2017－2018 年第一学期

《大数据统计基础》试题　答题纸

学校：中央财经大学　　学号：2017210785　　姓名：司徒雪颖　　成绩_____

二、数据预处理

1.（1）对数据作图估计预测变量和被解释变量之间的函数关系。

以 Rings 为纵轴，每个定量自变量分别与 Rings 作散点图，定性变量与 Rings 作箱线图。

从图 2-1 中可以看出，Height 与 Rings 呈现明显的线型正相关关系，其他定量变量虽然也是与 Rings 有线型关系，但同时也有明显的异方差，即 Rings 取值越大，自变量取值越离散。对于定性变量 type，type 取 I 值时 Rings 的均值比取 F、M 值时小。

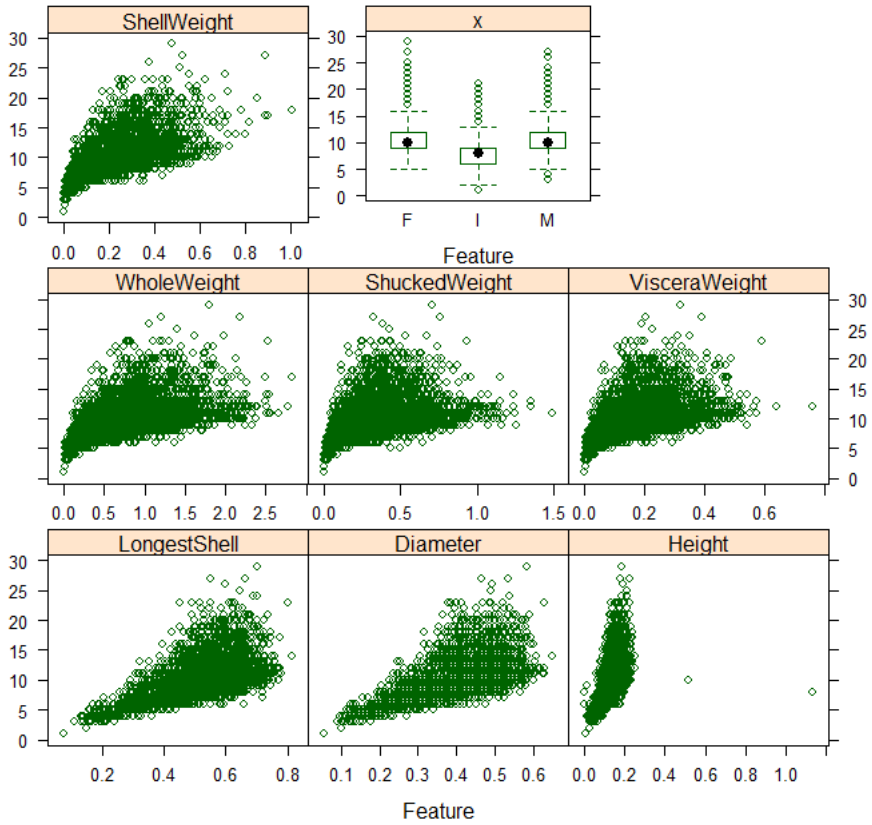

图 2- 1　各个自变量与因变量的散点图

（2）用散点图和相关系数图解释预测变量之间的相关性。

由相关矩阵图 2-2 和散点矩阵图 2-3 中可以看出，各个自变量之间高度正相关，其中

longestshell 与 diameter 的相关系数达到 0.99！

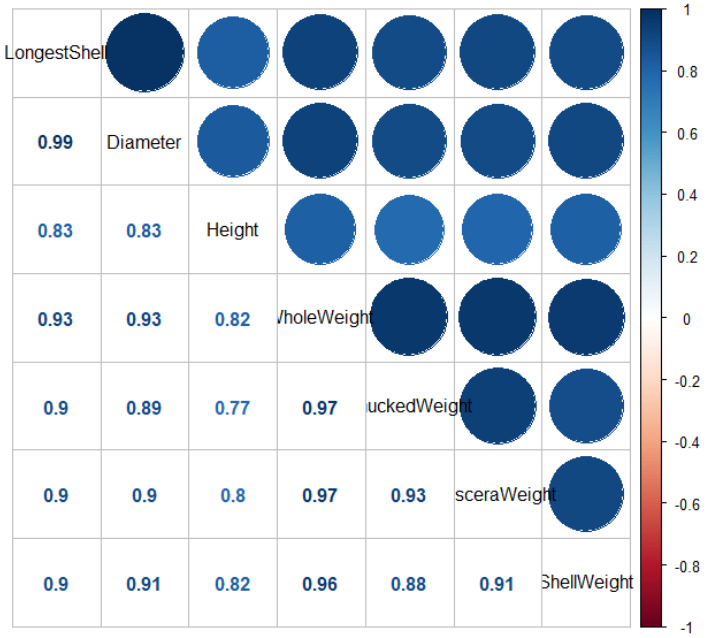

图 2- 2 各个自变量之间的相关图矩阵

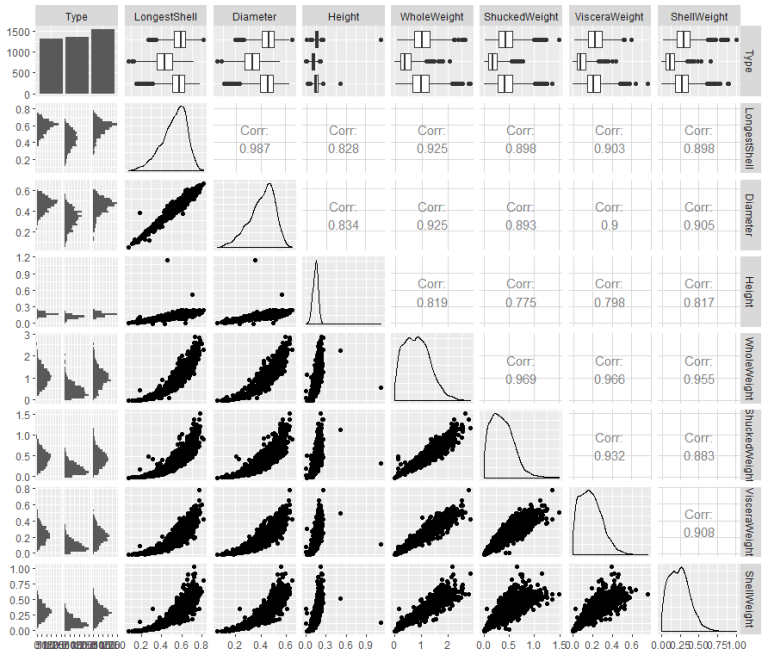

图 2- 3 自变量之间的散点图矩阵

（3）对预测变量估计重要性得分。找到一种筛选方法得到预测变量子集，该集合不含冗余变量。

A．估计重要性得分

用随机森林评估各自变量重要性得分，由表 2-1 可以看出，ShuckedWeight 和 type 对

于建立模型的作用最大。

表 2- 1 随机森林评估自变量重要性得分

排名	变量名	重要性指数
1	ShuckedWeight	78.944
2	Type	76.65459
3	ShellWeight	61.76333
4	Height	50.13624
5	VisceraWeight	45.80616
6	Diameter	39.77014
7	WholeWeight	39.27629
8	LongestShell	34.97381

B．找到一种筛选方法得到预测变量子集

