BostonHousing 数据集回归分析报告

20300180139 李哲彦 20307100180 卢威兆 19307110144 何昀翰
20307100049 包凯宇 20300180133 刘忻奕

1 案例背景

1.1 问题描述

在本选题中，我组代入波士顿房地产开发商的角度，需根据经济形势合理选址建造商品房，并需对房屋进行合理定价，以实现效益最大化。因此，通过利用统计手段分析 1970 年人口普查中得到的 506 个波士顿房价数据，我们希望能初步得出房价与各影响因素之间的关系。

1.2 所用数据集

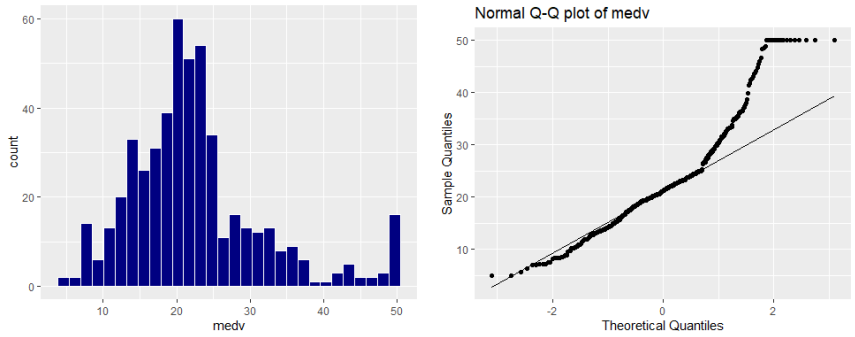
原始数据集来源于 UCI 机器学习数据库 [UCI Repository Of Machine Learning Databases. http://www.ics.uci.edu/ mlearn/MLRepository.html]，包含 14 个变量，共 506 个样本。其中变量具体含义如表 1所示。

表 1: 变量介绍

变量名	含义	取值范围/水平
crim	城镇人均犯罪率	$0.006 \sim 88.976$
zn	超过 25000 平方英尺的地块中住宅用地占比	$0 \sim 100$
indus	城镇非零售经营面积占比	$0.5 \sim 27.8$
chas	Charles River 虚拟变量（邻接河流取 1，否则取 0）	2 个水平：$0, 1$
nox	氮氧化物浓度（千万分之一）	$0.39 \sim 0.87$
rm	住宅平均房间数	$3.6 \sim 8.8$
age	1940 年以前建造的自住单位比例	$2.9 \sim 100.0$
dis	到波士顿五个就业中心的加权平均距离	$1.1 \sim 12.1$
rad	辐射状公路可达性指数	9 个水平：$1, 2, \ldots, 8, 24$
tax	按每 10000 美元计算的全职物业税税率	$187 \sim 711$
ptratio	城镇的学生-教师比	$12.6 \sim 22.0$
b	$1000(Bk - 0.63)^2$，其中 Bk 为城镇中黑人比例	$0.3 \sim 396.9$
lstat	底层阶级人口占比	$1.7 \sim 38.0$
medv	业主自住房屋的价值中位数（千美元）	$5 \sim 50$

2 描述性统计分析

2.1 因变量探索



(a) 直方图　　(b) 正态 Q-Q 图

图 1: 房价中位数 medv 分布特征

首先，我们作出因变量 medv 的直方图及正态 Q-Q 图如图 1所示。不难观察得到：medv 大体呈右偏分布，非正态，峰值约在 20000 与 24000 美元间，均值为 22530 美元；此外在最大值 50000 美元处有一个小的峰值，有 16 个数据（3%）的 medv 正好为 50000 美元，因此我们猜测 medv 的观测具有截尾性。

2.2 自变量探索

在影响 medv 的 13 个变量中，chas 是定性变量，只用作分类不具有数值意义，作出其余 12 个变量的直方图以及各自关于 medv 的散点图，具体参见附录图 12至图 24，其中 rad 变量性质较为特殊，补充了箱线图与 medv 在其各水平上的分布图。简要分析分析其分布特点如下：

1. 人均犯罪率（crim）：严重右偏，65.6% 的数据小于 1%，中位数为 0.26%，最大值达 89.0%；与 medv 呈一定负相关关系。

2. 超过 25000 平方英尺的地块中住宅用地占比（zn）：取值较为离散，绝大部分（73.5%）数据为 0%，其余几个峰值在 $12.5\%(2\%), 20\%(4\%), 22\%(2\%), 25\%(2\%)$ 和 $80\%(3\%)$；与 medv 呈一定正相关关系。

3. 非零售经营面积占比（indus）：大体呈取值较离散的双峰分布，第一个峰较为平缓，数据集中在 2% 到 11%；另外的峰存在主要由于有较多数据取值为 18.1%（26.1% 的数据）和 19.58%（6.0% 的数据）；与 medv 呈一定负相关关系。

4. 氮氧化物浓度（nox）：呈右偏分布，均值为 0.55，与 medv 呈一定负相关关系。

5. 住宅平均房间数（rm）：大体呈有尖峰的对称分布，均值为 6.29 个，与 medv 呈较明显正相关关系。

6. 1940 年以前建造的自住单位比例（age）：呈左偏分布，均值为 68.6%，有 43 个数据（8.5%）的 age 达到最大值 100%；与 medv 呈一定负相关关系。

7. 到就业中心的加权平均距离（dis）：呈右偏分布，均值为 3.80；与 medv 呈一定正相关关系。

8. 可达性指数（rad）：两个峰值在 4 和 5（共 44.5% 的数据）以及 24（26.1% 的数据）；medv 在 rad 取值 1 到 8 时似乎无显着区别，而 rad = 24 时，medv 显著下降。

9. 物业税税率（tax）：大体呈双峰分布，第一个峰约在 307，数据集中在 187 到 469；另外的峰存在由于有较多数据（26.1%）取值为 666；与 medv 呈一定负相关关系。

10. 学生-教师比（ptratio）：呈较离散的左偏分布，均值为 18.46，有峰值 20.2；与 medv 呈一定负相关关系。

11. 黑人比例系数（b）：严重左偏，有 121 个数据（23.9%）取最大值 396.9；与 medv 呈一定正相关关系。

12. 底层阶级人口占比（lstat）：呈右偏分布，均值为 12.65%；与 medv 呈较明显负相关关系。

综上，可以初步判断 medv 与 zn，rm，dis 和 b 呈一定正相关关系，与 crim，indus，nox，age，tax，ptratio 和 lstat 呈一定负相关关系；在 rad = 24 水平下，medv 的分布显着异于其他水平，其值较低，即 medv 与 rad 似存在负相关关系。

除前述的 chas 与 rad 外，变量 zn 的分量有大量零值，考虑到其定义为占地面积超过 25000 平方英尺的住宅用地比例，这是合理现象，但可能会影响后续数据处理的方式。

进一步考虑两个离散变量 chas 与 rad，我们作出相应的箱线图图 2（rad 参照图 24a）并进行方差分析，可知 medv 与二者均具有相关性，其中 chas 尤为显著。

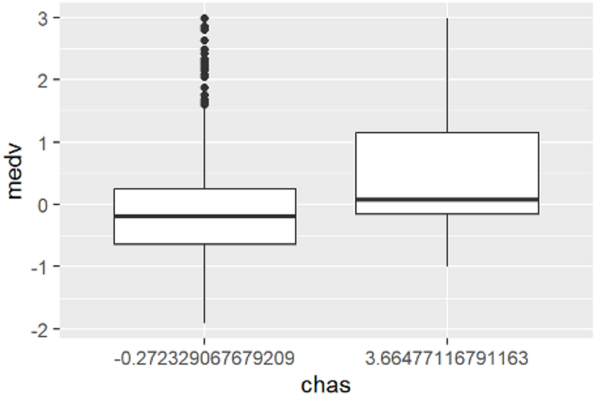
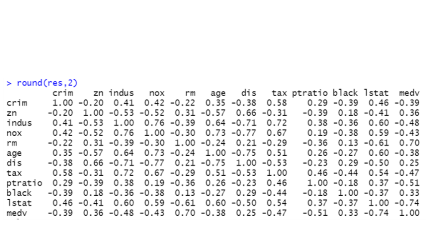


