回归分析作业

蹇傲霖 2018010919

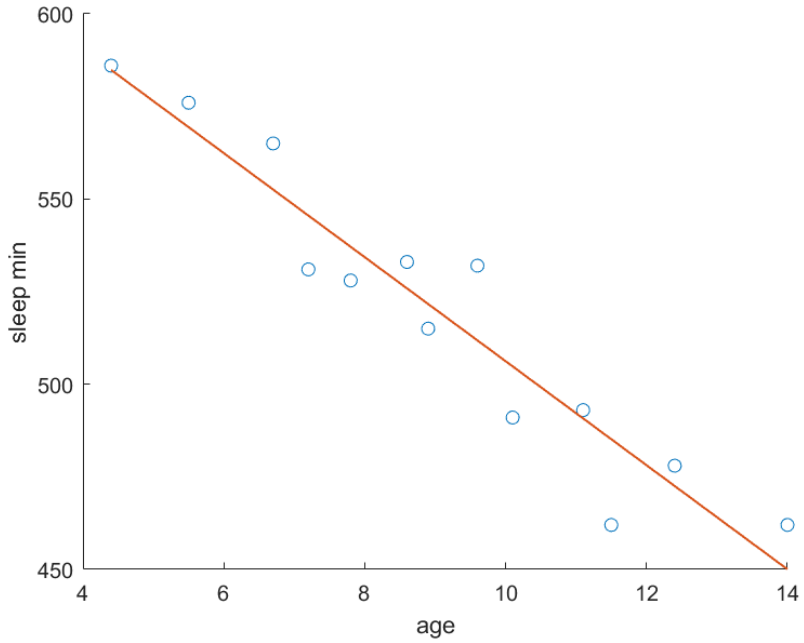
1.13 名少年儿童参加了一项睡眠时间与年龄关系的调查，下表中的（每天）睡眠时间（min）是根据连续 3 天记录的平均时间得到的。

序号	睡眠时间（min）	年龄（岁）	序号	睡眠时间（min）	年龄（岁）
1	586	4.4	8	515	8.9
2	462	14.0	9	493	11.1
3	491	10.1	10	528	7.8
4	565	6.7	11	576	5.5
5	462	11.5	12	533	8.6
6	532	9.6	13	531	7.2
7	478	12.4			

