大作业：房价预测问题

3170105743 李政达

目录

1 问题描述 **1**

2 使用模型 **2**

2.1 线性回归模型 2

2.2 自变量选取 2

3 数据分析 **2**

3.1 房价的分布 2

3.2 数据的清洗 3

3.3 建立回归模型 3

4 结果评价 **7**

4.1 评价指标 7

4.2 各模型的表现 7

4.3 改进思路 7

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。

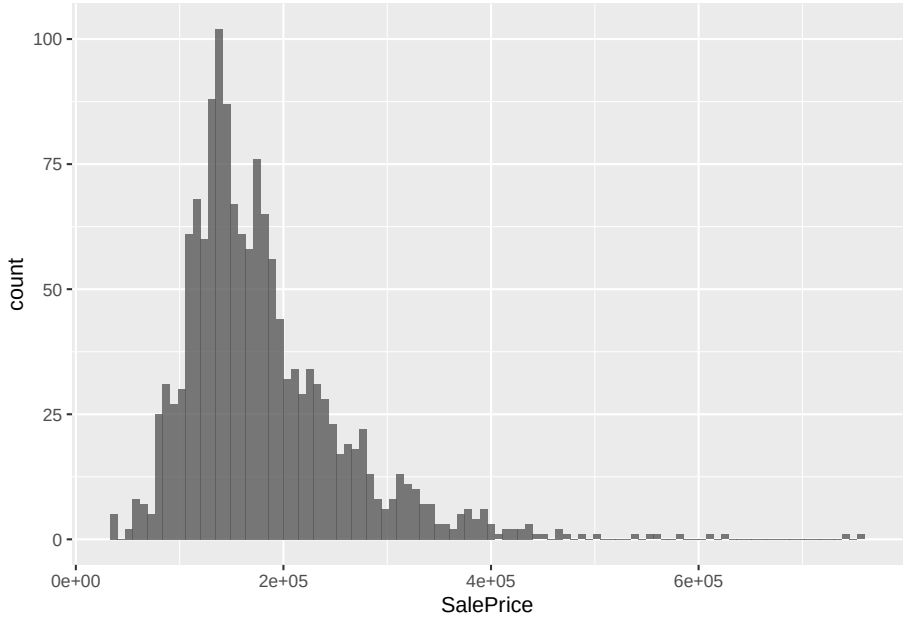
本文首先尝试了 kitchen sink 模型[3]。当回归的目标仅仅是预测并建模时，自变量对整个模型的决定程度并不是关键的考察因素。一个简单的想法是将所有可能的自变量全部作为预测变量，并通过多元回归的方法来建立模型。这种回归方式便是 kitchen sink 回归。

