通过房屋特征预测房价-基于 R 语言的回归分析

王昊 *(学号: 201821061107)* 指导教师: 段小刚

2019 年 11 月 30 日

目录

摘要

这个报告主要通过房屋特征预测房价探究了 R 语言中对于特征变量的处理，缺失值的补充等。随后使用了一种特征选择的方法, 并建立了最简单的线性模型来预测房价。

背景

给你提供一份有有关于美国 Lowa 市 Ames 的房价数据，其中包含 79 个 feature，提供 train 和 test 样本，要求对 test 中的房价进行预测。

研究问题

基于竞赛方所提供的爱荷华州埃姆斯的住宅数据信息，预测每间房屋的销售价格。

对数据的描述性分析

数据的网址为：https://www.kaggle.com/c/house-prices-advanced-regression-techniques

竞赛给了已经成交的近 1500 座房子的 80 个特征，然后让我们根据这些特征来预测房子的销售价格。数据集包含的特征字段相当多，除了地段、面积、层数等基本信息外，还有诸如地下室、离街道的距离、房屋的外墙材料等在国内完全不会关心的特征。

数据熟悉

在动手构造模型之前，先熟悉一下数据的缺失和分布情况。首先下载训练数据和测试数据，放在目录下，然后合并训练数据和测试数据。其中 SalePrice 就是这次要预测的房价字段。

读取训练数据集和测试数据集

```r
train <- read.csv("D:/我的坚果云/GitHubCode/R/R_Stat_Course/R_Homework/R_LR/house-prices-advanced-r
test <- read.csv("D:/我的坚果云/GitHubCode/R/R_Stat_Course/R_Homework/R_LR/house-prices-advanced-re
```

合并两个训练集

```r
test$SalePrice <- NA    # 把我们要预测的因变量置为空（未知量）
all <- rbind(train, test)
```

查看所有变量的结构

```r
str(all)
```

```
## 'data.frame':    2919 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...
##  $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...
##  $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...
##  $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...
##  $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...
##  $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...
##  $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...
##  $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
##  $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...
##  $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...
##  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...
##  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...
##  $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
##  $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
##  $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...
##  $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...
##  $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
##  $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...
##  $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
##  $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ..
```

其中，变量主要分为两类，一类为数字类型，一类为因子类型。

获取数据中 factor 变量的个数

```
res <- sapply(all, class)
# 获取 all 这个数据集所有数据的类型, 相当于 class(all(1,2,3,.....n))
#(apply 可以实现对数据的循环、分组、过滤、类型控制等)

table(res) #table 函数对应的就是统计学中的列联表, 是一种记录频数的方法。
```

```
## res
##  factor integer 
##      43      38
```

由以上结果可知，该数据集一共 81 个变量、2919 个记录，其中 43 个因子变量，38 个数字变量。

特征处理

从上面的变量取值情况可以看到数据集中有很多变量存在缺失值，所以第一步我们要处理缺失值。

统计所有变量的缺失值

首先按照各变量中的缺失值所占比例排序。

```
# 检查 all 里面所有的变量是否为缺省值, 如果是的话, 放入 res
res <- sapply(all, function(x)  sum(is.na(x)) )
```

按照缺失率排序

```
miss <- sort(res, decreasing=T)

# 打印: 缺失数据对应的变量 + 该变量缺失的数据
miss[miss>0]
```

```
##       PoolQC  MiscFeature        Alley        Fence    SalePrice  FireplaceQu 
##         2909         2814         2721         2348         1459         1420
##  LotFrontage  GarageYrBlt GarageFinish   GarageQual   GarageCond   GarageType
##          486          159          159          159          159          157
##     BsmtCond BsmtExposure     BsmtQual BsmtFinType2 BsmtFinType1   MasVnrType
##           82           82           81           80           79           24
##   MasVnrArea     MSZoning    Utilities BsmtFullBath BsmtHalfBath   Functional
##           23            4            2            2            2            2
##  Exterior1st  Exterior2nd   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF
##            1            1            1            1            1            1
##   Electrical  KitchenQual   GarageCars   GarageArea     SaleType
##            1            1            1            1            1
```