图 2: chas

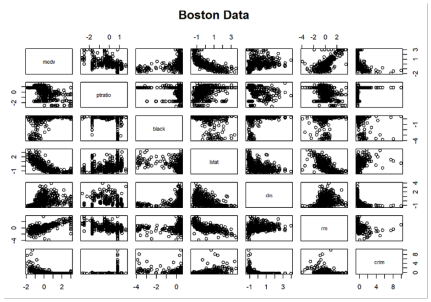
```
> summary(aov1)
              Df Sum Sq Mean Sq F value   Pr(>F)
chas           1   15.5  15.512   15.97 7.39e-05 ***
Residuals    504  489.5   0.971
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov2)
              Df Sum Sq Mean Sq F value Pr(>F)
rad            8  115.5  14.434   18.42 <2e-16 ***
Residuals    497  389.5   0.784
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

求出其余自变量和因变量的相关系数矩阵并作出散点图矩阵如图 3。其中相关系数矩阵表明一些自变量间存在较强的相关关系；而由散点图矩阵可以看出，一些自变量和因变量似乎不是简单的线性关系，可以尝试用取对数或二次方等方法转化为线性关系。例如，medv 和 lstat 以及 rm 显示出二次趋势，作出具体的散点图如图 4，并比较拟合曲线。



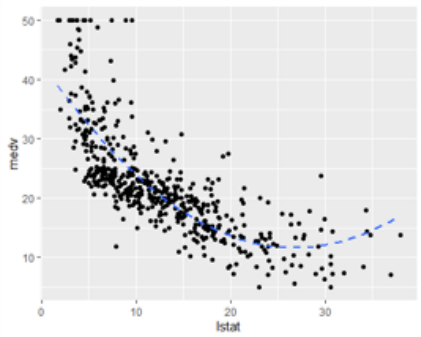
```
> round(res,2)
        crim    zn indus   nox    rm   age   dis   tax ptratio black lstat  medv
crim    1.00 -0.20  0.41  0.42 -0.22  0.35 -0.38  0.58    0.29 -0.39  0.46 -0.39
zn     -0.20  1.00 -0.53 -0.52  0.31 -0.57  0.66 -0.31   -0.39  0.18 -0.41  0.36
indus   0.41 -0.53  1.00  0.76 -0.39  0.64 -0.71  0.72    0.38 -0.36  0.60 -0.48
nox     0.42 -0.52  0.76  1.00 -0.30  0.73 -0.77  0.67    0.19 -0.38  0.59 -0.43
rm     -0.22  0.31 -0.39 -0.30  1.00 -0.24  0.21 -0.29   -0.36  0.13 -0.61  0.70
age     0.35 -0.57  0.64  0.73 -0.24  1.00 -0.75  0.51    0.26 -0.27  0.60 -0.38
dis    -0.38  0.66 -0.71 -0.77  0.21 -0.75  1.00 -0.53   -0.23  0.29 -0.50  0.25
tax     0.58 -0.31  0.72  0.67 -0.29  0.51 -0.53  1.00    0.46 -0.44  0.54 -0.47
ptratio 0.29 -0.39  0.38  0.19 -0.36  0.26 -0.23  0.46    1.00 -0.18  0.37 -0.51
black  -0.39  0.18 -0.36 -0.38  0.13 -0.27  0.29 -0.44   -0.18  1.00 -0.37  0.33
lstat   0.46 -0.41  0.60  0.59 -0.61  0.60 -0.50  0.54    0.37 -0.37  1.00 -0.74
medv   -0.39  0.36 -0.48 -0.43  0.70 -0.38  0.25 -0.47   -0.51  0.33 -0.74  1.00
```

(a) 相关系数矩阵

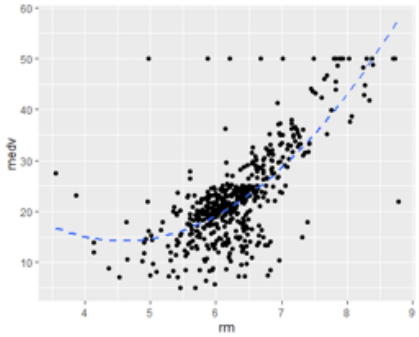


(b) 散点图矩阵

图 3



(a) lstat



(b) rm

图 4: 二次拟合图象

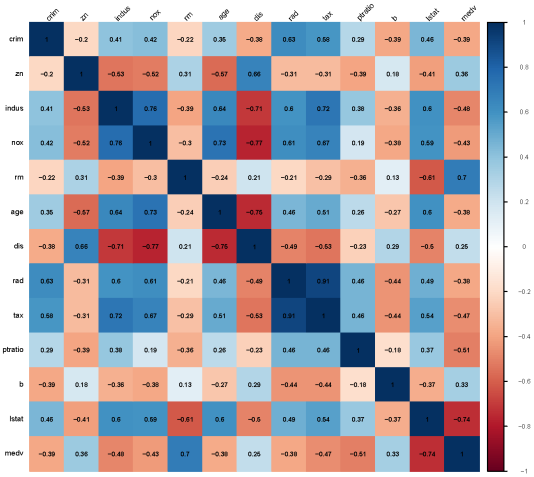


图 5: 热力图矩阵

最后，由相关系数图图 5，发现 medv 与 rm 有较强正相关性，与 lstat，ptratio，indus 和 tax 有较强负相关性，与此前分析相符。此外，tax 和 rad 有较高相关性，其潜在的多重共线性问题有待后续检验。

3 回归分析

3.1 模型建立

首先考虑全模型回归，在代码中以 model_0 表示，结果如图 6a，可知 indus 和 age 不显著，亦即每个城镇非零售营业面积的比例和 1940 年以前建造的业主自用单位的比例对房价影响不大。此时，以 0.7 : 0.3 的比例划分训练集和测试集，得出预测的误差平方和为 4511.62。

```
Call:
lm(formula = Boston$medv ~ ., data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69559 -0.29680 -0.05633  0.19322  2.84864

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.833e-15  2.294e-02   0.000 1.000000
crim        -1.010e-01  3.074e-02  -3.287 0.001087 **
zn           1.177e-01  3.481e-02   3.382 0.000778 ***
indus        1.534e-02  4.587e-02   0.334 0.738288
chas         7.420e-02  2.379e-02   3.118 0.001925 **
nox         -2.238e-01  4.813e-02  -4.651 4.25e-06 ***
rm           2.911e-01  3.193e-02   9.116  < 2e-16 ***
age          2.119e-03  4.043e-02   0.052 0.958229
dis         -3.378e-01  4.567e-02  -7.398 6.01e-13 ***
rad          2.897e-01  6.281e-02   4.613 5.07e-06 ***
tax         -2.260e-01  6.891e-02  -3.280 0.001112 **
ptratio     -2.243e-01  3.080e-02  -7.283 1.31e-12 ***
black        9.243e-02  2.666e-02   3.467 0.000573 ***
lstat       -4.074e-01  3.938e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.516 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

