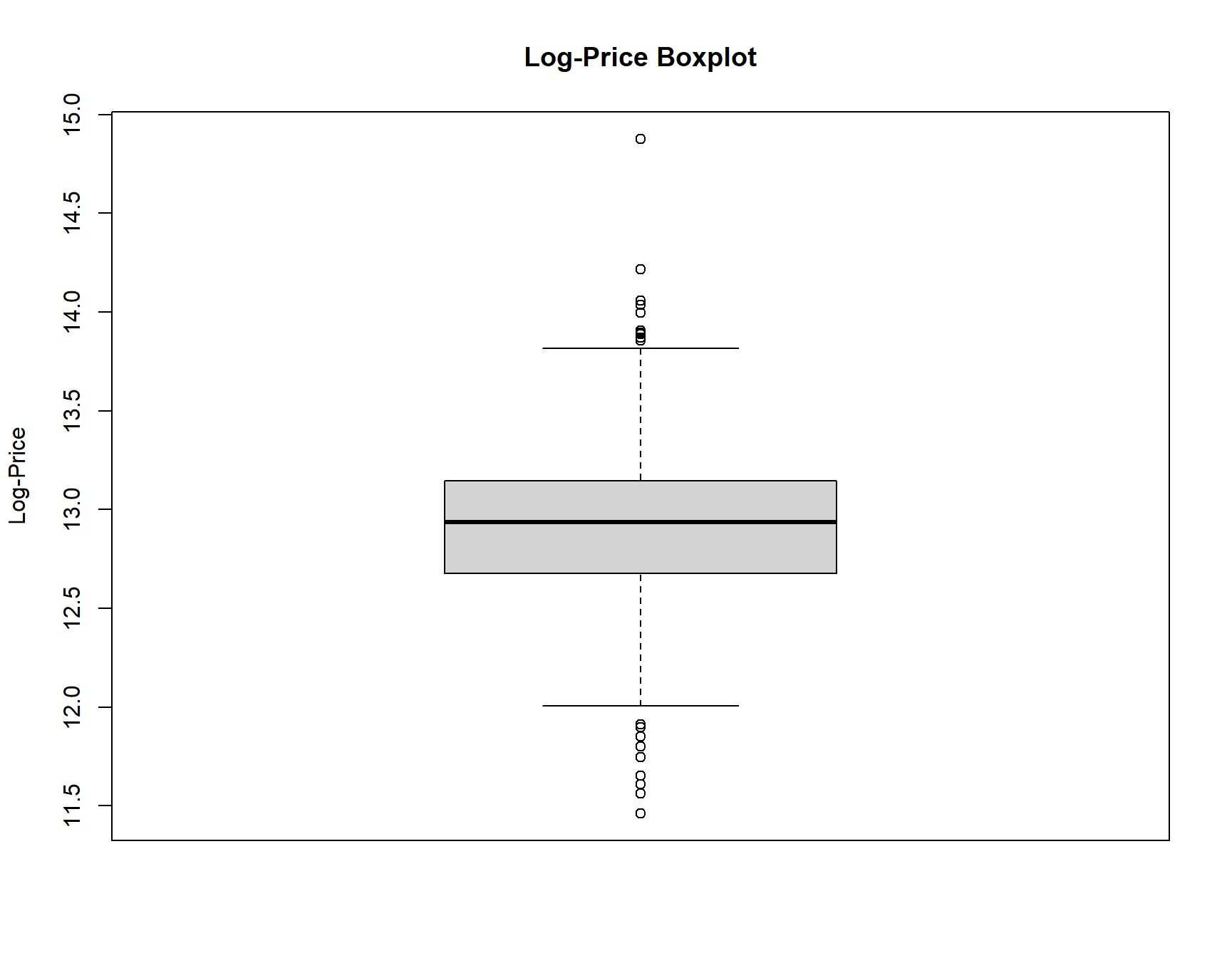
## 2.1 主要函数说明及参数设置

定义对数线性回归模型为

其中为常数项，为第i个自变量）的回归系数。

## 2.2 异常值处理

将961条数据读入后将price值取对数得到因变量，剔除缺失值后，对因变量绘制箱线图进行异常值检索，结果如下。



图x.x 帆船对数价格箱线图

删除离群点后共941条数据。同样地，数据集中变量ID与后续数据分析无关，因此做删除处理。

## 3.1 模型建立

## 以帆船价格的对数为响应变量，其余所有变量为解释变量进行对数线性回归模型model的建立，并对模型效果进行查看，结果如下。

表x.x：model相关参数分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 变量 | 估计系数 | 标准误 | t值 | p值 | 显著性 |
| 常数 | -9.059e+01 | 2.685e+00 | -33.745 | < 2e-16 | \*\*\* |
| Length | 3.190e-02 | 5.593e-03 | 5.705 | 1.57e-08 | \*\*\* |
| Year | 4.977e-02 | 1.327e-03 | 37.515 | < 2e-16 | \*\*\* |
| LWL | 1.748e-02 | 5.629e-03 | 3.105 | 0.00196 | \*\* |
| Beam | 3.817e-02 | 7.128e-03 | 5.355 | 1.08e-07 | \*\*\* |
| Draft | 2.631e-02 | 1.565e-02 | 1.681 | 0.09302 | . |
| Displacement | -1.917e-06 | 1.025e-06 | -1.871 | 0.06172 | . |
| SailArea | 1.301e-04 | 5.631e-05 | 2.310 | 0.02108 | \* |
| GDP | 3.811e-05 | 6.731e-06 | 5.662 | 1.99e-08 | \*\*\* |
| GDP.Capita | 1.299e-06 | 4.434e-07 | 2.930 | 0.00347 | \*\* |
| =0.7887 调整=0.7866 F(9,931)=386.1 p<2.2e-16 | | | | | |

