第六次生统作业

一、某地29名13岁儿童身高（cm），体重（kg）和肺活量（L）.数据见“homework-6.1-data.xlsx”，求：

（1）由身高，体重推算肺活量的回归方程；

（2）求出的方程是否有意义；

（3）剩余标准差

二、某农场通过试验取得早稻收获量与春季降雨量和春季温度的数据如下：

数据见“homework-6.2-data.csv”

|  |  |  |
| --- | --- | --- |
| 收获量y(kg/mm2) | 降雨量x1(mm) | 温度x2(℃) |
| 2250 | 25 | 6 |
| 3450 | 33 | 8 |
| 4500 | 45 | 10 |
| 6750 | 105 | 13 |
| 7200 | 110 | 14 |
| 7500 | 115 | 16 |
| 8250 | 120 | 17 |

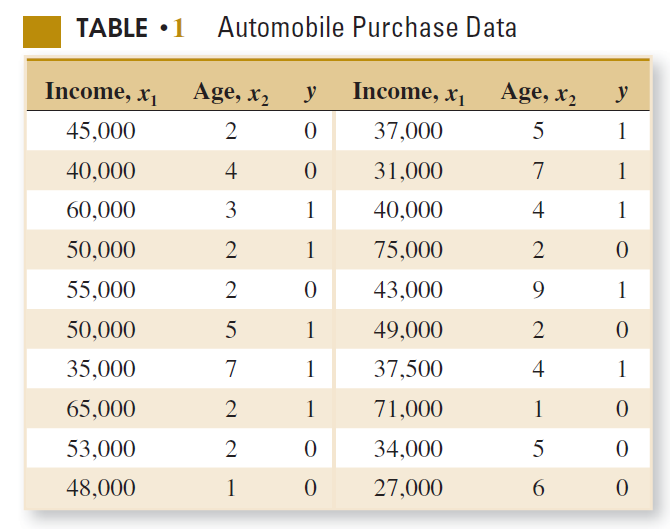
建立早稻收获量对春季降雨量和春季温度的二元线性回归方程，计算各回归系数的置信区间，并对回归模型的线性关系和回归系数进行检验（α=0.05）。

三、某葡萄酒爱好者想探索葡萄酒的品质与哪些因素相关。他有一个数据集包含了（1 -固定酸度，2 -挥发性酸度，3 -柠檬酸，4 -残余糖，5 -氯化物，6 -自由二氧化硫量，7 -二氧化硫总量，8 -密度，9 - pH值，10 -硫酸盐，11 -酒精浓度，和12 -品质(0 - 10分)。数据见“homework-6.3-winequality-red.csv” 。

1. 查看数据集的前五行和数据集的总结
2. 通过直方图展示固定酸度的分布和展示挥发性酸度与品质的散点图
3. 计算这些变量与品质的相关性
4. 通过方差分析不同品质的葡萄酒的酒精浓度是否有差异
5. 通过多元线性回归建立一个品质预测模型，并说明哪些变量与品质显著相关。

四、A study was performed to investigate new automobile purchases. A sample of 20 families was selected. Each family was surveyed to determine the age of their oldest vehicle and

their total family income. A follow-up survey was conducted six months later to determine if they had actually purchased a new vehicle during that time period (*y* = 1 indicates *yes* and *y* = 0 indicates *no*). The data from this study are shown in the Table1(“homework-6.4-data.txt”).



(a) Fit a logistic regression model to the data.(40’)

(b) Interpret the model coefficients β1 and β2 and write the logistic regression model formula.(20’)

(c) What is the estimated probability that a family with an income of $45,000 and a car that is five years old will purchase a new vehicle in the next six months?(40’)

五、数据文件“homework-6.5-Drivers.csv”为对45名司机的调查结果，其中四个变量的含义为：

1）x1：表示视力状况，它是一个分类变量，1表示好，0表示有问题；

2）x2：年龄，数值型；

3）x3：驾车教育，它也是一个分类变量，1表示参加过驾车教育，0表示没有；

4）y：一个分类型输出变量，表示去年是否出过事故，1表示出过事故，0表示没有；

问题：

1. 请在R语言中调用logistic回归函数，计算视力状况、年龄、驾车教育与是否发生事故的logistic回归模型，并以“odds=……”的形式写出回归公式。（10分）
2. 指出（1）得到的模型中哪些因素对是否发生事故有显著性影响。如果存在对是否发生事故没有显著性影响的因素，请去除这些因素后重新计算logistic回归模型，并以“p=……”的形式写出回归公式。（20分）
3. A是一名参加过驾车教育，但视力有问题的50岁老司机；B是一名没有参加过驾车教育，但视力良好的20岁新手。现在A、B都想在某保险公司投保，但按公司规定，被保险人必须满足“明年出事故的概率不高于40%”的条件才能予以承保。请预测A、B两者明年出事故的概率，并告诉保险公司谁可以投保。（20分）

六、Many digitized image of a fine needle aspirate (FNA) of a breast mass are collected and computed to predict the diagnosis of breast cancer((“homework-6.6-data.csv”)).

Attribute information

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the threelargest values) of these features were computed for each image,resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recorded with four significant digits.

In total, there are 357 benign and 212 malignant samples.

You may need to use proper regression algorithm to train your data, and make predictions.

Instructions:

1. Use all mean features(such as: radius\_mean,texture\_mean…) to construct a logistic regression model
2. Then try to reduce the number of features from your last model, construct another regression model, and you will need to write down the equation of your logistic regression model(Tips: Logit P = α+β1X1+β2X2+..+βpXp)
3. Use proper test to test the difference between two models
4. You may split the data properly, use part of them to train your regression model and use another part to make predictions. Lastly, you may try to calculate the accuracy of your model.(Tips: To split the data, you can use the first 398 rows as training data, use the last 171 rows as prediction data.The predict function return a value between 0 and 1, 0.~0.5 belong to the first class, and 0.5~1 belong to second class in binary classification problems)