(a) 全模型回归

```
Call:
lm(formula = Boston$medv ~ ., data = new)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69601 -0.29777 -0.05487  0.18780  2.85278

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.842e-15  2.289e-02   0.000 1.000000
crim        -1.014e-01  3.066e-02  -3.307 0.001010 **
zn           1.163e-01  3.429e-02   3.390 0.000754 ***
chas         7.508e-02  2.359e-02   3.183 0.001551 **
nox         -2.189e-01  4.454e-02  -4.915 1.21e-06 ***
rm           2.904e-01  3.104e-02   9.356  < 2e-16 ***
dis         -3.418e-01  4.252e-02  -8.037 6.84e-15 ***
rad          2.837e-01  6.003e-02   4.726 3.00e-06 ***
tax         -2.158e-01  6.180e-02  -3.493 0.000521 ***
ptratio     -2.228e-01  3.038e-02  -7.334 9.24e-13 ***
black        9.223e-02  2.654e-02   3.475 0.000557 ***
lstat       -4.057e-01  3.682e-02 -11.019  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.515 on 494 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7348
F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

(b) 向后逐步回归

图 6: 全模型与选模型

于是自然地，我们考虑删去 indus 与 age 两个变量后的选模型 model_1，进行向后逐步回归观察选模型的结果（图 6b）。我们发现自变量的删除对 R^2 几乎不产生影响，且调整的 R^2 统计量略有改善，说明变量选择具有合理性。此时各自变量统计意义上都显著，同时测试集预测的误差平方和缩小为 4481.712，但模型 R^2 表现不理想，可以进一步优化。

结合 2.2 中的结论，结合观察与反复测试，我们最终选择加入两个关于 lstat 与 rm 的二次项再做回归，得到模型 model_2，这也是改进最明显的一步：R^2 统计量从 0.74 上升到 0.82 左右，新加入的二次项在统计学意义上显著。此模型中测试集预测的误差平方和降低至 3240.365，而 zn 变量的显著性降低了，p 值略高于 0.05，可以进行进一步的变量筛选。

```
> new<-cbind(new,I(Boston$lstat^2),I(Boston$rm^2))
> model_2<-lm(Boston$medv~.,data=new)
> summary(model_2)

Call:
lm(formula = Boston$medv ~ ., data = new)

Residuals:
     Min       1Q   Median       3Q      Max
-2.88106 -0.24984 -0.01954  0.19595  2.87486

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.21843    0.02560  -8.532  < 2e-16 ***
crim            -0.14139    0.02601  -5.435 8.64e-08 ***
zn               0.05252    0.02938   1.788 0.074468 .
chas             0.06700    0.01985   3.374 0.000798 ***
nox             -0.18319    0.03775  -4.853 1.63e-06 ***
rm               0.22404    0.02652   8.447 3.39e-16 ***
dis             -0.26672    0.03621  -7.366 7.46e-13 ***
rad              0.22409    0.05068   4.421 1.21e-05 ***
tax             -0.15884    0.05215  -3.046 0.002443 **
ptratio         -0.16255    0.02590  -6.276 7.63e-10 ***
black            0.06948    0.02239   3.104 0.002022 **
lstat           -0.56764    0.03878 -14.638  < 2e-16 ***
`Boston$lstat^2`  0.11348    0.01816   6.250 8.93e-10 ***
`Boston$rm^2`     0.10538    0.01166   9.040  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4332 on 492 degrees of freedom
Multiple R-squared:  0.8172,    Adjusted R-squared:  0.8123
F-statistic: 169.2 on 13 and 492 DF,  p-value: < 2.2e-16
```

(a) 加入二次项

```
> model_3<-update(model_2,~.-zn)
> #变量选择
> summary(model_3)

Call:
lm(formula = Boston$medv ~ crim + chas + nox + rm + dis + rad +
    tax + ptratio + black + lstat + `Boston$lstat^2` + `Boston$rm^2`,
    data = new)

Residuals:
     Min       1Q   Median       3Q      Max
-2.90297 -0.24706 -0.03459  0.18386  2.88735

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.22404    0.02546  -8.799  < 2e-16 ***
crim            -0.13839    0.02602  -5.319 1.59e-07 ***
chas             0.06680    0.01990   3.357 0.000848 ***
nox             -0.18862    0.03771  -5.002 7.89e-07 ***
rm               0.22939    0.02641   8.685  < 2e-16 ***
dis             -0.23435    0.03143  -7.457 4.01e-13 ***
rad              0.21697    0.05064   4.285 2.20e-05 ***
tax             -0.13978    0.05116  -2.732 0.006513 **
ptratio         -0.17678    0.02470  -7.157 3.01e-12 ***
black            0.06942    0.02244   3.094 0.002087 **
lstat           -0.57467    0.03866 -14.864  < 2e-16 ***
`Boston$lstat^2`  0.11899    0.01793   6.635 8.58e-11 ***
`Boston$rm^2`     0.10549    0.01168   9.030  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4342 on 493 degrees of freedom
Multiple R-squared:  0.816,     Adjusted R-squared:  0.8115
F-statistic: 182.2 on 12 and 493 DF,  p-value: < 2.2e-16
```

(b) 初步模型建立完成

图 7: model_2 与 model_3

于是在删除 zn 变量后我们得到了模型 model_3。注意到删除 zn 带来的 R^2 的变化很小，可以认为 zn 变量贡献不大，而测试集上的预测结果甚至略有改善，误差平方和降至 3227.076。至此，我们得到了一个效果基本良好的模型，开始进行回归诊断。

3.2 回归诊断

3.2.1 强影响点与异常值点处理

先进行残差检验，具体表现见附录图 25，主体部分表现良好，但模型有几个明显的异常点，严重影响了模型的结果和残差的分布。

```
1 > w
2  365  369  370  371  372  373  506
```

通过简单的 cook 距离计算与 F 检验，以 p 值小于 0.01 作为标准筛选异常值点，查找到的异常值点如上所示，共 7 个，由于其数量有限，故采取直接删除的处理方法。

3.2.2 多重共线性

```
> vif(model_3)
      crim                chas                 nox                  rm ...
                         dis                 rad
  1.813534            1.060735            3.809152            1.868822 ...
                 2.645765            6.869687
       tax             ptratio               black               lstat ...
          `Boston$lstat^2`     `Boston$rm^2`
  7.011276            1.634469            1.348690            4.004796 ...
                 2.129767            1.408896
> kappa(XX, exact = T)
[1] 61.25595
```

vif 检验及 kappa 值均表明了模型多重共线性不明显，故不需要采取岭回归或主成分回归等方法。

3.2.3 自相关性

```
> dwtest(model_3, alternative = "two.sided")

        Durbin-Watson test

data:  model_3
DW = 1.1203, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is not 0
```

dw 检验显示，误差存在明显自相关性。然而该模型并不具有时间序列，因此自相关性并不寻常，需要进一步检查与讨论。

3.3 模型调整

根据 3.2.1 中的分析，先删去强影响点后重新回归，得到 model_4，其统计量及相关分析具体参见附录图 26。

对新模型诊断发现，强影响点的影响削弱了，同时 R^2 得到明显提升，说明少量的强影响点确实阻碍了准确的回归，但强自相关性依然没有消除。

在尝试使用 Box-Cox 变换消除自相关性并发现效果不佳后，我们采用了迭代法来消除自相关性，得到 model_5，如图 8a所示。不难发现，迭代法很有效地消除了自相关性，但是影响了 R^2 的表现。

```
> dwtest(model_5, alternative = "two.sided")

        Durbin-Watson test

data:  model_5
DW = 1.9187, p-value = 0.2699
alternative hypothesis: true autocorrelation is not 0
```

```
> summary(model_5)

Call:
lm(formula = y ~ ., data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97271 -0.18458 -0.02302  0.16438  1.33611

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.148764   0.015904  -9.354  < 2e-16 ***
crim            -0.096421   0.018763  -5.139 4.01e-07 ***
chas             0.009072   0.016153   0.562 0.574609
nox             -0.149292   0.035143  -4.248 2.58e-05 ***
rm               0.339170   0.020440  16.593  < 2e-16 ***
dis             -0.156708   0.030479  -5.142 3.96e-07 ***
rad              0.155245   0.046213   3.359 0.000843 ***
tax             -0.198621   0.046420  -4.279 2.27e-05 ***
ptratio         -0.148875   0.022917  -6.496 2.04e-10 ***
black            0.070442   0.019068   3.694 0.000246 ***
lstat           -0.397341   0.031872 -12.467  < 2e-16 ***
`Boston$lstat^2`  0.078842   0.013382   5.892 7.16e-09 ***
`Boston$rm^2`     0.125963   0.008595  14.655  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3029 on 485 degrees of freedom
Multiple R-squared:  0.8442,    Adjusted R-squared:  0.8403
F-statistic:   219 on 12 and 485 DF,  p-value: < 2.2e-16
```

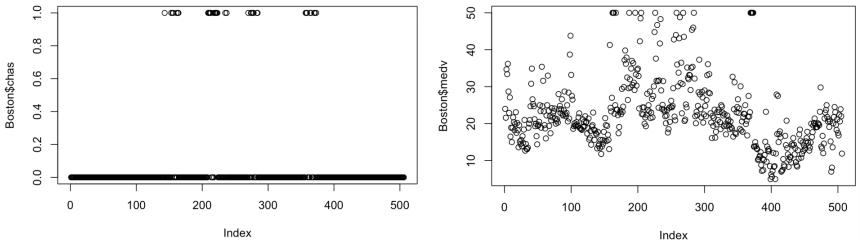
(a) model_5