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

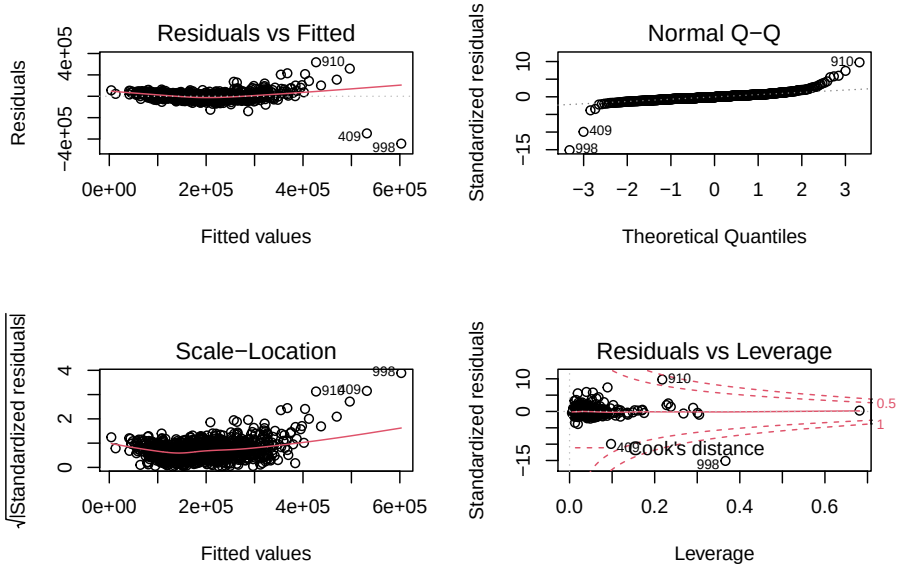
3.3.1 加入较多自变量

使用 kitchen sink model 得到结果如下：

```
##
## Call:
## lm(formula = SalePrice ~ . - Id, data = train_values)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -442865  -16873   -2581   14998  318042
##
## Coefficients: (2 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.232e+05  1.701e+06  -0.190 0.849317
## MSSubClass    -2.005e+02  3.449e+01  -5.814 8.03e-09 ***
## LotFrontage   -1.161e+02  6.124e+01  -1.896 0.058203 .
## LotArea        5.454e-01  1.573e-01   3.466 0.000548 ***
## OverallQual    1.870e+04  1.478e+03  12.646  < 2e-16 ***
## OverallCond    5.227e+03  1.367e+03   3.824 0.000139 ***
## YearBuilt      3.170e+02  8.762e+01   3.617 0.000311 ***
## YearRemodAdd   1.206e+02  8.661e+01   1.392 0.164174
## MasVnrArea     3.160e+01  7.006e+00   4.511 7.15e-06 ***
## BsmtFinSF1     1.739e+01  5.835e+00   2.980 0.002947 **
## BsmtFinSF2     8.362e+00  8.763e+00   0.954 0.340205
## BsmtUnfSF      5.006e+00  5.275e+00   0.949 0.342890
## TotalBsmtSF           NA         NA      NA       NA
## `1stFlrSF`     4.591e+01  7.356e+00   6.241 6.21e-10 ***
## `2ndFlrSF`     4.668e+01  6.099e+00   7.654 4.28e-14 ***
## LowQualFinSF   3.415e+01  2.788e+01   1.225 0.220788
## GrLivArea             NA         NA      NA       NA
## BsmtFullBath   8.980e+03  3.194e+03   2.812 0.005018 **
## BsmtHalfBath   2.490e+03  5.071e+03   0.491 0.623487
## FullBath       5.390e+03  3.529e+03   1.527 0.126941
## HalfBath      -1.119e+03  3.320e+03  -0.337 0.736244
## BedroomAbvGr  -1.023e+04  2.154e+03  -4.750 2.30e-06 ***
## KitchenAbvGr  -2.193e+04  6.704e+03  -3.271 0.001105 **
## TotRmsAbvGrd   5.440e+03  1.486e+03   3.661 0.000263 ***
## Fireplaces     4.375e+03  2.188e+03   2.000 0.045793 *
## GarageYrBlt   -4.914e+01  9.093e+01  -0.540 0.589011
## GarageCars     1.679e+04  3.487e+03   4.815 1.68e-06 ***
```