（1） 画出散点图，建立回归模型并检验模型的有效性，解释所得结果。

设年龄为自变量 x，睡眠分钟数为因变量 y。

根据所给数据，绘出散点图和线性回归直线。



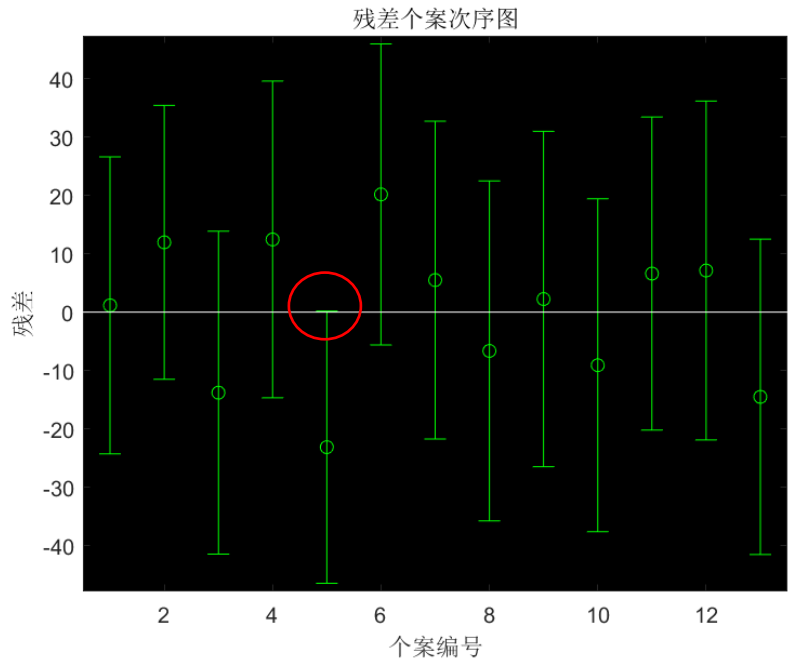
$$y = 646.6229 - 14.0416x$$

线性回归的各项统计量如下表：

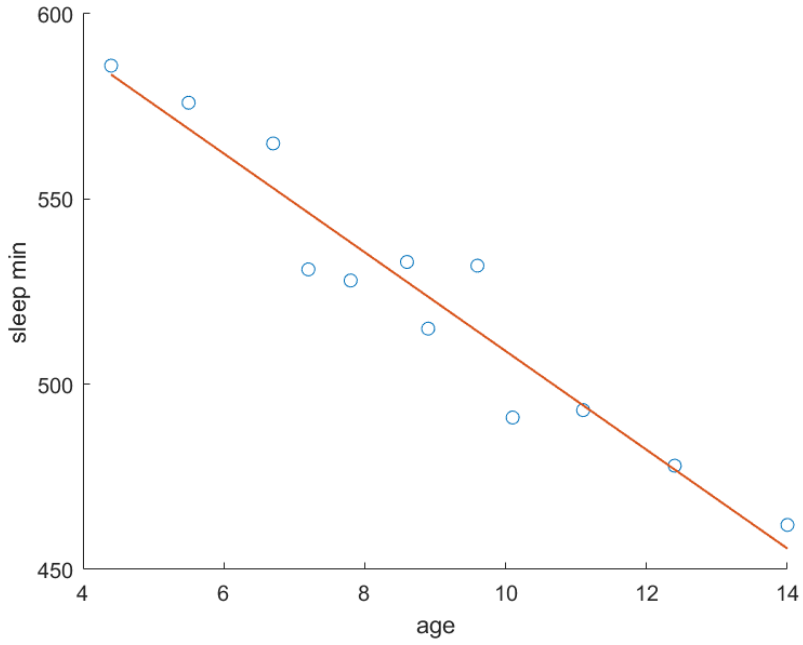
R-sqr	F	p(badF)	S^2_{error}
0.9054	105.3213	0.0000	172.7721

可见 R 方=0.9054，线性回归不显著的 p 值=0，因此可以认为线性回归是显著的。

画出残差分析图如下，所有样本点的残差 95%置信区间均包含 0 点，但样本点-5 的残差置信区间上界=0.1835，逼近 0，下面尝试去除样本点-5 之后再做分析。



去除样本点-5 之后:



$$y = 642.2428 - 13.3294x$$

统计量:

R-sqr	F	p(badF)	S^2_{error}
0.9228	119.4791	0.0000	127.6628

可见 R 方有所提升，残差方差有所减小。另一角度来说，线性回归对于数据的拟合效果变好，但是可能因此损失有效信息，需要做一定权衡。

（2） 给出 10 岁孩子的平均睡眠时间及预测区间。

利用模型-1 预测：

点预测——506.2071

t 分布 95%置信区间预测——[476.0520　536.3623]

正态分布 95%区间预测——[480.4448　531.9695]

利用模型-2 预测：

点预测——508.9490

t 分布 95%置信区间预测——[482.5627　535.3352]

正态分布 95%区间预测——[486.8037　531.0942]

注解：

1. 模型-2 比模型-1 预测区间更小，需要对于“异常”数据特征和预测精度进行权衡；
2. 正态分布是对于 t 分布的近似，适用于 n 较大且 x_0 接近于 x 均值的情况。本例题 n=12/13，不够大，且 x_0 与 x 均值有一定距离，因而正态近似不适用。

2. 社会学家认为犯罪与收入低、失业及人口规模有关，对 20 个城市的犯罪率 y（每 10 万人中犯罪的人数）与年收入低于 5000 美元家庭的百分比 x_1、失业率 x_2 和人口总数 x_3（千人）进行了调查，结果如下表。

序号	y	x_1	x_2	x_3	序号	y	x_1	x_2	x_3
1	11.2	16.5	6.2	587	11	14.5	18.1	6.0	7895
2	13.4	20.5	6.4	643	12	26.9	23.1	7.4	762
3	40.7	26.3	9.3	635	13	15.7	19.1	5.8	2793
4	5.3	16.5	5.3	692	14	36.2	24.7	8.6	741
5	24.8	19.2	7.3	1248	15	18.1	18.6	6.5	625
6	12.7	16.5	5.9	643	16	28.9	24.9	8.3	854
7	20.9	20.2	6.4	1964	17	14.9	17.9	6.7	716
8	35.7	21.3	7.6	1531	18	25.8	22.4	8.6	921
9	8.7	17.2	4.9	713	19	21.7	20.2	8.4	595
10	9.6	14.3	6.4	749	20	25.7	16.9	6.7	3353