```
> summary(model_6)

Call:
lm(formula = y ~ crim + nox + rm + dis + rad + tax + ptratio +
    black + lstat + `Boston$lstat^2` + `Boston$rm^2`, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97593 -0.17913 -0.02205  0.16291  1.33526

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.148655   0.015892  -9.354  < 2e-16 ***
crim            -0.096603   0.018747  -5.153 3.73e-07 ***
nox             -0.148561   0.035094  -4.233 2.75e-05 ***
rm               0.339724   0.020402  16.652  < 2e-16 ***
dis             -0.157575   0.030418  -5.180 3.25e-07 ***
rad              0.156375   0.046137   3.389 0.000757 ***
tax             -0.200687   0.046242  -4.340 1.73e-05 ***
ptratio         -0.150163   0.022786  -6.590 1.15e-10 ***
black            0.070778   0.019045   3.716 0.000226 ***
lstat           -0.396219   0.031787 -12.465  < 2e-16 ***
`Boston$lstat^2`  0.078405   0.013350   5.873 7.94e-09 ***
`Boston$rm^2`     0.125970   0.008589  14.666  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3027 on 486 degrees of freedom
Multiple R-squared:  0.8441,    Adjusted R-squared:  0.8405
F-statistic: 239.2 on 11 and 486 DF,  p-value: < 2.2e-16
```

(b) model_6

图 8: 应对自相关性与改进变量选择

注意到此时 chas 不再显著，猜测原因可能在于：街区排列是相邻的，邻河的房子也相邻，迭代时减轻了是否临河的影响。于是我们进一步删去变量 chas 得到模型 model_6（图 8b）。



(a) chas 值按序排列　　(b) medv 按序排列

图 9: chas 与 medv 的相关性

按序号顺序分别画出 chas（是否临河）和 medv（房价）的图象，可以看到临河在直观上十分连续（图 9a），并且影响房价（图 9b）。这验证了先前猜想的合理性，也说

明相邻序号的房子具有地理上的相关性，指示房地产商寻找相邻房价信息，来辅助新房的定价预测。

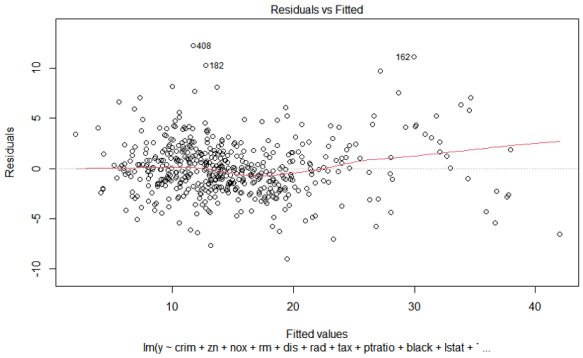


图 10: model_6 的残差分布

最后进行异方差性的分析，异方差性检验揭示了模型具有异方差性。根据最终模型的残差分布（图 10），异方差性在房价较集中的部分是轻微的。

对模型使用 Box-Cox 变换，计算不同 lambda 值对应 Box-Cox 变换的似然函数，结合附录图 28求最值可知 $\lambda = 0.11$，于是采取 Box-Cox 变换得到 model_7（图 11a），再次进行残差分析与异方差性检验发现异方差性并没有消除。

随后我们考虑使用 WLSE 来处理异方差性，首先求出等级相关系数最大的自变量，为 rm。随后以 0.5 为步长在 $[-2, 2]$ 中选出使对数似然函数最大化的权重函数幂指数，求得最优指数为 0；缩小步长至 0.01 在 $[-0.5, 0.5]$ 中查找，求得 0.15。以 0.15 为权重函数指数做 WLSE，得到模型 model_8，如图 11b所示。

```
> summary(lm1_bc) # 输出拟合结果

Call:
lm(formula = y_bc ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12, data = data0)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98594 -0.14019 -0.02438  0.14043  1.19928

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.0619327  0.2863421  17.678  < 2e-16 ***
x1          -0.0135637  0.0018423  -7.362 7.64e-13 ***
x2           0.0018130  0.0007704   2.353    0.019 *
x3           0.0041824  0.0034333   1.218    0.224
x4          -1.0703700  0.2142746  -4.995 8.17e-07 ***
x5           0.1415834  0.0234460   6.039 3.06e-09 ***
x6           0.0003413  0.0007405   0.461    0.645
x7          -0.0719225  0.0111936  -6.425 3.11e-10 ***
x8           0.0209742  0.0037023   5.665 2.50e-08 ***
x9          -0.0009371  0.0002095  -4.473 9.59e-06 ***
x10         -0.0562493  0.0073108  -7.694 7.83e-14 ***
x11          0.0006053  0.0001506   4.021 6.72e-05 ***
x12         -0.0400722  0.0028415 -14.102  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2663 on 493 degrees of freedom
Multiple R-squared:  0.785,     Adjusted R-squared:  0.7798
F-statistic:   150 on 12 and 493 DF,  p-value: < 2.2e-16
```

(a) model_7

```
Call:
lm(formula = y ~ ., data = data0, weights = w)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-16.0695  -3.1777  -0.7362   2.2289  30.2391

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.039113   5.159238   6.985 9.25e-12 ***
x1           -0.112624   0.033165  -3.396 0.000739 ***
x2            0.046149   0.013789   3.347 0.000880 ***
x3            0.043339   0.061710   0.702 0.482823
x4          -17.254288   3.862890  -4.467 9.86e-06 ***
x5            3.971351   0.422495   9.400  < 2e-16 ***
x6            0.002304   0.013323   0.173 0.862799
x7           -1.474702   0.200935  -7.339 8.92e-13 ***
x8            0.325292   0.066695   4.877 1.45e-06 ***
x9           -0.013769   0.003770  -3.652 0.000288 ***
x10          -0.992209   0.131300  -7.557 2.03e-13 ***
x11           0.009817   0.002713   3.618 0.000327 ***
x12          -0.532319   0.051455 -10.345  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.492 on 493 degrees of freedom
Multiple R-squared:  0.739,     Adjusted R-squared:  0.7326
F-statistic: 116.3 on 12 and 493 DF,  p-value: < 2.2e-16
```

(b) model_8

图 11: 利用 Box-Cox 变换与加权最小二乘应对异方差性

考察 model_8 的表现，我们发现虽然模型的异方差性得到了很好的处理，但是 R^2 及预测效果受到了相应的影响，结合作为房地产商开发普通住宅的需求，最终仍然选用在这一方面表现良好的 model_6 作为总结结论时使用的模型。

4 结果解读

根据最终的 model_6，我们拟合出如下线性模型：

$$\text{medv} = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon,$$

$$\boldsymbol{x} = (\text{crim}, \text{nox}, \text{rm}, \text{dis}, \text{rad}, \text{tax}, \text{ptratio}, \text{b}, \text{lstat}, \text{lstat}^2, \text{rm}^2)^T, \boldsymbol{\beta} = (-0.096603, -0.148561,$$

$$0.339724, -0.157575, 0.156375, -0.200687, -0.150163, 0.070778, -0.396219, 0.078405, 0.125970)^T.$$

通过这一模型，我们可以推断，以下因素会分别对波士顿房地产开发商的选择造成各种不同程度的影响：

- 人口阶层：lstat 表示底层阶级人口占比，与 medv 呈二次函数关系，在 $\text{lstat} > 0$ 处有一个极小值，说明房价与社区的整体社会阶层有关。一般而言底层人口越多，房价越低；但如果不同阶层人口混住，这样的社区可能对上下阶层的居民吸引力都更差。

- 环境：nox 表示氮氧化物浓度，氮氧化物浓度越高，说明空气质量差，居住环境不宜人，相应地房价较低。

- 交通：rad 表示可达性指数，rad 值越高说明交通越便利，房价相应就高。居住成本：tax 表示物业税税率，物业税税率越高，居住成本越高，人们可能不乐意负担这部分生活成本而不愿购买此类房屋，导致需求降低，价格也降低；dis 表示到就业中心的平均加权距离，dis 的值越低，住户通勤时间就越短，由于存在一部分住户会选择用金钱换时间，dis 值低的住宅价格就更高。

- 教育：ptratio 表示学生-教师比，该比例越小，单位学生拥有的教育资源越丰富，有孩子的居民可能更愿意以高价购入此类房屋。

- 房屋本身：rm 表示单位住宅的平均房间数，rm 越大面积越大，总价也越高，而且增长是二次的，这也符合我们的一般认识：当房屋面积大到一般范围外后，由于其稀缺性和房屋整体品质的提升，价格往往不会简单随面积线性增长。

A 代码

主程序