查看有缺失数据的变量的描述性统计量

```
summary(all[,names(miss)[miss>0]])
```

```
##   PoolQC     MiscFeature  Alley        Fence        SalePrice      FireplaceQu
##  Ex  :   4   Gar2:   5   Grvl: 120   GdPrv: 118   Min.   : 34900   Ex  :  43
##  Fa  :   2   Othr:   4   Pave:  78   GdWo : 112   1st Qu.:129975   Fa  :  74
##  Gd  :   4   Shed:  95   NA's:2721   MnPrv: 329   Median :163000   Gd  : 744
##  NA's:2909   TenC:   1               MnWw :  12   Mean   :180921   Po  :  46
##              NA's:2814               NA's :2348   3rd Qu.:214000   TA  : 592
##                                                   Max.   :755000   NA's:1420
##                                                   NA's   :1459
##   LotFrontage      GarageYrBlt    GarageFinish GarageQual  GarageCond
##  Min.   : 21.00   Min.   :1895   Fin : 719    Ex  :   3   Ex  :   3
##  1st Qu.: 59.00   1st Qu.:1960   RFn : 811    Fa  : 124   Fa  :  74
##  Median : 68.00   Median :1979   Unf :1230    Gd  :  24   Gd  :  15
##  Mean   : 69.31   Mean   :1978   NA's: 159    Po  :   5   Po  :  14
##  3rd Qu.: 80.00   3rd Qu.:2002                TA  :2604   TA  :2654
##  Max.   :313.00   Max.   :2207                NA's: 159   NA's: 159
##  NA's   :486      NA's   :159
##    GarageType   BsmtCond    BsmtExposure BsmtQual    BsmtFinType2 BsmtFinType1
##  2Types :  23   Fa  : 104   Av  : 418    Ex  : 258   ALQ :  52    ALQ :429
##  Attchd :1723   Gd  : 122   Gd  : 276    Fa  :  88   BLQ :  68    BLQ :269
##  Basment:  36   Po  :   5   Mn  : 239    Gd  :1209   GLQ :  34    GLQ :849
##  BuiltIn: 186   TA  :2606   No  :1904    TA  :1283   LwQ :  87    LwQ :154
##  CarPort:  15   NA's:  82   NA's:  82    NA's:  81   Rec : 105    Rec :288
##  Detchd : 779                                        Unf :2493    Unf :851
##  NA's   : 157                                        NA's:  80    NA's: 79
##    MasVnrType     MasVnrArea        MSZoning     Utilities     BsmtFullBath
##  BrkCmn :  25   Min.   :   0.0   C (all):  25   AllPub:2916   Min.   :0.0000
##  BrkFace: 879   1st Qu.:   0.0   FV     : 139   NoSeWa:   1   1st Qu.:0.0000
##  None   :1742   Median :   0.0   RH     :  26   NA's  :   2   Median :0.0000
##  Stone  : 249   Mean   : 102.2   RL     :2265                 Mean   :0.4299
##  NA's   :  24   3rd Qu.: 164.0   RM     : 460                 3rd Qu.:1.0000
##                 Max.   :1600.0   NA's   :   4                 Max.   :3.0000
##                 NA's   :23                                    NA's   :2
##   BsmtHalfBath       Functional    Exterior1st    Exterior2nd
##  Min.   :0.00000   Typ    :2717   VinylSd:1025   VinylSd:1014
##  1st Qu.:0.00000   Min2   :  70   MetalSd: 450   MetalSd: 447
##  Median :0.00000   Min1   :  65   HdBoard: 442   HdBoard: 406
##  Mean   :0.06136   Mod    :  35   Wd Sdng: 411   Wd Sdng: 391
##  3rd Qu.:0.00000   Maj1   :  19   Plywood: 221   Plywood: 270
##  Max.   :2.00000   (Other):  11   (Other): 369   (Other): 390
##  NA's   :2         NA's   :   2   NA's   :   1   NA's   :   1
##    BsmtFinSF1       BsmtFinSF2        BsmtUnfSF       TotalBsmtSF
##  Min.   :   0.0   Min.   :   0.00   Min.   :   0.0   Min.   :   0.0
##  1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.: 220.0   1st Qu.: 793.0
##  Median : 368.5   Median :   0.00   Median : 467.0   Median : 989.5
##  Mean   : 441.4   Mean   :  49.58   Mean   : 560.8   Mean   :1051.8
##  3rd Qu.: 733.0   3rd Qu.:   0.00   3rd Qu.: 805.5   3rd Qu.:1302.0
##  Max.   :5644.0   Max.   :1526.00   Max.   :2336.0   Max.   :6110.0
##  NA's   :1        NA's   :1         NA's   :1        NA's   :1
##  Electrical   KitchenQual   GarageCars      GarageArea        SaleType
##  FuseA: 188   Ex  : 205   Min.   :0.000   Min.   :   0.0   WD     :2525
##  FuseF:  50   Fa  :  70   1st Qu.:1.000   1st Qu.: 320.0   New    : 239
##  FuseP:   8   Gd  :1151   Median :2.000   Median : 480.0   COD    :  87
##  Mix  :   1   TA  :1492   Mean   :1.767   Mean   : 472.9   ConLD  :  26
##  SBrkr:2671   NA's:   1   3rd Qu.:2.000   3rd Qu.: 576.0   CWD    :  12
##  NA's :   1               Max.   :5.000   Max.   :1488.0   (Other):  29
##                           NA's   :1       NA's   :1        NA's   :   1
```

```
# 获取描述性统计量：最小值、最大值、四分位数和数值型变量的均值，因子向量和逻辑型向量：频数统计。
```

