# 《机器学习》课程第 3 次作业

姓名：刘哲　　学号：2022103691

## 1　Auto

### a

Auto数据集中的所有变量的散点图矩阵如下：


其中，mpg是每加仑汽油英里数，cyl是气缸数，displace是发动机排量，hp是发动机马力， weight是车重，acc是60英里加速时间，year是出品年份，origin是汽车原产地。

### b

排除车名变量，其余各变量之间的相关系数矩阵为

| | mpg | cyl | displace | hp | weight | acc | year | origin |
|---|---|---|---|---|---|---|---|---|
| mpg | 1 | -0.7776 | -0.8051 | -0.7784 | -0.8322 | 0.4233 | 0.5805 | 0.5652 |
| cylinders | -0.7776 | 1 | 0.9508 | 0.8430 | 0.8975 | -0.5047 | -0.3456 | -0.5689 |
| displacement | -0.8051 | 0.9508 | 1 | 0.8973 | 0.9330 | -0.5438 | -0.3699 | -0.6145 |
| horsepower | -0.7784 | 0.8430 | 0.8973 | 1 | 0.8645 | -0.6892 | -0.4164 | -0.4552 |
| weight | -0.8322 | 0.8975 | 0.9330 | 0.8645 | 1 | -0.4168 | -0.3091 | -0.5850 |
| acceleration | 0.4233 | -0.5047 | -0.5438 | -0.6892 | -0.4168 | 1 | 0.2903 | 0.2127 |
| year | 0.5805 | -0.3456 | -0.3699 | -0.4164 | -0.3091 | 0.2903 | 1 | 0.1815 |
| origin | 0.5652 | -0.5689 | -0.6145 | -0.4552 | -0.5850 | 0.2127 | 0.1815 | 1 |

**c**

以mpg作为响应变量，进行多元线性回归。由于origin是分类变量，将其转换为哑变量(American, European)。其中，(1, 0)表示origin=1，即汽车原产地为American；(0, 1)表示origin=2，即汽车原产地为European；(0, 0)表示origin=3，即汽车原产地为Janpanese。回归结果如下：

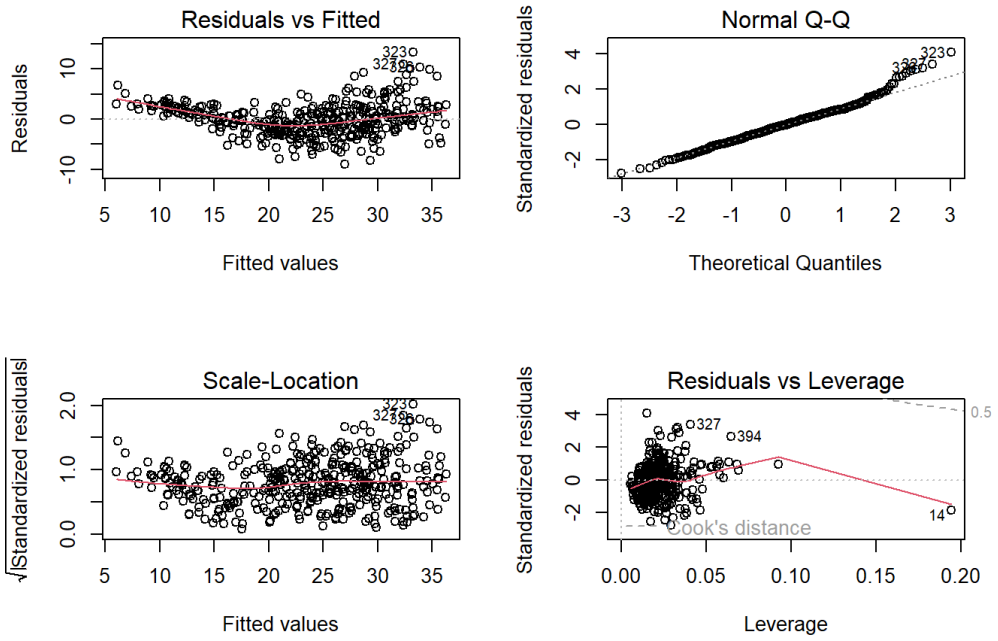
```
##
## Call:
## lm(formula = mpg ~ ., data = auto.dum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.510e+01  4.681e+00  -3.226  0.00136 **
## cylinders    -4.897e-01  3.212e-01  -1.524  0.12821
## displacement  2.398e-02  7.653e-03   3.133  0.00186 **
## horsepower   -1.818e-02  1.371e-02  -1.326  0.18549
## weight       -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration  7.910e-02  9.822e-02   0.805  0.42110
## year          7.770e-01  5.178e-02  15.005  < 2e-16 ***
## American     -2.853e+00  5.527e-01  -5.162 3.93e-07 ***
## European     -2.232e-01  5.661e-01  -0.394  0.69355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

由回归方程F检验的p-value极小可知，回归方程显著，但并不是所有预测变量都与响应变量具有显著相关性。其中预测变量displacement、weight、year、American和截距项的回归系数显著，说明其与mpg在统计上具有显著关系。

发动机排量（displacement）具有正回归系数，表明发动机排量越大，mpg越大；车重（weight）具有负回归系数，表明汽车越重，mpg越小；出品年份（year）具有正回归系数，表明新款汽车的mpg更大，汽车的mpg具有逐年上升的趋势；哑变量American具有负回归系数，表明美国产汽车的mpg更小。

**d**

多元线性回归的诊断图如下：




由Residuals图可以看出，残差点基本上在一条水平直线附近，但当拟合值较大时，残差点变得更加分散，出现大量离群点。 Normal Q-Q图基本上形成了一条直线，表明拟合残差满足正态性。 Scale-Location图中出现一条水平直线，表明残差沿着预测值的范围平均分布，符合同方差性假设。 Leverage图显示了离群点对回归的影响，因为图中几乎看不到Cook距离线，说明所有样本点都在Cook距离线内部，表明不存在异常高杠杆作用的点，离群点对回归几乎没有影响。

**e**

使用逐步选择法建立回归模型，结果如下：

