.18637/jss.v039.i05).

Tay, J. Kenneth, Balasubramanian Narasimhan, and Trevor Hastie (2023). “Elastic Net Regularization Paths for All Generalized Linear Models”. In: *Journal of Statistical Software* 106.1, pp. 1–31. DOI: [10.18637/jss.v106.i01](https://doi.org/10.18637/jss.v106.i01).