先删除相关系数达 95%以上的变量 WholeWeight Diameter，再使用逐步回归筛选变量，剔除了 VisceraWeight。因此筛选出的变量为 Type LongestShell，Height ShuckedWeight，ShellWeight，这些变量的系数均显著，因此没有冗余变量。

```
> summary(m3)

Call:
lm(formula = Rings ~ Type + LongestShell + Height + ShuckedWeight +
    ShellWeight, data = abalone[, -c(highcor + 1)])

Residuals:
     Min       1Q   Median       3Q      Max
-11.6038  -1.3420  -0.3616   0.8553  16.1217

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.87886    0.29453  13.170  < 2e-16 ***
TypeI          -0.91937    0.10310  -8.918  < 2e-16 ***
TypeM           0.04370    0.08495   0.514    0.607
LongestShell    7.62643    0.80822   9.436  < 2e-16 ***
Height         11.85475    1.55624   7.618 3.18e-14 ***
ShuckedWeight -11.61793    0.38751 -29.981  < 2e-16 ***
ShellWeight    20.34578    0.64303  31.641  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.241 on 4170 degrees of freedom
Multiple R-squared:  0.5178,    Adjusted R-squared:  0.5171
F-statistic: 746.3 on 6 and 4170 DF,  p-value: < 2.2e-16
```