```
library(MASS)
library(ggplot2)
library(Hmisc)
library(car)
library(lmtest)
rm(list=ls())
data(Boston)
summary(Boston)
head(Boston)
Boston=data.frame(scale(Boston))
#观察boston数据集的性质
#由观察以及资料显示，medv是因变量，代表房价中位数
#在自变量中，chas是分类变量，有且只有01两种取值，代表是否临河,需要特殊处理
#rad取值是等级，不连续，需要处理
#zn数据集零值较多，检查定义为占地面积超过25,000平方英尺的住宅用地比例，是合理的
#以下四行是划分训练集和测试集的代码，在迭代法上用起来不方便，暂时不用了
#set.seed(2022)
#index<-sort(sample(nrow(Boston),nrow(Boston)*0.7))
#train<-Boston[index,]
#test<-Boston[-index,]
tempchas<-Boston$chas
Boston$chas=as.factor(Boston$chas)
ggplot(Boston,aes(x=chas,y=medv))+ geom_boxplot()
aov1<-aov(medv~chas,Boston)
aov1
summary(aov1)
#由箱线图和方差分析，可以看出medv和chas有明显相关性
temprad=Boston$rad
Boston$rad=as.factor(Boston$rad)
ggplot(Boston,aes(x=rad,y=medv))+ geom_boxplot()
aov2<-aov(medv~rad,Boston)
aov2
summary(aov2)
#由方差分析，medv和rad有相关性，但箱线图值为24的情况表现比较奇怪
mydata<-(Boston[,c(1,2,3,5,6,7,8,10,11,12,13,14)])
#在刨去两个分类变量
res<-cor(mydata)
round(res,2)
#相关系数矩阵，观察相关性
pairs(~ medv + ptratio + black + lstat + dis + rm + crim,data=Boston, main = ...
    "Boston Data")
#根据相关系数矩阵以及一些试验，最终选定了效果较好的散点图矩阵
#相关系数矩阵表明一些自变量间存在较强的相关关系
```

```
#由散点图矩阵可以看出，一些自变量和因变量并不是简单线性关系，可以尝试用取对数或二次方等方法转化为线性关系
#例如，medv和istat以及rm显示出二次趋势
ggplot(Boston,aes(x=lstat,y=medv))+geom_point(shape=5)+geom_smooth(method=lm)
ggplot(Boston,aes(x=rm,y=medv))+geom_point(shape=5)+geom_smooth(method=lm)
#两张具体的散点图，效果探索
Boston$rad<-temprad
Boston$chas<-tempchas
#重新把分类变量作为数值变量，作多元线性回归，因为这些分类变量定义是有逻辑顺序的
#尝试过对分类变量作回归，效果不佳。尝试将rad值24重新赋值0，回归失去显著性。（这些代码已经删除）
model_0<-lm(Boston$medv~.,data=Boston)
summary(model_0)
#全变量回归，有部分自变量不显著
model_0.back<-step(model_0,direction="backward")
#向后逐步回归
new<-Boston[,c(1,2,4,5,6,8,9,10,11,12,13)]
model_1<-lm(Boston$medv~.,data=new)
#根据向后逐步回归的结果，做变量选择
summary(model_1)
#自变量删除几乎不对r2产生影响，调整的r2统计量还略有改善，说明变量选择是有道理的
new<-cbind(new,I(Boston$lstat^2),I(Boston$rm^2))
model_2<-lm(Boston$medv~.,data=new)
summary(model_2)
#作回归
model_3<-update(model_2,~.-zn)
#变量选择
summary(model_3)
new<-new[,-2]
#观察模型性质，各自变量统计意义上都显著，但模型R2表现不是很好
plot(model_3)
#模型有几个明显的强影响点
#残差异方差性不明显
cooks.distance(model_3)
#计算cook距离，进一步判断强影响点
p=12
n=dim(Boston)[1]
r=rstandard(model_3)
F = (n - p - 1) * r ^ 2 / (n - p -r ^ 2)
p.value = 1 - pf(F, 1, n - p - 1)
w=which(p.value < 0.01)
 #用cook距离作F检验，以p值小于0.01w筛选出异常点
new<-new[-w,]
Boston<-Boston[-w,]
#对强影响点尝试直接删除。这是比较草率的，因为强影响点较多。可能存在未被纳入考量的因素，需要修改模型，再重新考虑。
vif(model_3)
#vif检验的结果不支持多重共线性。
```

```
#但由之前相关系数矩阵，相关性值得再深入研究。可以考虑引入交互项
delete=c(3,7,14)
XX=cor(Boston[,-delete])
kappa(XX, exact = T)
#多重共线性不明显，故不适用岭回归
dwtest(model_3, alternative = "two.sided")
#dw检验显示，误差存在明显自相关性
#这非常奇怪，模型并不具有时间序列，自相关性并不寻常
model_4<-lm(Boston$medv~.,data=new)
summary(model_4)
#删除强影响点后重新回归
plot(model_4)
vif(model_4)
dwtest(model_4, alternative = "two.sided")
#对新模型诊断发现，强影响点的影响削弱了。
#R2得到提升，这是建立在删除大量数据的基础上，效果仍然不是很好。
#残差自相关性问题还没有解决。
#以下用迭代法消除自相关性
p1=1-1.3372/2
n <- length(Boston[,1])
y<- Boston$medv[2:n]-p1*Boston$medv[1:n-1]
x<-new[2:n,]-p1*new[1:n-1,]
model_5<-lm(y~.,data=x)
summary(model_5)
dwtest(model_5, alternative = "two.sided")
#自相关性得到消除，chas不再显著
model_6<-update(model_5,~.-chas)
summary(model_6)
#删除chas，猜测是街区排列是相邻的，邻河的房子也相邻，再迭代的时候减轻了是否临河的
影响
data(Boston)
plot(Boston$chas)
plot(Boston$medv)
Boston<-Boston[-w,]
#发现，邻河确实是比较连续的
#以下是boxcox，好像不太行
b2=boxcox(Boston$medv~crim+chas+nox+rm,data=new)
l=which(b2$y==max(b2$y))
b2$x[l]
lm.boxcox<-lm(Boston$medv^0.303~.,data=new)
summary(lm.boxcox)
dwtest(lm.boxcox,alternative = "two.sided")
```

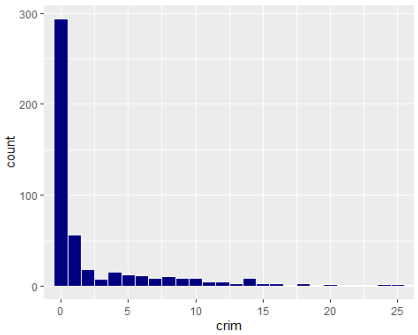
针对异方差性所做的处理