```
##
## Call:
## lm(formula = mpg ~ weight + year + American + displacement +
##     horsepower + cylinders, data = auto.dum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0599 -2.0866 -0.1227  1.9076 13.5115
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -13.839999   4.113203  -3.365 0.000843 ***
## weight        -0.006505   0.000564 -11.534  < 2e-16 ***
## year           0.777686   0.050646  15.355  < 2e-16 ***
## American      -2.751258   0.481795  -5.710 2.26e-08 ***
## displacement   0.023493   0.007598   3.092 0.002133 **
## horsepower    -0.024557   0.010705  -2.294 0.022329 *
## cylinders     -0.496154   0.319883  -1.551 0.121712
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.301 on 385 degrees of freedom
## Multiple R-squared:  0.8238, Adjusted R-squared:  0.8211
## F-statistic: 300.1 on 6 and 385 DF,  p-value: < 2.2e-16
```

根据回归系数的显著性，选择最佳变量weight、year、American建立回归模型，结果如下：

```
##
## Call:
## lm(formula = mpg ~ weight + year + American, data = auto.dum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4841 -2.1141 -0.0192  1.7795 13.6261
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.639e+01  3.921e+00  -4.180 3.60e-05 ***
## weight      -5.893e-03  2.593e-04 -22.731  < 2e-16 ***
## year         7.725e-01  4.823e-02  16.017  < 2e-16 ***
## American    -2.095e+00  4.361e-01  -4.804 2.22e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.334 on 388 degrees of freedom
## Multiple R-squared:  0.8189, Adjusted R-squared:  0.8176
## F-statistic:   585 on 3 and 388 DF,  p-value: < 2.2e-16
```

使用连续正交法得到每个预测变量在排除其他预测变量的干扰后，对响应变量的相关系数。

| var | coef |
|---|---|
| **cylinders** | **-3.5581** |
| displacement | -0.0511 |
| horsepower | -0.0599 |
| weight | -0.0053 |
| acceleration | -0.0291 |
| **year** | **0.7534** |
| **American** | **-2.7469** |
| European | -0.2232 |

系数绝对值的大小代表了对应预测变量对响应变量的影响大小，因此选择最佳变量cylinders、year、American建立回归模型，结果如下：

```
##
## Call:
## lm(formula = mpg ~ cylinders + American + year, data = auto.dum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.0529  -2.5583  -0.2539   2.2336  14.5046
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.85015    4.79913  -4.345 1.78e-05 ***
## cylinders    -2.44655    0.15958 -15.331  < 2e-16 ***
## American     -3.03313    0.53189  -5.703 2.34e-08 ***
## year          0.78415    0.05908  13.274  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.017 on 388 degrees of freedom
## Multiple R-squared:  0.7371, Adjusted R-squared:  0.7351
## F-statistic: 362.6 on 3 and 388 DF,  p-value: < 2.2e-16
```

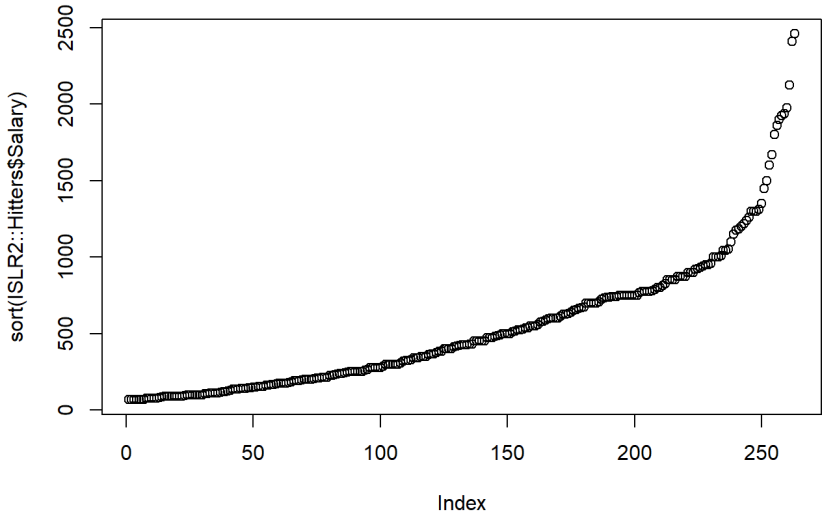
比较两个回归模型，根据逐步选择法筛选出变量，由此建立的回归模型具有更大的$R^2$和调整$R^2$，认为使用变量weight、year、American作为预测变量建立的回归模型具有更优的性能。

# 2 Hitters

### a

Hitters 数据集中有 59 名球员的 Salary 变量为空，不利于后续基于 Salary 的分析研究，故将这部分球员删去。

将剩余样本的 Salary 变量排序，其散点图如下：




选取阈值为 500，即认为 Salary 大于 500 的击球手属于高薪球员，记为1；其余则属于低薪球员，记为0。数据集中球员薪资水平的分布如下：

| Salary | 1 | 0 |
|---|---|---|
| n | 112 | 151 |

属于高薪和低薪的球员数量相近，有利于建立二分类模型。

**b**

使用逐步选择法建立回归模型，结果如下：