缺失数据的处理

直接删除存在大量缺失值的变量　PoolQC、MiscFeature、Alley、Fence、FireplaceQu 等变量缺失值比较多，是由于房子没有泳池、特殊的设施、旁边的小巷、篱笆、壁炉等这些特殊设施。

```
# 缺失量比较多，我们直接移除这几个变量。
Drop <- names(all) %in% c("PoolQC","MiscFeature","Alley","Fence","FireplaceQu") # 做逻辑判断，如果这
```

```r
all <- all[!Drop]    #all 只留下没被选中的那些变量
```

将 NA 作为新的一个因子 查看变量的描述文件可以知道，车库相关的 5 个变量 (GarageType、GarageYrBlt、GarageFinish、GarageQual、GarageCond) 也是由于房子没有车库而缺失。同理，BsmtExposure、BsmtFinType2、BsmtQual、BsmtCond、BsmtFinType1 这 5 个变量是关于地下室的，都是由于房子没有地下室而缺失。此类变量缺失的数量比较少，直接用 None 来替换缺失值, 代表这个房子不具备这个属性。

```r
# 将如下变量的 NA 值填充为 None
Garage <- c("GarageType","GarageQual","GarageCond","GarageFinish")
Bsmt <- c("BsmtExposure","BsmtFinType2","BsmtQual","BsmtCond","BsmtFinType1")
for (x in c(Garage, Bsmt) )
{
  all[[x]] <- factor( all[[x]], levels= c(levels(all[[x]]),c('None')))
  all[[x]][is.na(all[[x]])] <- "None"
}
```

其中 GarageYrBlt 为车库的年份，我们用房子的建造年份来替代.

```r
# 单独处理车库年份
all$GarageYrBlt[is.na(all$GarageYrBlt)] <- all$YearBuilt[is.na(all$GarageYrBlt)]
```

人工补齐缺失值 对剩下的变量我们依次查看其详细数据，可以分别如下处理。变量 LotFrontage 是房子到街道的距离这是一个数值变量，我们用中位数 Median 来补充。

```r
# 用中位数来填充
all$LotFrontage[is.na(all$LotFrontage)] <- median(all$LotFrontage, na.rm = T)
```

变量 MasVnrType 外墙装饰材料这个变量对价钱的影响应该不大，MasVnrType 中的 NA 用它本身的 None 来代替

```r
# 用 None 补充
all[["MasVnrType"]][is.na(all[["MasVnrType"]])] <- "None"
```

变量 MasVnrArea 外墙装饰材料的面积这个缺失值对应着 MasVnrType 的 None 值，应该将 NA 用 0 来替代

```r
# 用 0 补充
all[["MasVnrArea"]][is.na(all[["MasVnrArea"]])] <- 0
```