（4）对连续型预测变量应用主成分分析，决定多少个不相关的主成分能够代表数据中的信息？

由碎石图 2-4 可以看出，只需要选取一个主成分。且第一个主成分的方差贡献率就已经

达到 97%

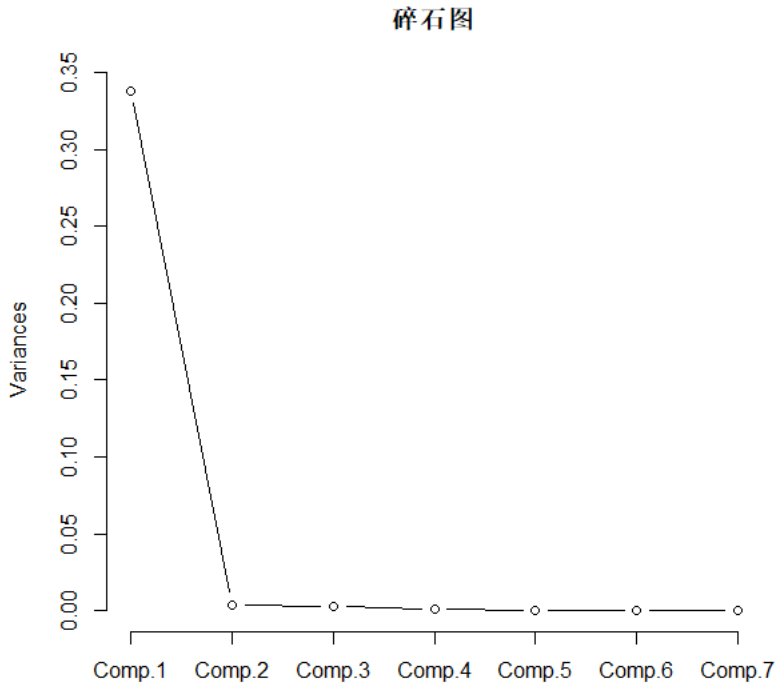

图 2- 4 碎石图

表 2- 2 方差贡献表

	Comp.1	Comp.2	Comp.3
Standard deviation	0.581455	0.062953	0.053917
Proportion of Variance	0.974101	0.011418	0.008376
Cumulative Proportion	0.974101	0.985519	0.993895

2.（1）写一个 R 函数从该模型中模拟数据。

```
sim = function(n)
{
  set.seed(1994)
  res = matrix(0,n,6)
  res = as.data.frame(res)
  colnames(res) = c("x1","x2","x3","x4","x5","y")
  for(i in 1:5)
  {
    res[,i] = runif(n,0,1)
  }
  res$y = 10*sin(pi*res$x1*res$x2)+20*(res$x3-
0.5)^2+10*res$x4+5*res$x5+rnorm(n,0,1)
```

```
  return(res)
}
```