分析上述表格，由和调整值接近于1，可知本模型拟合效果良好，且p值小于0.05，说明在检验水平为0.05条件下，因变量与其他变量之间的线性关系显著。

## 3.2 模型改进

运用方差膨胀因子法对本模型进行多重共线性诊断，检验结果如下：

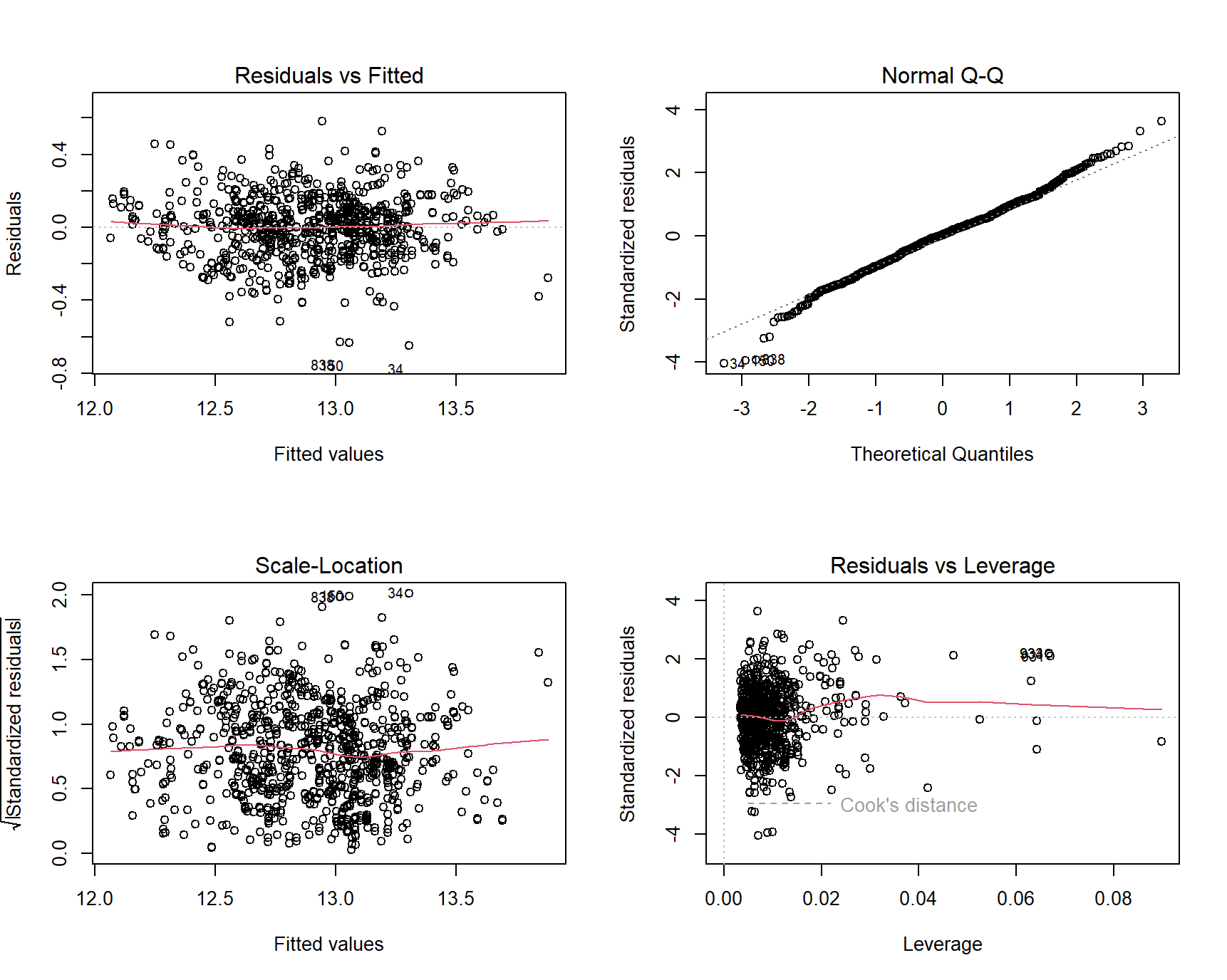
表x.x：方差膨胀因子法

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 变量 | Length | Year | LWL | Beam | Draft |
| vif值 | 15.04 | 1.17 | 10.40 | 4.51 | 1.49 |
| 变量 | Displacement | SailArea | GDP | GDP.Capita |  |
| vif值 | 5.10 | 7.90 | 1.01 | 1.41 |  |

由检验结果知，变量Length与LWL之间存在严重的多重共线性，且Length的vif值最大，因此考虑将其从模型中移除，进行重新建模得到model1。下面将展示对模型model1潜在问题的诊断及修正过程。

## 3.3 模型诊断

## 3.3.1 整体诊断



图x.x.x 模型诊断图

绘制线性模型的整体散点图进行模型整体分析。观察左下图可知散点分布基本随机；观察Q-Q图可知，随横坐标增大，散点分散近似成一条直线，因此可认为残差服从正态分布；观察右下图可知存在较多离群点。下面将对模型进行检验。

## 3.3.2 残差正态性检验

由上面残差Q-Q图可直观判断，残差服从正态分布，下面针对模型model1的标准化残差进行K-S检验，检验结果如下。

表x.x.x：K-S检验

|  |  |  |
| --- | --- | --- |
| 检验变量 | 统计量D值 | P值 |
| lm1模型残差（标准化） | 0.035694 | 0.1817 |

由检验结果，检验p值并不显著，因此在检验水平为0.05条件下不能拒绝原假设，即模型model1的标准化残差服从正态分布。

## 3.3.3 方差齐性检验

1）异方差检验

由整体检验时已经分析残差图，散点随机性较良好，因此直观认为残差之间不存在异方差，对模型采用ncv计分检验，检验结果如下表。

表x.x.x：ncv检验

|  |  |  |  |
| --- | --- | --- | --- |
| 检验变量 | Chisquare | df | p |
| lm1 | 1.399706 | 1 | 0.23677 |

由检验p值大于0.05可认为检验不显著，则认为不存在异方差。

## 3.3.4 多重共线性诊断

运用方差膨因子法对模型model1进行多重共线性诊断，结果如下：

