大作业：房价预测问题

3170105743 李政达

目录

1 问题描述

对于一个房屋购买者，决定其心仪的房子的因素是多样的。例如，尽管房屋地下室到天花板的高度或者房屋与铁路的距离并不是一个特别重要的因素，但是亦有数据表明这些因素对于房价的影响力是高于房屋卧室的数量或是栅栏的颜色等因素的。

具体而言，本文试图通过已有的包含 79 个解释变量与的 1460 个观察值训练数据集[1]，建立房价关于其中若干解释变量的模型，并通过此模型对房价进行预测。

2 使用模型

本文考虑使用较为基本的线性回归模型，并在回归变量的选取上进行了一定的研究。

2.1 线性回归模型

在统计学中，线性回归是利用称为线性回归方程的最小二乘函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种函数是一个或多个称为回归系数的模型参数的线性组合。只有一个自变量的情况称为简单回归，大于一个自变量情况的叫做多元回归（multivariable linear regression）[2]。

2.2 自变量选取

自变量的选取是回归分析中很重要的环节之一。

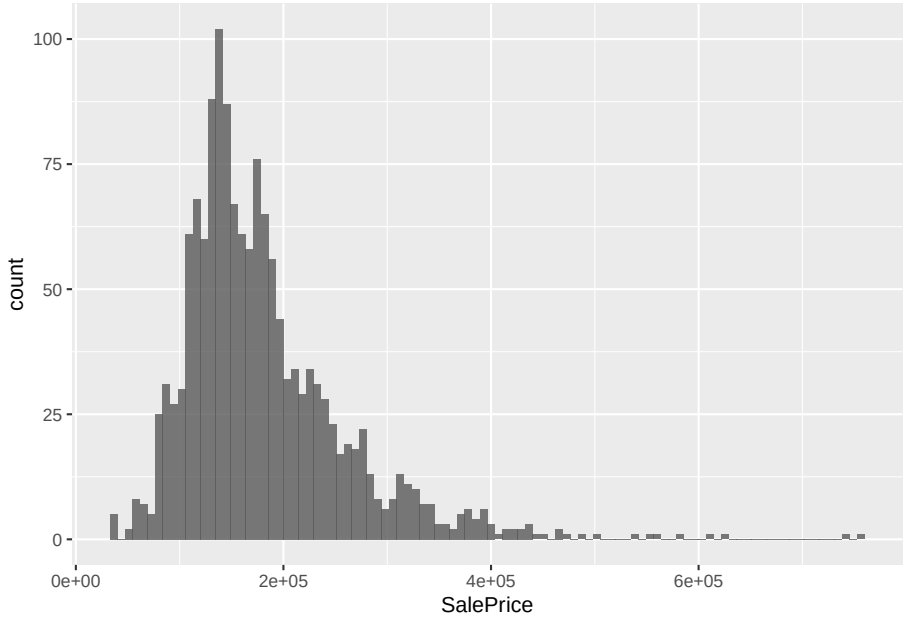
本文首先尝试了 kitchen sink 模型[3]。当回归的目标仅仅是预测并建模时，自变量对整个模型的决定程度并不是关键的考察因素。一个简单的想法是将所有可能的自变量全部作为预测变量，并通过多元回归的方法来建立模型。这种回归方式便是 kitchen sink 回归。

此外，本文亦试图对预测变量进行精简。相关系数是反映两组变量相关性的重要指标，一般而言，相关系数的绝对值越高，可认为两组变量相关程度较大。本文从相关系数入手，选取与房价相关系数绝对值较高的两组指标作为预测变量，建立了线性回归模型。

3 数据分析

3.1 房价的分布

首先考虑房价的大致分布情况。使用 `ggplot` 包提供的函数可绘制其分布的直方图：



可见房价的分布呈正偏态，多数房价位于 2×10^5 内，而仅有极少的房价超过 4×10^5。

3.2 数据的清洗

本文主要考虑有具体数值的影响因素，因而首先对非数值项进行了筛选。

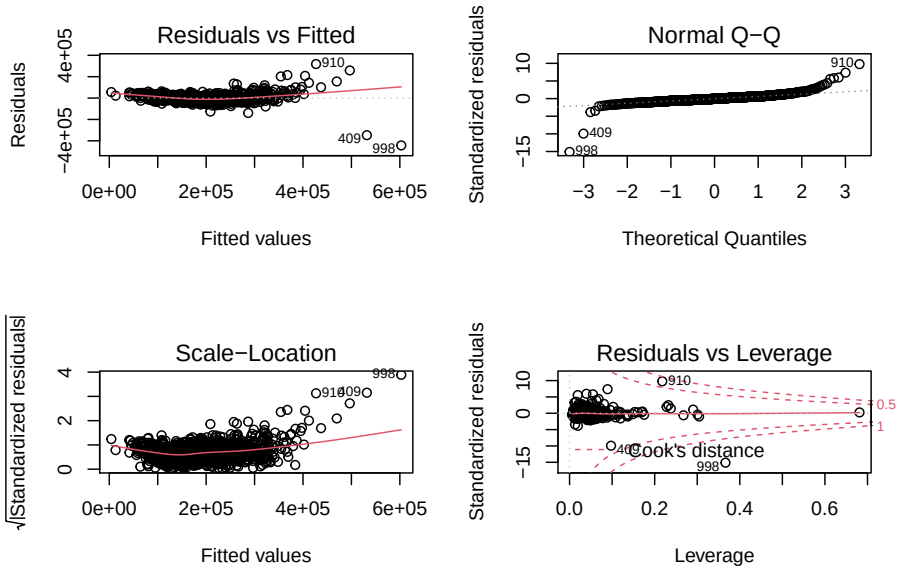
此外，在训练集中有 NA 项的数据观察值不能进行训练，因此考虑去除所有含有 NA 项的观察值。