```
abse = abs(resid(lm1)) # 计算残差绝对值
for(i in 1:p)
{out=cor.test(data0[,i], abse, method = "spearman",exact=FALSE) # ...
```

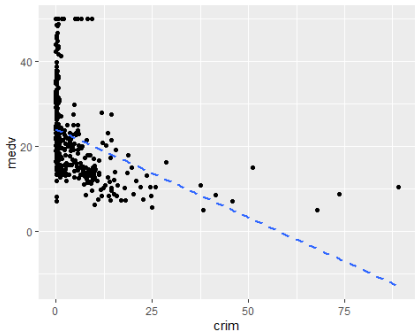
```
    计算残差与xi的相关系数
if(out$p.value<0.05)
  print(names(data0)[i])
}
# 结果表明残差绝对值与自变量存在显著相关关系,即数据存在异方差性

bc = boxcox(y ~ ., data = data0, lambda = seq(-2, 2, 0.01))
# 计算不同lambda值对应BoxCox变换的似然函数，lambda取值区间为[-2, 2]，步长为0.01
# 输出的图像展示对数似然函数随lambda增长的变化
lambda = bc $ x[which.max(bc $ y)] # 选取使似然函数达到最大值的lambda值
lambda # 输出lambda值
y_bc = (data0 $ y ^ lambda - 1) / lambda # 计算变换后的y值

lm1_bc = lm(y_bc ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12, data = data0) # ...
    使用变换后的y值拟合线性模型
summary(lm1_bc) # 输出拟合结果
abse_bc = abs(resid(lm1_bc)) # 计算残差绝对值
for(i in 1:p)
{out=cor.test(data0[,i], abse_bc, method = "spearman",exact=FALSE) # ...
    计算残差与xi的相关系数
if(out$p.value<0.05)
  print(names(data0)[i])
}
# 借助boxcox变换，数据异方差性并未消除。

s = seq(-2, 2, 0.5) # 产生数列(-2,-1.5,...,2)作为权重函数幂指数备选
result1 = vector(length = length(s), mode = "list") # ...
    产生一个与s维度相同的空向量以储存后续结果
result2 = vector(length = length(s), mode = "list") # ...
    产生一个与s维度相同的空向量以储存后续结果

corr2 = vector(length = p, mode = "list") # ...
    产生一个长度为p的向量，用来存储rho的绝对值
for(i in 1:p)
{corr2[i]=abs(cor.test(data0[,i], abse, method = ...
    "spearman",exact=FALSE)$estimate) # 把rho的绝对值存储在向量corr2中
}
xk=which.max(corr2) # corr2的最大的元素对应等级相关系数最大的自变量
xk # 找出等级相关系数最大的自变量x5
for(j in 1 : length(s))
{

  w = data0 $ x5 ^ (-s[j]) # 计算权重
  lm1_wlse = lm(y ~ ., weights = w, data0) # 用WLSE拟合线性模型
  result1[[j]] = logLik(lm1_wlse) # 储存拟合模型对应的对数似然函数值
  result2[[j]] = summary(lm1_wlse) # 储存模型拟合结果

}
```

```
result1 # 输出对数似然函数值
s[which.max(result1)]
result2[which.max(result1)] # 输出对应对数函数最大值的模型

# 发现m=0时loglikelihood取max,下面更改s，再次做wlse

s = seq(-0.5, 0.5, 0.01) # 产生数列(-0.49,-0.48,...,0.5)作为权重函数幂指数备选
result1 = vector(length = length(s), mode = "list") # ...
    产生一个与s维度相同的空向量以储存后续结果
result2 = vector(length = length(s), mode = "list") # ...
    产生一个与s维度相同的空向量以储存后续结果

for(j in 1 : length(s))
{

  w = data0 $ x5 ^ (-s[j]) # 计算权重
  lm1_wlse = lm(y ~ ., weights = w, data0) # 用WLSE拟合线性模型
  result1[[j]] = logLik(lm1_wlse) # 储存拟合模型对应的对数似然函数值
  result2[[j]] = summary(lm1_wlse) # 储存模型拟合结果