表x.x.x：方差膨胀因子法

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量 | Year | LWL | Beam | Displacement | GDP | SailArea | Draft | GDP.Capita |
| vif值 | 1.14 | 5.53 | 4.07 | 4.36 | 1.39 | 7.77 | 1.49 | 1.41 |

由检验结果知，各解释变量之间不存在严重的多重共线性，即多重共线性检验通过。

## 3.4 模型应用

经上述检验，可确定最终模型model1，对其效果进行查看，结果如下。

表x.x：model1相关参数分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 变量 | 估计系数 | 标准误 | t值 | p值 | 显著性 |
| 常数 | -8.821e+01 | 2.696e+00 | -32.714 | < 2e-16 | \*\*\* |
| Year | 4.859e-02 | 1.332e-03 | 36.467 | < 2e-16 | \*\*\* |
| LWL | 3.947e-02 | 4.171e-03 | 9.463 | < 2e-16 | \*\*\* |
| Beam | 5.093e-02 | 6.882e-03 | 7.400 | 3.03e-13 | \*\*\* |
| Draft | 3.092e-02 | 1.589e-02 | 1.946 | 0.05192 | . |
| Displacement | 3.081e-07 | 9.639e-07 | 0.320 | 0.74927 |  |
| SailArea | 1.716e-04 | 5.678e-05 | 3.023 | 0.00257 | \*\* |
| GDP | 3.952e-05 | 6.840e-06 | 5.778 | 1.03e-08 | \*\*\* |
| GDP.Capita | 1.287e-06 | 4.509e-07 | 2.856 | 0.00439 | \*\* |
| = 0.7813 调整= 0.7794 F(8,932)=416.2 p<2.2e-16 | | | | | |

由于与调整的值会随着变量个数的减少而减小，model1的和调整略低于改进前模型model，而model1的和调整均大于线性回归模型lm2，且p值仍旧显著，证明模型基本可靠。建模最终得到回归方程

该模型输出的因变量为价格的对数，在进行预测错误率计算之前，需要将其通过指数函数转化为价格数据，以此代入进行计算。计算出该对数线性回归模型的预测错误率为，相较于线性回归模型没有改善。