清洗后的数据共含有 36 个解释变量与 1121 个有效观察值。

3.3 建立回归模型

3.3.1 加入较多自变量

使用 kitchen sink model 得到结果如下：

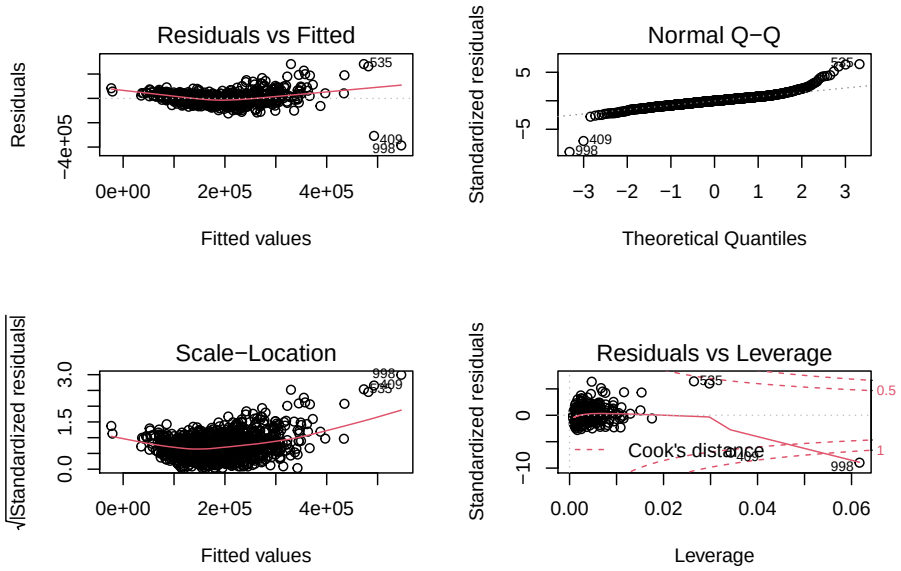


在回归的系数中存在 NA 项，需要在预测时将其去除；修正后的 R^2 值为 0.8036，可见该模型尽管使用了较多的自变量，但拟合效果仍较为一般。

3.3.2 通过相关系数寻找自变量

为了寻找与 SalePrice 最为相关的变量，首先应计算各变量间的相关系数，尤其需要考虑各个变量与 SalePrice 的相关系数。

由上图可看出，OverallQual 与 GrLivArea 两项与 SalePrice 有较高的相关系数，可考虑以这两项为自变量进行线性回归，结果如下：



虽然其修正后 R^2 项低于 kitchen sink model，但由于该模型自变量较少，计算较为简单，且不含有 NA 项，也有一定应用价值。

4 结果评价

4.1 评价指标

均方根误差（root-mean-square deviation、root-mean-square error、RMSD、RMSE）是一种常用的测量数值之间差异的度量[4]，可以通过计算准确值与预测值的 RMSE 来判断预测的好坏。

另一方面，对于房价较高的观察值，其预测值可能会有较大的误差；而对于房价较低的观察值的预测值的误差也可能较低。为了消除其影响，实际上进行评价的指标是房价准确值与预测值的对数值的 RMSE，该指标越低说明预测模型越优秀。

为便于调用，程序将此评价指标实现为了一个函数。

4.2 各模型的表现

kitchen sink 模型最终指标较低，为 0.1928299，这与上文中较低的修正后 R^2 值是相符的，因此考虑使用该模型预测测试集样本。

在使用该模型预测测试集时，有部分数据显示为 NA（附件 1：original_sub.csv），这与部分观察值某项指标的缺失有关。考虑使用训练集的均值取代 NA 作为最终提交结果（附件 2：adjusted_sub.csv），在比赛网站上得到最终评价指标为 0.38025，该数值虽高于训练集的指标，但由于训练集的准确性一般高于测试集，因此这仍是一个可以接受的结果。

4.3 改进思路

至本文完成前，该模型总排名为 4551 名，并不是一个特别理想的结果。考虑改进思路如下：

1. 引入对非数值变量的分析：本文在数据处理过程中直接去除全部非数值变量，而实际上非数值变量蕴含着非常重要的信息，若合理分析其含义可取得一定改进；
2. 引入更加先进的模型：目前本文使用模型为线性回归模型，较为基础；而先进的模型如 random forest 模型、gradient boosting 模型等会取得较好的效果。

5 参考文献

[1] Kaggle, House Prices: Advanced Regression Techniques[EB/OL], https://www.kaggle.com/c/house-prices-advanced-regression-techniques

[2] 维基百科, 线性回归 [EB/OL], https://zh.wikipedia.org/wiki/%E7%B7%9A%E6%80%A7%E5%9B%9E%E6%AD%B8

[3] Tony Fischetti, Data Analysis with R, 2015.12

[4] Wikipedia, Root-mean-square deviation[EB/OL], https://en.wikipedia.org/wiki/Root-mean-square_deviation