**实习报告**

By miao 2017.09.01

**概述**

住宅是由各种特征要素所组成的，因此这些特征要素也决定了房产的价格高低，通过对大量房产数据进行分析，最终得到住宅中每个特征要素对于房产价格的影响权重系数，从而得到特征价格模型。

［注］特征要素即下文的特征变量／解释变量／自变量。

成交价（预测价）即下文的被解释变量／因变量。

1. **数据预处理**

* **数据表的重新制作（处理缺失值并将房产表格与小区表格链接）**

**1.房产部分**

1. 筛选房产特征变量（akey值）：
2. 将筛选出来的akey值中aval不存在的r\_id剔除
3. 检查各个aval的数量是否均衡：

筛选过程中发现房屋用途为普通住宅的数据占极大比例，因此为了保证预测结果的准确性，模型建立的前提设置为房屋用途为**普通住宅。**

1. 同时因各类因素印象筛选出半年以内的房产成交数据来进行训练和预测。时间有效性

**2.小区部分**

与房产类似，特征变量选择小区名称，小区均价与物业费用三项。

**3.链接表格**

通过相同的akey小区名称将房产表格与小区表格链接起来，形成完整的每一套房产对有与之相对应的小区信息。左连接 右连接

* **表格中数据的检查与清洗**

1. **错误数据排查**

**对各个特征变量（数值型如建筑面积 小区均价等）进行散点图分析，将散点图较为分散的数据进行人工排查是否为错误数据。**

最终得到数据表apart\_a

1. **在建立回归模型前仍需对数据表apart\_a中各akey值对应的aval值进行处理**

**由于特征变量的属性分为两种：定量变量与定性变量（即数值型与文字描述型）如建筑面积与装修程度。因此处理方法据此分为两类。**

1. **定量变量：去除数字后的单位等等，将其转换为可以进行计算的 数值常量。**
2. **定性变量：通过转化为因子，将其转化为虚拟变量*（此处详见material文件夹中tips文档中有关虚拟变量的描述）（R语言对于虚拟变量的设定非常方便 详见tips文档中的各种链接）***

**二．建立回归模型**

1.将数据表apart\_a划分为训练集与测试集（大概70%和30%）这个比例相对合理 训练集要占多数，测试集合少数供后期的模型的修正调整，和准确率的判断。好像是书上写的比例。

2.利用训练集数据首先建立最基本的多元线性回归模型

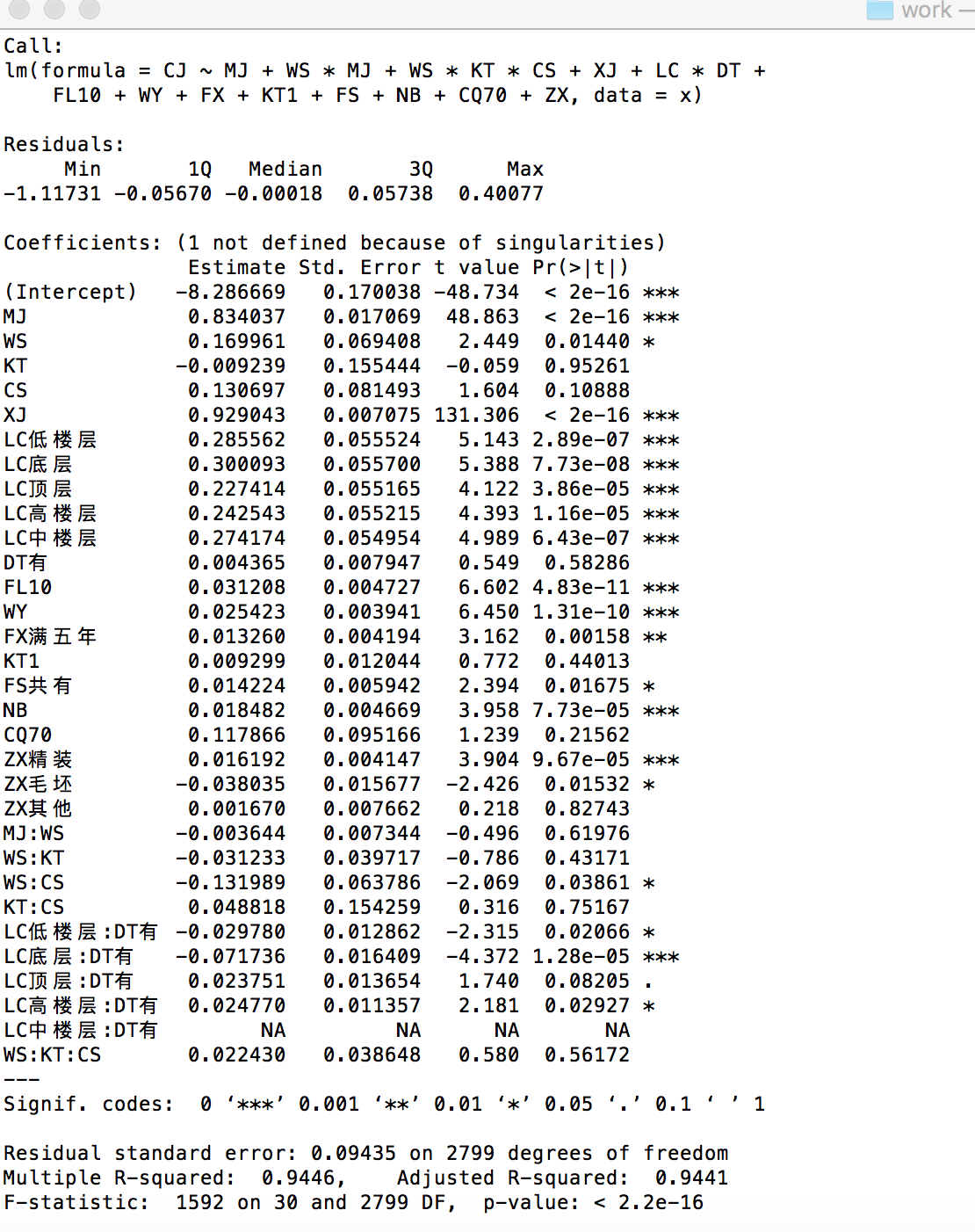
（根据该模型检验各个变量之间的相关性，共线性，以及解释变量与被解释变量之间的相关程度，筛选和过滤出最终变量，然后再进行后续的回归模型的性能提高等操作。）