}
result1 # 输出对数似然函数值
s[which.max(result1)]
result2[which.max(result1)] # 输出对应对数函数最大值的模型
# 此时R square=0.739,F-statistic=116.3,普通最小二乘估计R ...
    square=0.7355,F-statistic=114.3
# 说明对于这个数据集加权最小二乘估计的拟合效果好于普通最小二乘估计
```

B 图表

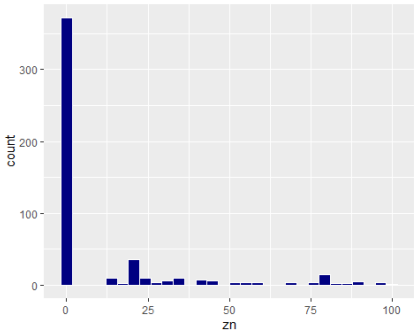


(a) 直方图

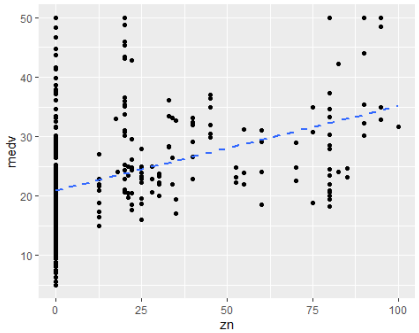


(b) 散点图及线性拟合结果

图 12: crim

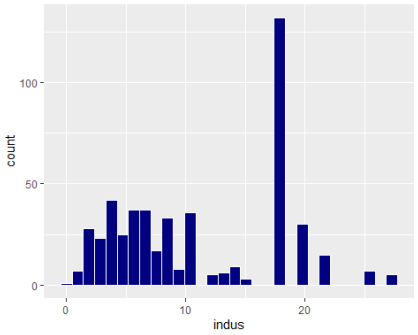


(a) 直方图

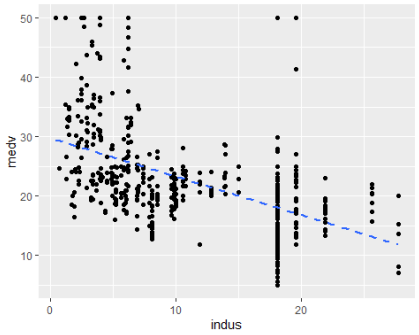


(b) 散点图及线性拟合结果

图 13: zn

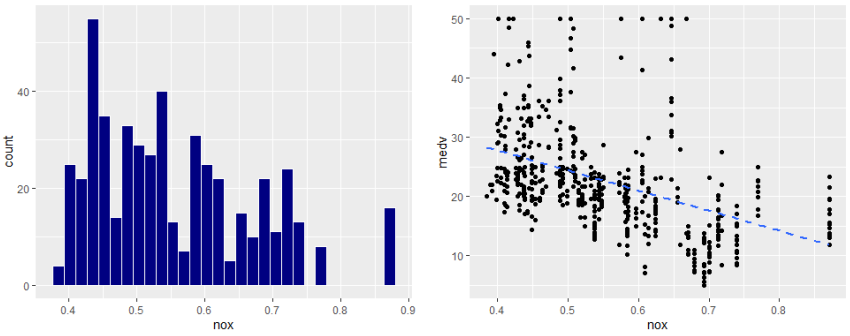


(a) 直方图



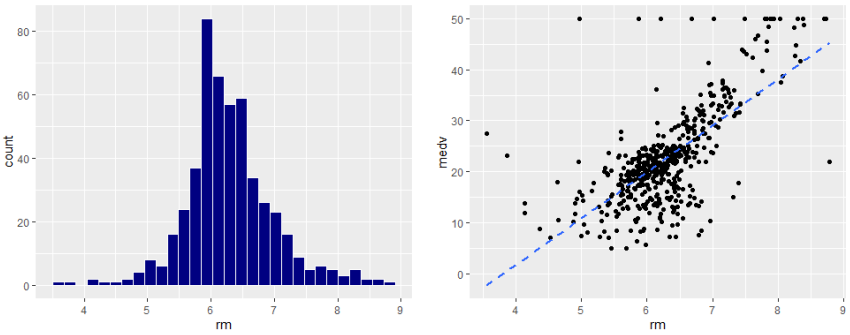
(b) 散点图及线性拟合结果

图 14: indus



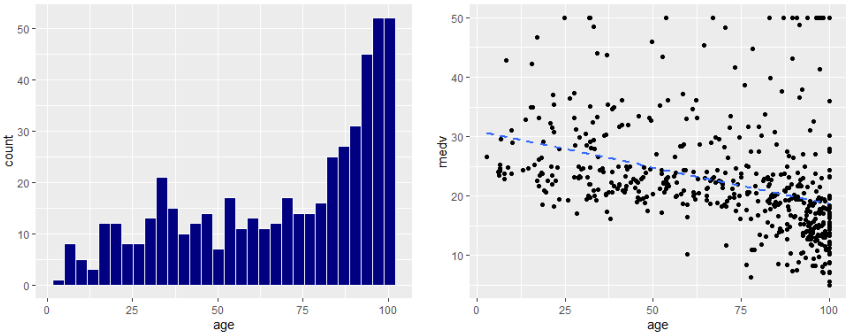
(a) 直方图　　(b) 散点图及线性拟合结果

图 15: nox



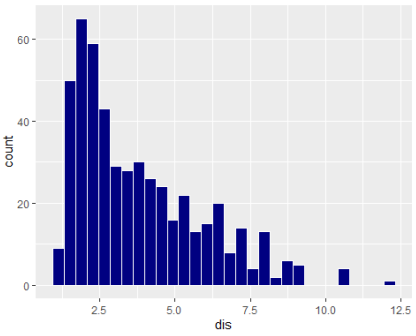
(a) 直方图　　(b) 散点图及线性拟合结果

图 16: rm

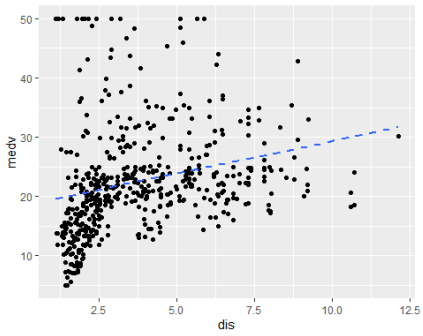


(a) 直方图　　(b) 散点图及线性拟合结果

图 17: age

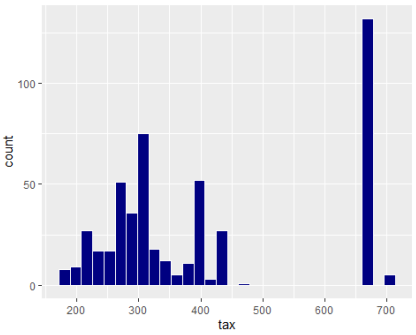


(a) 直方图

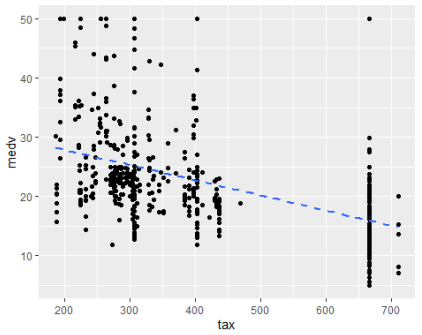


(b) 散点图及线性拟合结果

图 18: dis

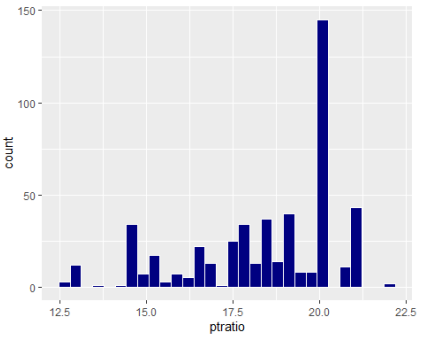


(a) 直方图

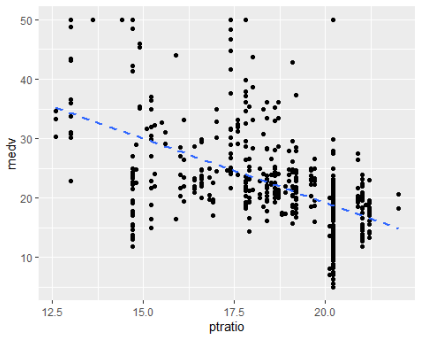


(b) 散点图及线性拟合结果

图 19: tax

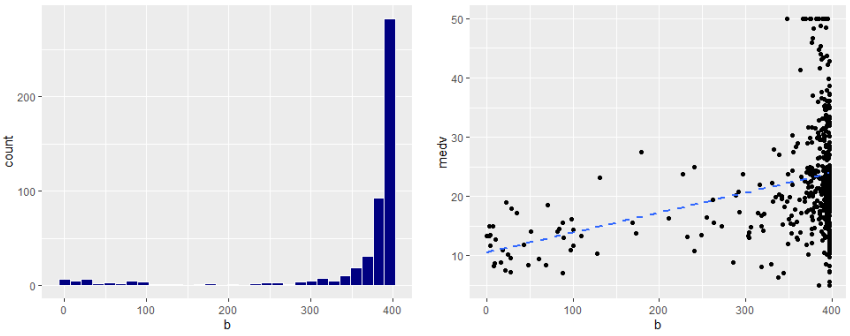


(a) 直方图



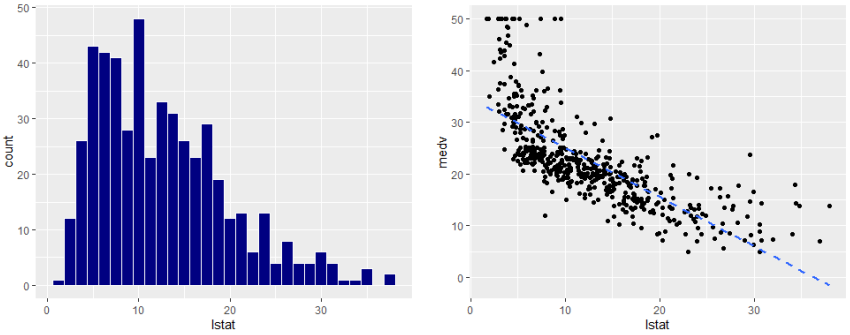
(b) 散点图及线性拟合结果

图 20: ptratio



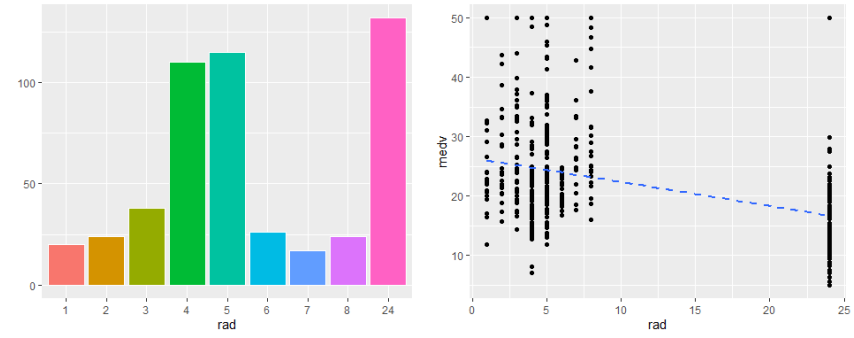
(a) 直方图　　(b) 散点图及线性拟合结果

图 21: b



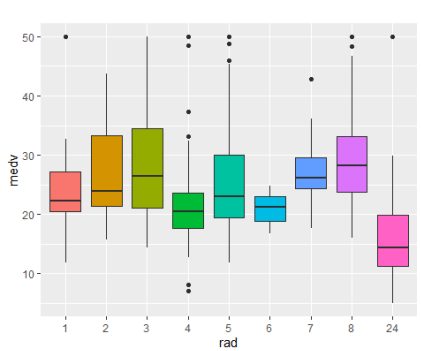
(a) 直方图　　(b) 散点图及线性拟合结果

图 22: lstat

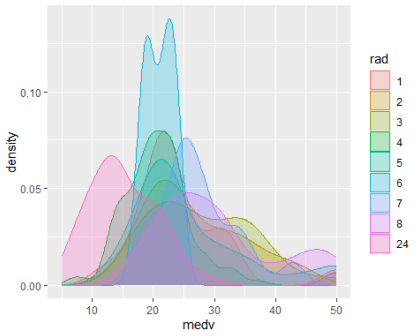


(a) 直方图　　(b) 散点图及线性拟合结果

图 23: rad12



(a) 箱线图



(b) medv 在 rad 每个水平上的分布

图 24: rad34

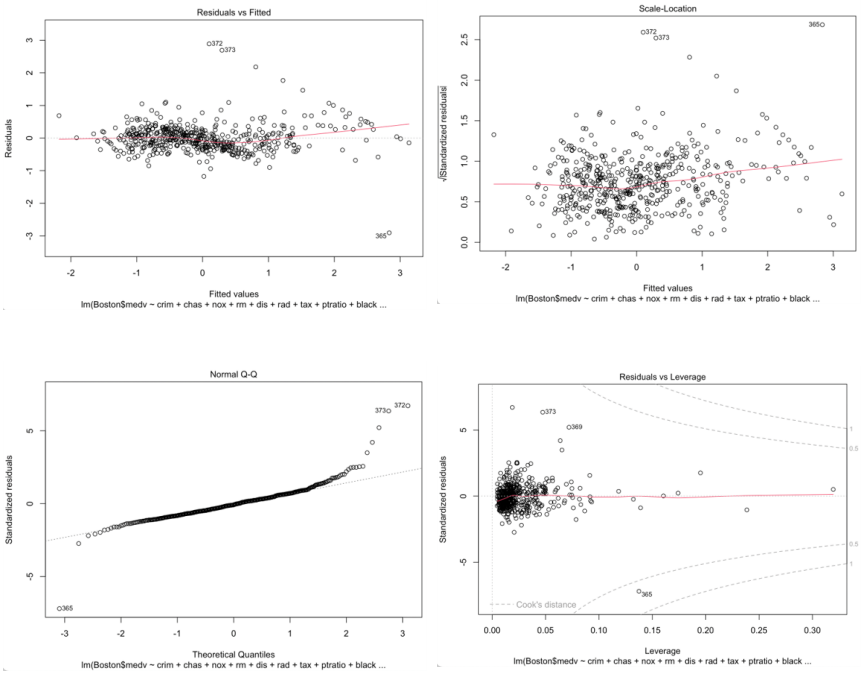


图 25: model_3 的残差分布图