```
##
## Call:
## lm(formula = Salary ~ CHits + Hits + Division + NewLeague + Walks,
##     data = hitters.class)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97970 -0.26103 -0.05952  0.31275  1.33527
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.447e-01  7.008e-02  -3.491 0.000566 ***
## CHits        3.092e-04  3.834e-05   8.063 2.83e-14 ***
## Hits         3.219e-03  6.590e-04   4.885 1.82e-06 ***
## Division     1.114e-01  4.771e-02   2.334 0.020372 *
## NewLeague   -1.072e-01  4.790e-02  -2.239 0.026023 *
## Walks        2.505e-03  1.375e-03   1.822 0.069630 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3853 on 257 degrees of freedom
## Multiple R-squared:  0.4067, Adjusted R-squared:  0.3951
## F-statistic: 35.23 on 5 and 257 DF,  p-value: < 2.2e-16
```

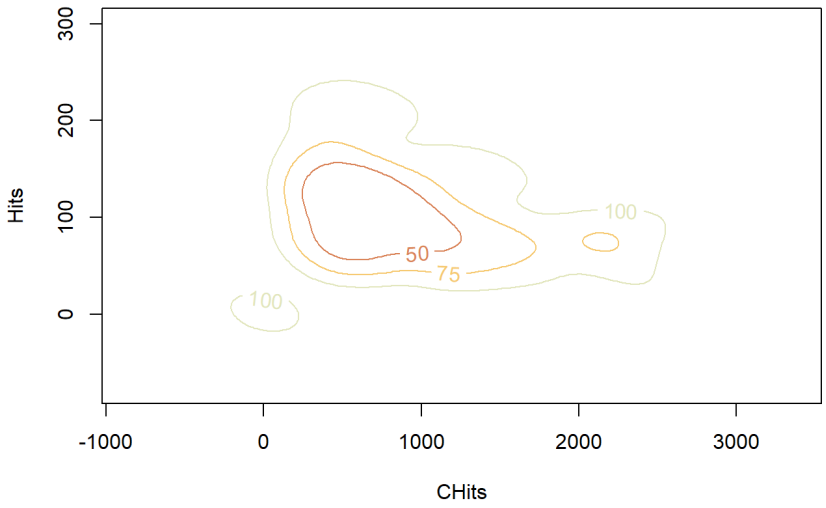
选择其中显著性最高的两个变量 CHits 和 Hits 建立 logistic 回归模型，结果如下：

```
##
## Call:
## glm(formula = Salary ~ CHits + Hits, family = binomial(link = "logit"),
##     data = hitters.class)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4942  -0.6542  -0.3463   0.7403   2.9237
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.3009111  0.5536668  -7.768 7.97e-15 ***
## CHits        0.0020230  0.0003179   6.364 1.97e-10 ***
## Hits         0.0226544  0.0040287   5.623 1.87e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 358.79  on 262  degrees of freedom
## Residual deviance: 241.51  on 260  degrees of freedom
## AIC: 247.51
##
## Number of Fisher Scoring iterations: 5
```

显然，logistic 回归模型的参数均显著，且模型准确率为0.8137。

**c**

共有49个球员的薪资水平被预测错误，其中将低薪球员预测成高薪球员的有20个，将高薪球员预测成低薪球员的有29个。对预测错的样本进行核密度估计，两个变量 CHits 和 Hits 的分布图如下：




从图中可以看出，出错点在 CHits 和 Hits 上覆盖的范围比较广，而且形状不连续，另外模型本身的预测准确度已经不低，由此认为预测出错的原因很可能是出现离群值。