3.得到训练模型结果 summary（）

**三．模型的检验与修正**

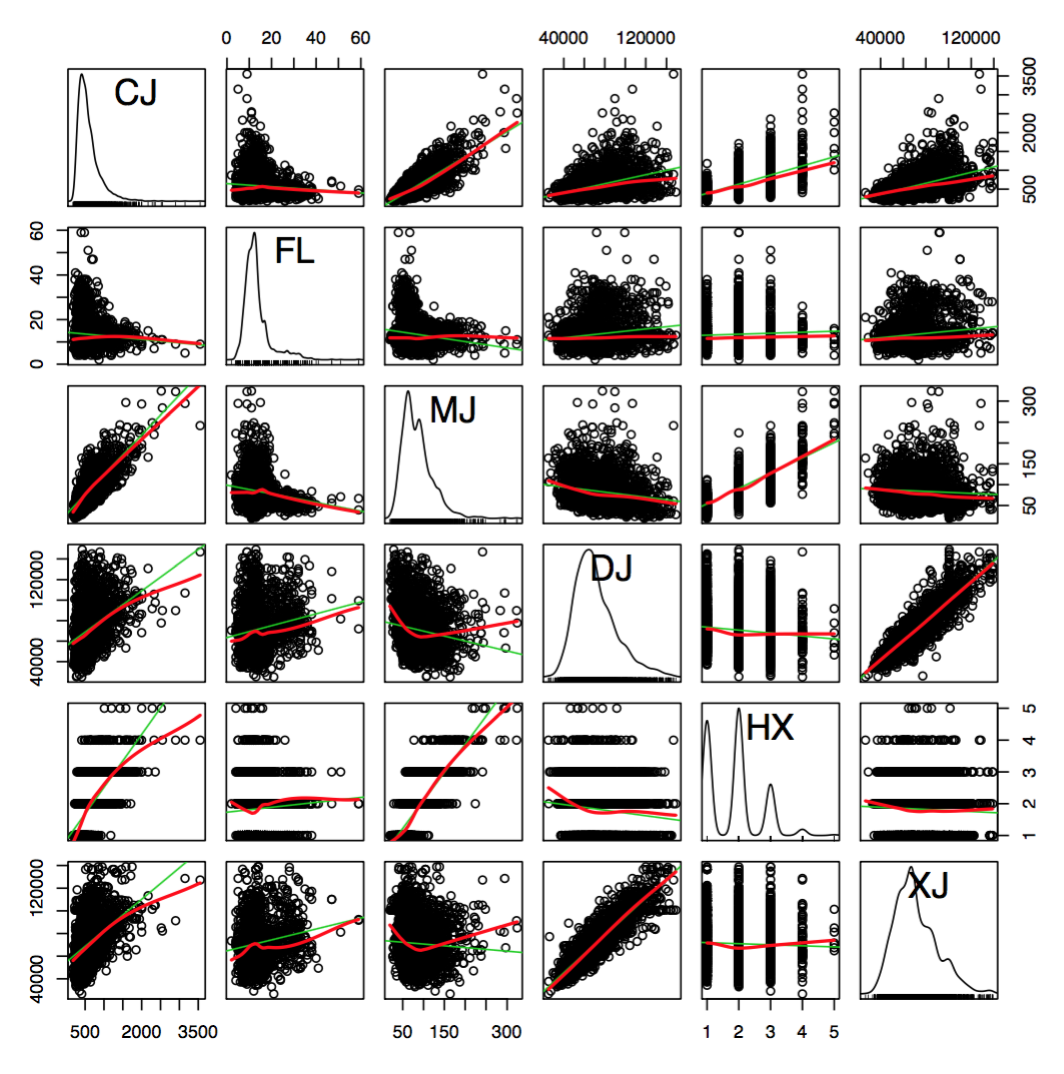
**1.经济学意义：**对应的特征变量系数符号是否与预期的符号一致（如果不一致需要考虑是否需要修正该特征变量或者是否前期的数据有错误）