变量 Utilities 没有区分度，直接丢弃

```r
# 删除变量 Utilities
all$Utilities <- NULL
```

变量 BsmtFullBath BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF GarageCars GarageArea 则是由于不存在相应的设施而缺失，这些变量都是数字变量，所以都补充为 0 即可。

```
# 由于设施缺失，导致数量缺失，补充为 0
Param0 <- c("BsmtFullBath","BsmtHalfBath","BsmtFinSF1","BsmtFinSF2","BsmtUnfSF","TotalBsmtSF","Gar
for (x in Param0 )    all[[x]][is.na(all[[x]])] <- 0
```

变量 MSZoning,Functional,Exterior1st,Exterior2nd,KitchenQual,Electrical,SaleType 这些变量都是因子变量，并且只有几个缺失值，直接用最多的因子来代替

```
# 用最高频的因子来补充
Req <- c("MSZoning","Functional","Exterior1st","Exterior2nd","KitchenQual","Electrical","SaleType"
for (x in Req )    all[[x]][is.na(all[[x]])] <- levels(all[[x]])[which.max(table(all[[x]]))]
```

生成训练集

经过一系列的缺失值补齐之后，我们看到最后剩余 75 个变量，并且不存在缺失数据。我们通过 SalePrice 是否为 NA 来将数据集拆分为训练集和测试集，为后面的模型训练做准备。

```
# 通过 SalePrice 是否为空来区分训练集和测试集
train <- all[!is.na(all$SalePrice), ]
test <- all[is.na(all$SalePrice), ]
```

统计模型的具体形式

线性回归的最主要的问题就是自变量的选择。选择那些与最后预测的响应变量相关度比较高的特征变量是模型成功的第一步。变量选择有很多方法，其中最关键同时也是最直接的方法就是分析师根据业务场景人工筛选。我们首先尝试这种变量选择的方法，作为我们模型的第一步。

初步模型

这次题目给的自变量有很多，我们需要从中挑选对房价影响最大的变量。我们的思路是先人工挑选一些对房价影响比较重要的因素，然后再慢慢的添加新的变量来看是否会改变模型的精度。

以国内的房价为例，影响房价的因素主要是房子面积、房子所在的区域、小区等，房龄、房型（小高层、多层、别墅等）、特殊场景（地铁房、学区房等）、装修等也会影响价格。这个数据是美国的房屋信息，不过基本的影响因素应该差不多。

我们先来 j 建立简单的模型来，选择如下变量：

- LotArea 房子的面积
- Neighborhood 城市街区用来初步代替区域、小区
- Condition1 Condition2 附近的交通情况
- BldgType 房屋类型独栋别墅、联排别墅

- HouseStyle 房子的层数
- YearBuilt 房子建造的年份
- YearRemodAdd：房子的改造年份
- OverallQual：房子整体质量，考量材料和完成度
- OverallCond：房子整体条件

装修越好的房子价格越高。

查看各变量之间的相关性:

```
# 相关系数画图
#install.packages('corrgram')
#library(corrgram)
# <- c("LotArea","Neighborhood","BldgType","HouseStyle","YearBuilt","YearRemodAdd","OverallQual","

#corrgram(train[,sel], order=TRUE, lower.panel=panel.shade, upper.panel=panel.pie, text.panel=pane
```