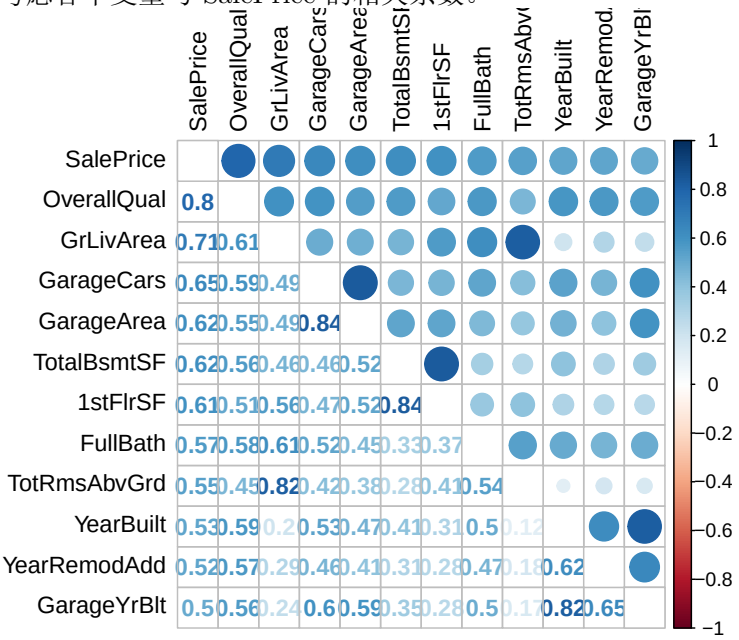
```
## GarageArea      6.488e+00  1.211e+01   0.536 0.592338
## WoodDeckSF      2.155e+01  1.002e+01   2.151 0.031713 *
## OpenPorchSF    -2.315e+00  1.948e+01  -0.119 0.905404
## EnclosedPorch   7.233e+00  2.061e+01   0.351 0.725733
## `3SsnPorch`     3.458e+01  3.749e+01   0.922 0.356593
## ScreenPorch     5.797e+01  2.040e+01   2.842 0.004572 **
## PoolArea       -6.126e+01  2.984e+01  -2.053 0.040326 *
## MiscVal        -3.850e+00  6.955e+00  -0.554 0.579980
## MoSold         -2.240e+02  4.227e+02  -0.530 0.596213
## YrSold         -2.536e+02  8.454e+02  -0.300 0.764216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36790 on 1086 degrees of freedom
## Multiple R-squared:  0.8095, Adjusted R-squared:  0.8036
## F-statistic: 135.7 on 34 and 1086 DF,  p-value: < 2.2e-16
```



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

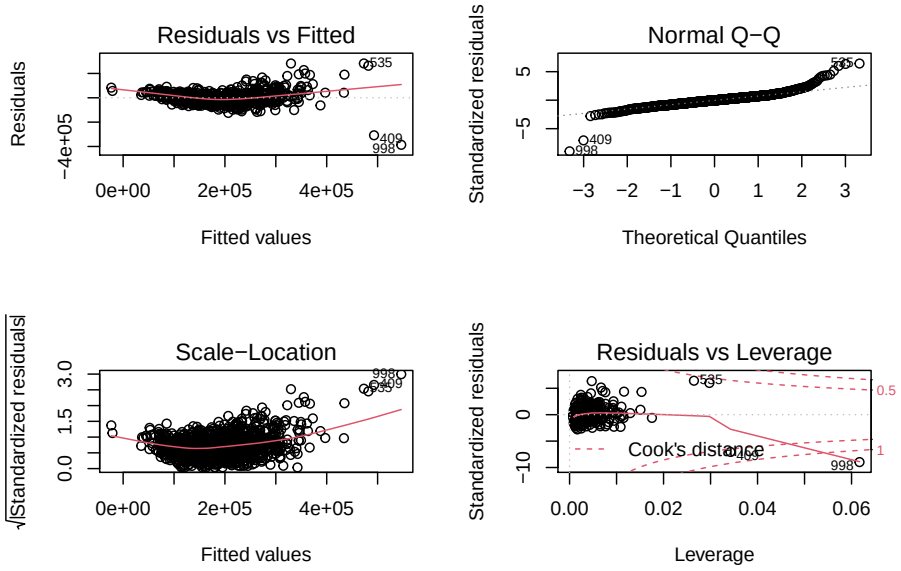
为了寻找与 SalePrice 最为相关的变量，首先应计算各变量间的相关系数，尤其需要考虑各个变量与 SalePrice 的相关系数。



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：

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea, data = train_values)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -386454  -24291   -1002   21150  281971
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.180e+05  6.130e+03  -19.25   <2e-16 ***
## OverallQual  3.520e+04  1.211e+03   29.07   <2e-16 ***
## GrLivArea    5.537e+01  3.193e+00   17.34   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44450 on 1118 degrees of freedom
## Multiple R-squared:  0.7136, Adjusted R-squared:  0.7131
## F-statistic:  1393 on 2 and 1118 DF,  p-value: < 2.2e-16
```



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

4.2 各模型的表现

4.3 改进思路