**2.统计学检验：**



1. 拟合优度：决定系数：Multiple R2以及adjusted R2，越接近1说明模型的解释性越高
2. 方程显著性检验——F值：给定5%显著水平 看P值是否小于5%
3. 变量显著性检验——t值：给定5%显著水平 看P值是否小于5%
4. 看变量之间的相关程度：通过cor（）函数得到相关系数矩阵

或者通过散点图矩阵可视化相关性。（下图是之前随手做的 不过不知如何解读）



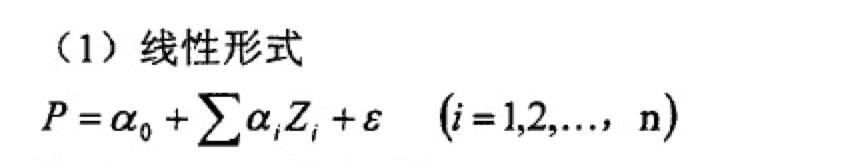
1. 计量经济学检验：多重共线性检验，vif

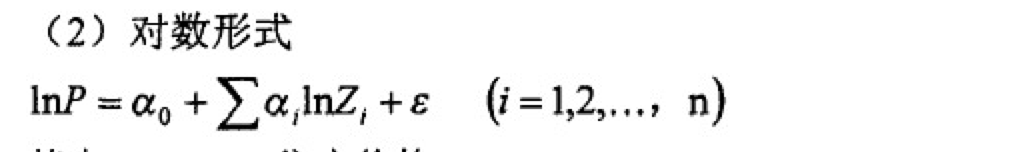
若各特征变量的vif值均<10，解释变量的容忍度>0.1,则不存在共线性。

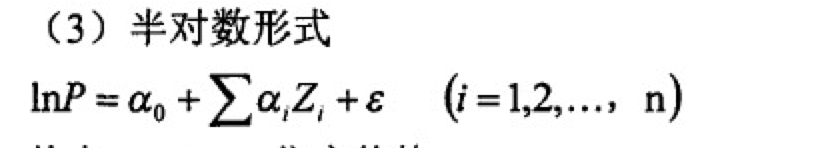
**四．提高模型性能（该部分可以详见机器学习与R语言中的P130页）**

**1.模型的设定：**

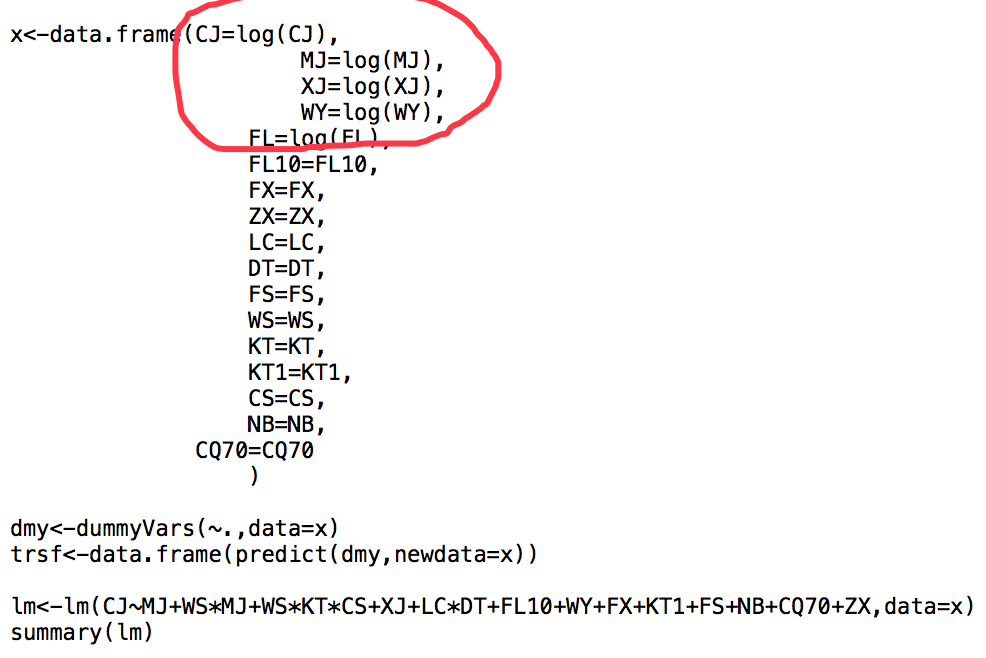
添加非线性关系







这一步则对应论文中的三个公式，其实后两个所谓的半对数与对数公式相比较于线性形式，个人认为就是添加了非线性关系，因为各个变量之间的影响并不一定是纯粹的线性恒定的，所以在模型中添加一些非线性关系非常重要。