模型训练

先用我们挑选的变量来建立一个 lm 模型，作为我们的 base 模型

```
# 通过人工选择的变量来构造一个公式
fm.base <- SalePrice ~ LotArea + Neighborhood + BldgType + HouseStyle + YearBuilt + YearRemodAdd +

# 训练模型
lm.base <- lm(fm.base, train)

# 查看模型概要
summary(lm.base)
```

```
##
## Call:
## lm(formula = fm.base, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -208970  -20882   -2917   15544  351199
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.455e+06  1.850e+05  -7.862 7.42e-15 ***
## LotArea              1.084e+00  1.156e-01   9.375  < 2e-16 ***
## NeighborhoodBlueste -1.068e+03  2.953e+04  -0.036 0.971141
## NeighborhoodBrDale  -1.440e+04  1.518e+04  -0.949 0.342806
## NeighborhoodBrkSide -1.876e+04  1.278e+04  -1.468 0.142460
## NeighborhoodClearCr -2.352e+03  1.332e+04  -0.177 0.859842
## NeighborhoodCollgCr -2.917e+04  1.086e+04  -2.685 0.007335 **
## NeighborhoodCrawfor  1.747e+04  1.246e+04   1.402 0.161225
## NeighborhoodEdwards -2.813e+04  1.165e+04  -2.414 0.015924 *
## NeighborhoodGilbert -4.030e+04  1.157e+04  -3.484 0.000508 ***
## NeighborhoodIDOTRR  -3.357e+04  1.343e+04  -2.499 0.012570 *
## NeighborhoodMeadowV  1.338e+04  1.446e+04   0.925 0.354867
## NeighborhoodMitchel -2.819e+04  1.196e+04  -2.356 0.018617 *
## NeighborhoodNAmes   -2.202e+04  1.130e+04  -1.950 0.051426 .
## NeighborhoodNoRidge  6.105e+04  1.226e+04   4.980 7.13e-07 ***
## NeighborhoodNPkVill  6.340e+03  1.650e+04   0.384 0.700928
## NeighborhoodNridgHt  4.876e+04  1.104e+04   4.417 1.08e-05 ***
## NeighborhoodNWAmes  -2.126e+04  1.166e+04  -1.823 0.068457 .
## NeighborhoodOldTown -2.915e+04  1.243e+04  -2.344 0.019194 *
## NeighborhoodSawyer  -2.575e+04  1.188e+04  -2.168 0.030350 *
## NeighborhoodSawyerW -2.224e+04  1.154e+04  -1.927 0.054234 .
## NeighborhoodSomerst -1.228e+04  1.093e+04  -1.123 0.261764
## NeighborhoodStoneBr  5.984e+04  1.249e+04   4.790 1.84e-06 ***
## NeighborhoodSWISU   -2.365e+04  1.433e+04  -1.651 0.099024 .
## NeighborhoodTimber  -1.326e+04  1.236e+04  -1.073 0.283489
## NeighborhoodVeenker  2.303e+04  1.555e+04   1.481 0.138905
## BldgType2fmCon       1.230e+03  7.413e+03   0.166 0.868218
## BldgTypeDuplex      -7.231e+02  5.831e+03  -0.124 0.901330
## BldgTypeTwnhs       -6.675e+04  7.811e+03  -8.546  < 2e-16 ***
## BldgTypeTwnhsE      -4.916e+04  4.892e+03 -10.049  < 2e-16 ***
## HouseStyle1.5Unf    -2.835e+04  1.102e+04  -2.573 0.010184 *
## HouseStyle1Story    -3.981e+03  3.977e+03  -1.001 0.316972
## HouseStyle2.5Fin     5.328e+04  1.472e+04   3.619 0.000306 ***
## HouseStyle2.5Unf    -4.606e+03  1.250e+04  -0.368 0.712613
## HouseStyle2Story     4.069e+03  4.205e+03   0.968 0.333393
## HouseStyleSFoyer    -1.173e+04  7.791e+03  -1.505 0.132424
## HouseStyleSLvl      -6.438e+03  6.197e+03  -1.039 0.299077
## YearBuilt            4.285e+02  8.428e+01   5.084 4.20e-07 ***
## YearRemodAdd         3.114e+02  7.505e+01   4.149 3.53e-05 ***
## OverallQual          2.849e+04  1.187e+03  24.010  < 2e-16 ***
## OverallCond          1.613e+03  1.150e+03   1.402 0.161035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38880 on 1419 degrees of freedom
## Multiple R-squared:  0.7671, Adjusted R-squared:  0.7605
## F-statistic: 116.8 on 40 and 1419 DF,  p-value: < 2.2e-16
```

模型的结果分析与解释

结果解读

针对模型 summary 之后的结果，我们简单解读一下输出结果含义

残差统计量

> Residuals: Min 1Q Median 3Q Max -208970 -20882 -2917 15544 351199
>
> 线性回归的计算基于一些假设，其中一个假设就是误差符合相互独立、均值为 0 的正态分布。

从本例可以看出这个残差的中位数为负数，数据整体左偏。其中的 1Q 和 3Q 是第一四分位（first quartile）和第三四分位（third quartile）。残差的最大值和最小值附近对应的记录则可能是异常值。