（1） 若 x_1, x_2, x_3 中至多只许选择 2 个变量，最好的模型是什么？

根据所给数据，进行双变量线性回归，得到统计量如下表所示：

	R-sqr	F	p(badF)	S^2_{error}
x2&x3	0.7672	28.0054	0.0000	25.4100
x1&x3	0.7103	20.8433	0.0000	31.6120

x1&x2	0.8020	34.4278	0.0000	21.6084

经过比较，从 R 方和残差方差的角度认为让 y 和 x1、x2 进行线性回归效果最显著。

（2） 包含 3 个自变量的模型比上面的模型好吗？确定最终模型。

	R-sqr	F	p(badF)	S^2_{error}
x2&x3	0.7672	28.0054	0.0000	25.4100
x1&x3	0.7103	20.8433	0.0000	31.6120
x1&x2	0.8020	34.4278	0.0000	21.6084
x1&x2&x3	0.8183	24.0220	0.0000	21.0661

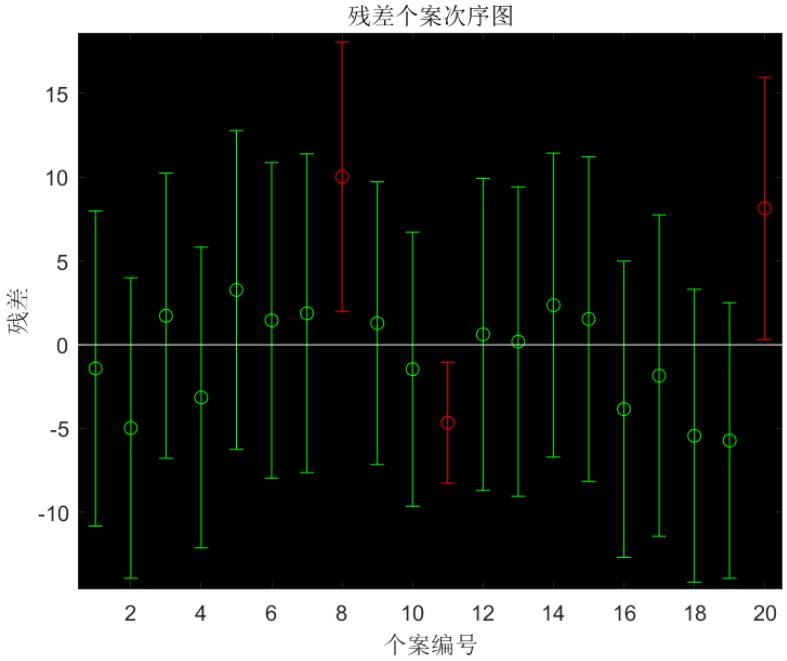
从 R 方和残差方差的角度，3 自变量线性回归结果好于 2 自变量的情况。虽然 F 值比 x1&x2 的情况有所下降，但是 p=0，仍然可以说明回归模型是显著的。

$$f(x1,x2,x3) = -36.7649 + 1.1922x_1 + 4.7198x_2 + 0.0008x_3$$

从线性回归系数也可以看出，相比于 x1 和 x2，因变量与 x3 的相关度较弱（仍然显著相关）。

（3）对最终模型观察残差，有无异常点，若有，剔除后如何。

根据（2）确定的三自变量回归模型，对于拟合残差进行分析：



发现样本点-8、11、20 属于异常点，拟合残差的置信区间不包含 0。因此尝试剔除这三个样本点，再做线性回归。

	R-sqr	F	p(badF)	S^2_{error}
x2&x3	0.7672	28.0054	0.0000	25.4100

x1&x3	0.7103	20.8433	0.0000	31.6120
x1&x2	0.8020	34.4278	0.0000	21.6084
x1&x2&x3	0.8183	24.0220	0.0000	21.0661
3-去除异常	0.9199	49.7697	0.0000	9.5576

可以发现去除异常样本点后 R 方升高、残差方差下降，拟合效果变好。

$$f(x1, x2, x3) = -37.8675 + 1.4575x_1 + 3.8905x_2 + 0.0016x_3$$