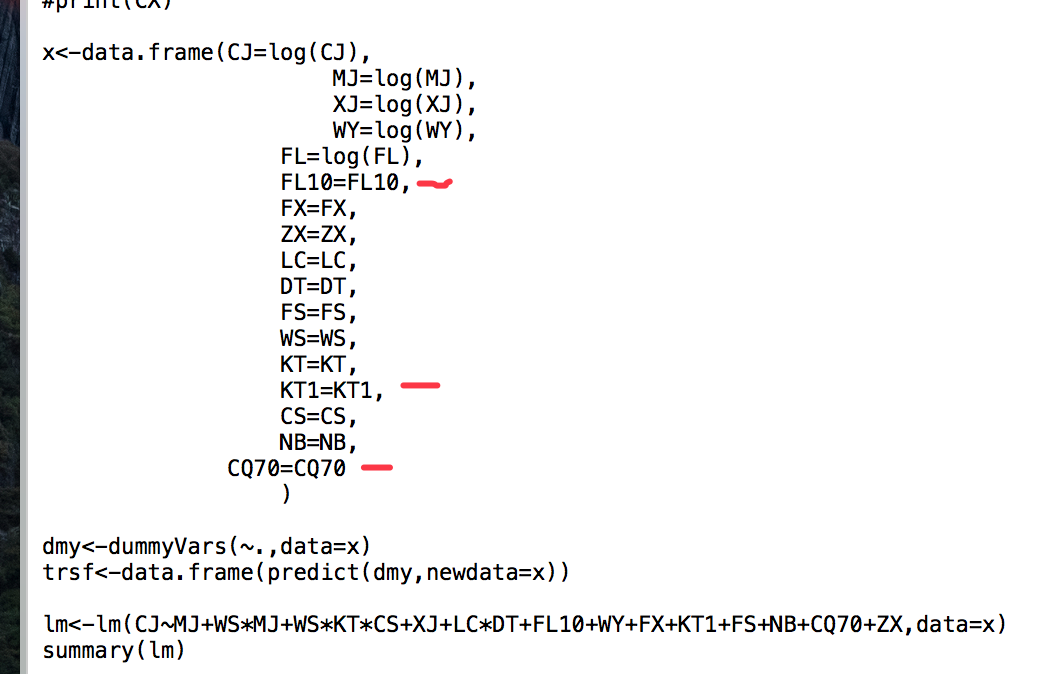


本次模型中，加入了一些对数形式，对成交价（CJ）与面积（MJ）等取了对数形式。加了对数形式之后 训练好的模型再次用测试集进行检验，发现准确率确实挺高了。

**2.将数值型变量转换为二进制指标（量变引起质变？？）**

一个特征的影响不是积累的，而是当特征的取值达到一个给定的阈值才会产生影响。

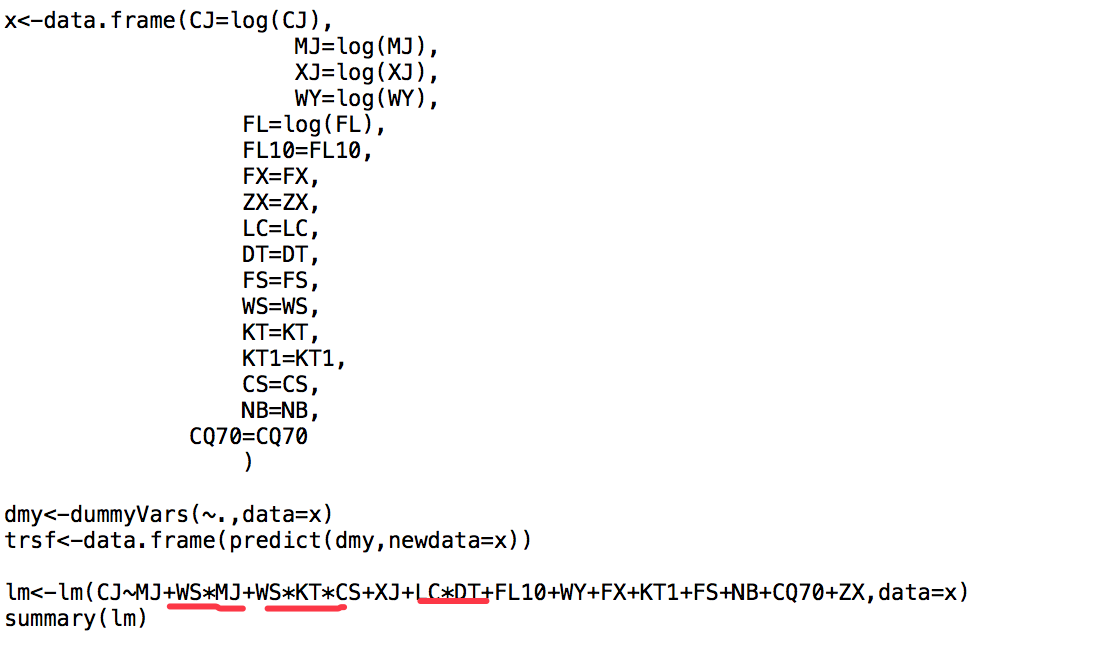
本次模型中选择对房龄（FL），客厅数量（KT）与产权年限（CQ）进行二进制转变，即房龄在十年以上为1，十年以下为0；客厅数小于1为1，大于等于1为0；产权年限70年以上和70年以下分别为1，0。



这里好像是因为之前的特征量检验时候 房龄在10年以上才三颗星，之后的模型是房龄用数值计算时候，没有星（反正就是相关度很低），修改后相关度提高，而且用测试集测试的时候准确率也大幅提高。

1. 在模型中加入相互作用的影响

由于某些特征对被解释变量即成交价有综合影响，或者说两个变量之间可能是有相互作用的，所以需要加入相互作用。



本次模型如图加入了卧室（WS）与面积（MJ）之间的相互影响，卧室（WS）客厅（KT）与卫生间（CS）之间的相互影响，楼层和电梯。