（2）随机模拟一个数据集，样本量是 500，绘制图形研究预测变量和被解释变量之间的关系。

从散点图 2-5 中可以看出，除了 x3，其他自变量与因变量都呈线性正相关关系，其中 x4 最明显。

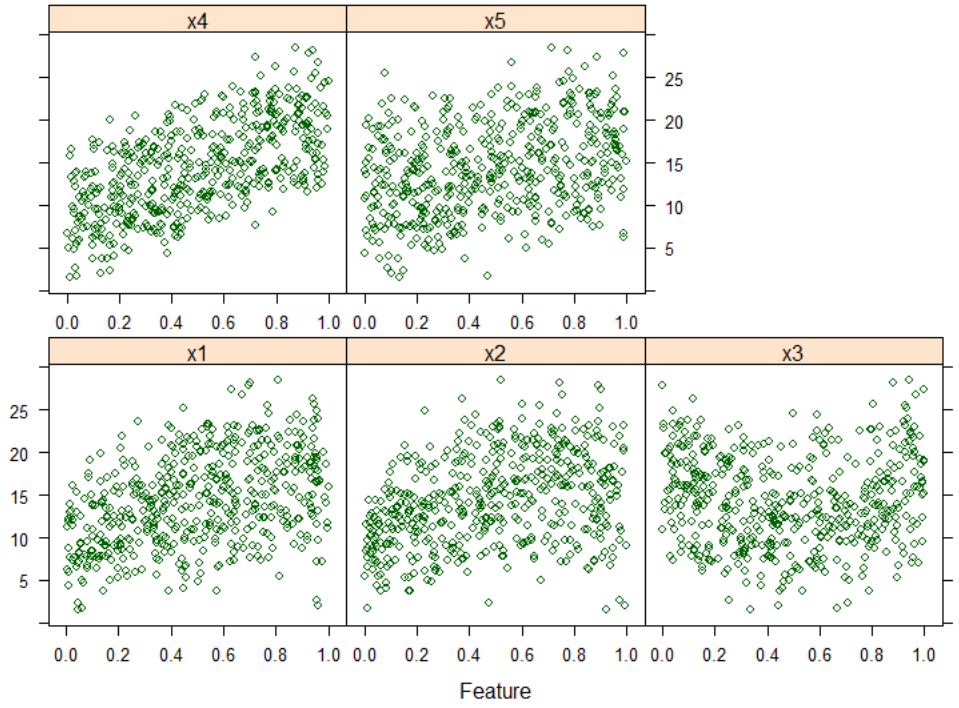

图 2- 5 X 关于 Y 的散点图

（3）使用线性回归中的向前法、向后法和逐步回归等变量选择方法，最终模型选择了哪些变量？

无论是向前、向后、逐步回归，三种模型都选择了所有变量，一个都没有剔除。

```
> library(MASS)
> m1 = stepAIC(initial,direction ="forward",trace = 1)
Start:  AIC=994.61
y ~ x1 + x2 + x3 + x4 + x5

> m2 = stepAIC(initial,direction ="backward",trace = 1)
Start:  AIC=994.61
y ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq    RSS     AIC
<none>              3568.2  994.61
- x3    1      19.2 3587.4  995.28
- x5    1    1315.4 4883.6 1149.52
- x2    1    1879.9 5448.1 1204.21
- x1    1    2233.9 5802.2 1235.69
- x4    1    4565.5 8133.7 1404.58
> m3 = stepAIC(initial,direction = "both",trace = 1)
Start:  AIC=994.61
y ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq    RSS     AIC
<none>              3568.2  994.61
- x3    1      19.2 3587.4  995.28
- x5    1    1315.4 4883.6 1149.52
- x2    1    1879.9 5448.1 1204.21
- x1    1    2233.9 5802.2 1235.69
- x4    1    4565.5 8133.7 1404.58
```

（4）应用不同的过滤法，逐个评估变量。一些过滤法同时评估多个变量（如 ReliefF 算法），两个有交互效应的预测变量x_1和x_2否被选中了？是否倾向于选择其中某一个变量？

单独评估每个自变量的方法。Loess 方法的 pseudo-R2 排序，则 x4 重要性得分最高，x3，x5 最低；MIC 系数排序，则还是 x4 最高，x5,x3 最低。

表 2- 3 重要性得分排序

vars	loess	vars	MIC
x4	0.389966	x4	0.4312
x2	0.187934	x2	0.251706
x1	0.183782	x1	0.2463
x3	0.120801	x5	0.225878
x5	0.109036	x3	0.218269

使用过滤法中 sbf(方法为随机森林)同时评估多个变量，则选中的变量为 x1, x2, x4, x5，这四个变量在重抽样中选中的概率为 100%。X1,x2 确实同时选中了。

```
Selection By Filter

Outer resampling method: Cross-Validated (10 fold, repeated 5 times)

Resampling performance:

  RMSE Rsquared   MAE RMSESD RsquaredSD MAESD
 2.431   0.8241 1.949 0.1984    0.03599 0.187

Using the training set, 4 variables were selected:
   x1, x2, x4, x5.

During resampling, the top 4 selected variables (out of a possible 4):
   x1 (100%), x2 (100%), x4 (100%), x5 (100%)
```

使用封装法中的 rfe（方法为随机森林）同时评估多个变量，则选中的变量为 x1,x2,x3,x4,x5。X1,x2 确实同时选中了。

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold, repeated 5 times)

Resampling performance over subset size:

 Variables  RMSE Rsquared   MAE RMSESD RsquaredSD  MAESD Selected
         1 4.873   0.2482 3.980 0.4164    0.08428 0.3590
         2 3.921   0.4700 3.152 0.3394    0.08466 0.2943
         3 2.632   0.7659 2.111 0.2724    0.04376 0.2177
         4 2.434   0.8229 1.949 0.2628    0.03226 0.2070
         5 2.241   0.8970 1.795 0.2357    0.01990 0.1783        *

The top 5 variables (out of 5):
   x4, x1, x2, x5, x3
```