由于残差代表预测值和真实值之间的差别，也就是说最大值 351199 表示我们预测的最大误差有 35 万美元之多。

仅仅从残差的五数概括上看不出什么关键信息，后续可以通过残差图来检查残差是否符合正态分布的趋势。

回归系数

> Coefficients: Estimate Std. Error t value Pr(>|t|)
> (Intercept) -1.455e+06 1.850e+05 -7.862 7.42e-15 LotArea 1.084e+00 1.156e-01 9.375 < 2e-16 NeighborhoodBlueste -1.068e+03 2.953e+04 -0.036 0.971141 NeighborhoodBrDale -1.440e+04 1.518e+04 -0.949 0.342806 NeighborhoodBrkSide -1.876e+04 1.278e+04 -1.468 0.142460 … Signif. codes: 0 ‘’ ***0.001*** ‘’ *0.01* “ 0.05 ‘.’ 0.1 ‘ ’ 1

线性回归拟合完成后得出的回归系数并不是准确的值，而是对于真实回归系数的估计值。

既然是估计值则必然存在误差，上述结果中的 > Estimate 表示回归系数的估计 >Std. Error 表示回归系数的标准误差 >t value 表示假设此回归系数为 0 时的 T 检验值 >Pr(>|t|) 则是上述假设成立的置信度 p-value

> P-value 越小则说明假设（假设回归系数为 0）越不容易出现，反过来就是此变量的回归系数不为 0 的几率越大，故此变量在整个回归拟合中作用越显著。一般用置信度 0.05 作为判断依据。

- 最后的三颗星表示此变量显著，星号越多越显著，最多三个。
- 最后一行 Signif. codes 标识着显著标识编码当 P-value 小于 0.001 时三颗星，小于 0.01 时两颗星，大于 0.05 则认为不太显著。

R^2 和 Adjusted R^2

> Multiple R-squared: 0.7671, Adjusted R-squared: 0.7605
>
> R-squared（判定系数，coefficient of determination）也称为模型拟合的确定系数，取值 0~1 之间，越接近 1，表明模型的因变量对响应变量 y 的解释能力越强。Adjusted R-squared: 当自变量个数增加时，尽管有的自变量与 y 的线性关系不显著，R square 也会增大。Adjusted R square 增加了对变量增多的惩罚，故我们以 Adjusted R square 为判断模型好坏的基本标准。

本例中 Adjusted R-squared: 0.7605 表示响应变量有 76% 的方差被此模型解释了。

模型整体的 F 检验

> F-statistic: 116.8 on 40 and 1419 DF, p-value: $<$ 2.2e-16。

F 统计量用来检验模型是否显著。

假设模型所有的回归系数均为 0，即该模型是不显著的。对此假设做 F 检验，在 p-value 的置信度下拒绝了此假设，则模型为显著的。

在本例中 p-value: $<$ 2.2e-16，远远低于 0.05，所以模型是显著的。

变量选择 - 人工筛选 模型 lm.base 的 Adjusted R-squared: 是 0.7605。从第一个模型的结果看到变量 OverallCond 并不显著，所以我们去掉变量 OverallCond，重新进行拟合。拟合结果 Adjusted R-squared: 0.7603 和之前相差不大，并且所有的变量都显著。故我们将第一个模型定为：

```r
# 初步决定的 lm.base 模型的变量
fm.base <- SalePrice ~ LotArea + Neighborhood + BldgType + HouseStyle + YearBuilt + YearRemodAdd +

# 训练模型
lm.base <- lm(fm.base, train)
```

模型出来了，我们把计算结果写入文件。

```r
# 用 lm.base 模型预测
lm.pred <- predict(lm.base, test)

# 写出结果文件
res <- data.frame(Id = test$Id, SalePrice = lm.pred)
write.csv(res, file = "D:/我的坚果云/GitHubCode/R/R_Stat_Course/R_Homework/R_LR/house-prices-advanc
```

参考文献

[1] Regression Analysis by Example . Samprit Chatterjee, Ali S. Hadi

[2] R in Action, Second Edition . Robert I. Kabacoff

[3] Machine Learning With R Cookbook . Yu-Wei, Chiu (David Chiu)

[4] Glmnet Vignette：http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html#qs

附录 (代码)

(将以文件形式附在压缩包内)