这个是当时思考怎么能再提高模型准确率，然后根据书里和咱们之前学过的这么弄的。（你可以说点根据咱们专业的经验判断出来卧室和面积之间也是有关系 还有楼层和电梯？）

**五．需要改进处与后续建议**

1.数据中各个属性的数量要均衡且充足（均衡性）

2.因为小区均价的不确定性（同一小区房子的差价可大可小）所以可以增加小区的区位条件作为特征变量加入到模型中来进行计算 排除其他不确定性的干扰。（小区均价太笼统了，当时是时间来不及了 只能用小区均价作为区位条件的代表。其实可以引入地图的api等等来具体到下面说的这些内容）

其中小区区位以北京地区为例：环数，距离各商圈中心的长度，距离地铁站的长度，周边配套设施的完善度（幼儿园 小学 超市 集市 车库等）周边环境的特殊性（如附近是否有污染源 是否有风景区）

3.朝向的界定

此次仅仅采用了是否为南北的属性规则。

4.各变量之间的耦合关系

本次模型只考虑了比较简单的关系，但变量之间应该还存在着较为复杂的连带关系 如厅 卧室 厨房的数量配比及其与建筑面积的关系。

5.预测时还要考虑相关的政策等软指标～如326限购之后大部分房价有所降低。

6.现阶段人为的修正还是挺重要的，即使是同一小区中之差距一层的楼房因视野高度等也可能会有部分差价，具体可以参考下考取房产估价师的相关资料。

Over 还有很多不足需要指正！感谢感谢哈XD～～