```
> summary(model_4)

Call:
lm(formula = Boston$medv ~ ., data = new)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99013 -0.21059 -0.01805  0.19042  1.17680

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.228831   0.019312 -11.849  < 2e-16 ***
crim             -0.127472   0.019721  -6.464 2.49e-10 ***
chas              0.038497   0.015910   2.420 0.015900 *
nox              -0.153695   0.028606  -5.373 1.20e-07 ***
rm                0.309335   0.020654  14.977  < 2e-16 ***
dis              -0.164872   0.024093  -6.843 2.34e-11 ***
rad               0.148673   0.038471   3.865 0.000126 ***
tax              -0.158143   0.038818  -4.074 5.40e-05 ***
ptratio          -0.160752   0.018942  -8.487 2.59e-16 ***
black             0.060432   0.016980   3.559 0.000409 ***
lstat            -0.423991   0.031049 -13.655  < 2e-16 ***
`Boston$lstat^2` 0.072227   0.013937   5.182 3.22e-07 ***
`Boston$rm^2`    0.139467   0.009204  15.153  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3282 on 486 degrees of freedom
Multiple R-squared:  0.8859,    Adjusted R-squared:  0.883
F-statistic: 314.3 on 12 and 486 DF,  p-value: < 2.2e-16
```

(a) summary

```
> vif(model_4)
            crim             chas              nox               rm
        1.819118         1.058908         3.799623         1.928257
             dis              rad              tax          ptratio
        2.674262         6.695276         6.855985         1.664017
           black            lstat `Boston$lstat^2`    `Boston$rm^2`
        1.350373         4.454585         2.244943         1.416818
> dwtest(model_4, alternative = "two.sided")

        Durbin-Watson test

data:  model_4
DW = 1.2897, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is not 0
```

(b) 多重共线性与自相关性检验

图 26: model_4

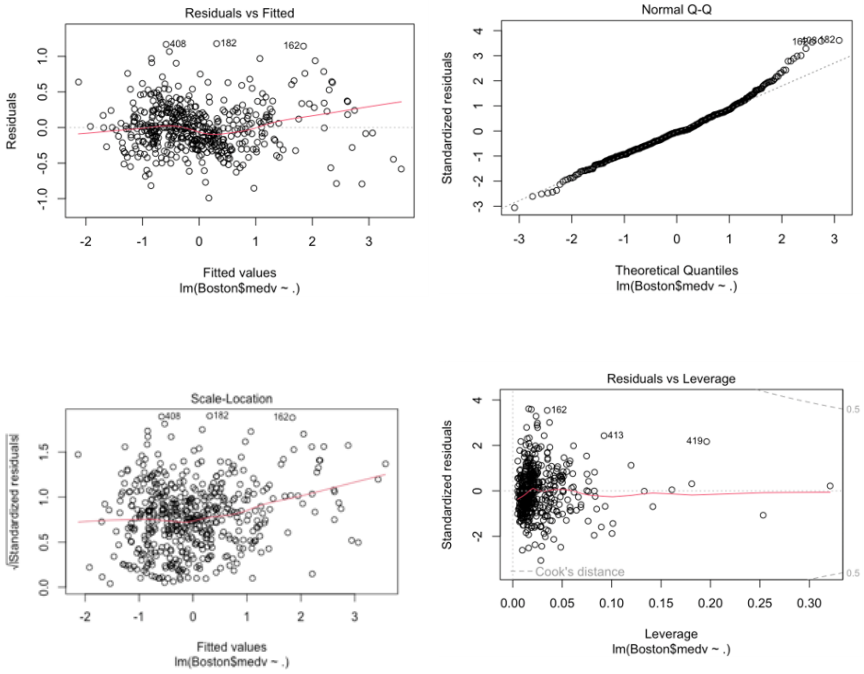


图 27: model_4 的残差分析

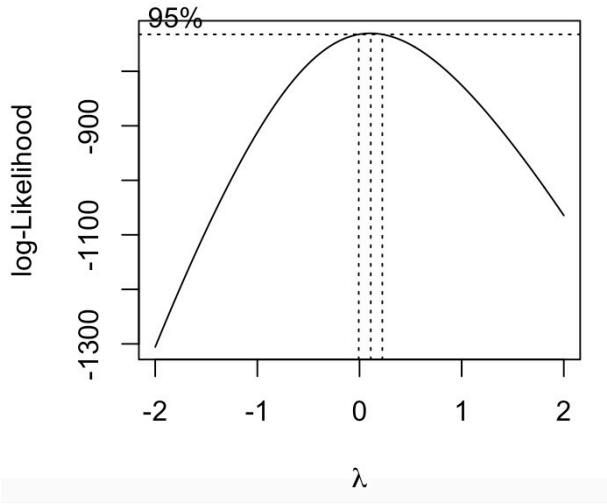


图 28: 对数似然函数随 λ 增长的变化