帕金森病情数据预测

个人大作业

邱叶红-2050347

目录

前15名受试者的语音记录条数统计

subject_	频数	百分比	累积频数	累积百分比
1	149	2.54	149	2.54
2	145	2.47	294	5.00
3	144	2.45	438	7.46
4	137	2.33	575	9.79
5	156	2.66	731	12.44
6	156	2.66	887	15.10
7	161	2.74	1048	17.84
8	150	2.55	1198	20.39
9	152	2.59	1350	22.98
10	148	2.52	1498	25.50
11	138	2.35	1636	27.85
12	107	1.82	1743	29.67
13	112	1.91	1855	31.57
14	136	2.31	1991	33.89
15	143	2.43	2134	36.32

数据来源

HTTPS://ARCHIVE.ICS.UCI.EDU/DATASET/189/PARKINSONS+TELEMONITORING

该数据集由一系列生物医学语音测量数据组成，这些数据来自 42 名早期帕金森病患者，他们参加了为期 6 个月的远程病情监测试验。录音是在患者家中自动采集的。

数据结构

表中各列包含受试者编号、年龄、性别、与基线招募日期的时间间隔、运动型UPDRS评分、UPDRS 总评分和 16 项生物医学语音测量指标。每一行对应这些人的 5,875 份语音记录中的一份。

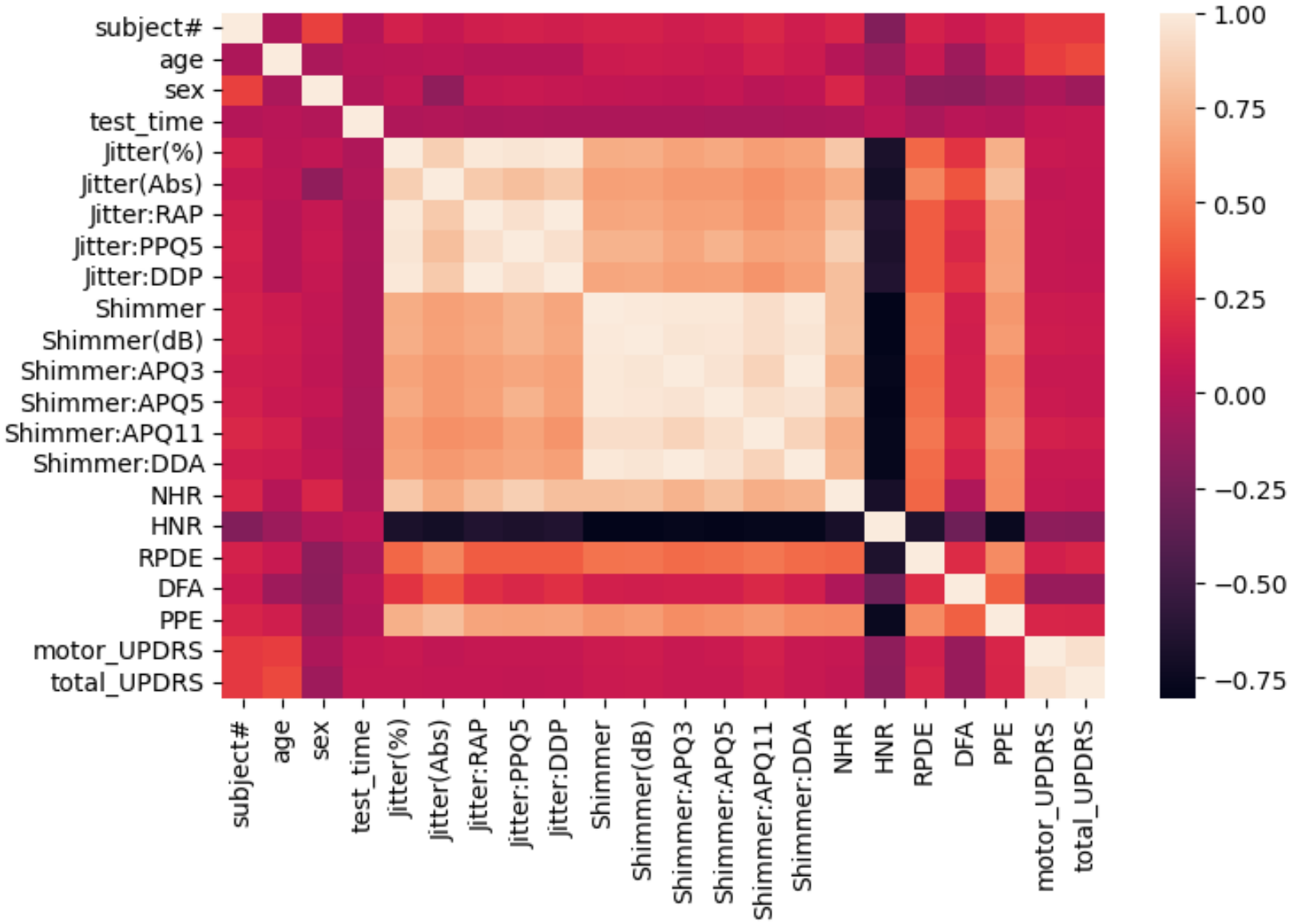
多输出回归任务

数据分析的主要任务是通过 16 项语音测量指标来预测运动型UPDR和UPDRS 总分（"motor_UPDRS"和"total_UPDRS"）。

变量名称

变量(组)名称	变量类型	数据类型	变量说明
subject#	ID	整型	受试者编号
age	自变量	整型	年龄
sex	自变量	0或1	性别
test_time	自变量	连续型	基线招募日期的时间间隔
Jitter组(共5个)	自变量	连续型	语音基频变化的几种测量指标
Shimmer组(共6个)	自变量	连续型	语音振幅变化的几种测量指标
NHR,HNR	自变量	连续型	声音中噪音与音调成分比例的两种测量指标
RPDE	自变量	连续型	非线性动态复杂度指标
DFA	自变量	连续型	信号分形缩放指数
PPE	自变量	连续型	语音基频变化的非线性指标
motor_UPDRS	因变量1	连续型	经过线性插值的临床运动UPDRS评分
total_UPDRS	因变量2	连续型	经过线性插值的临床UPDRS总评分

探索性数据分析



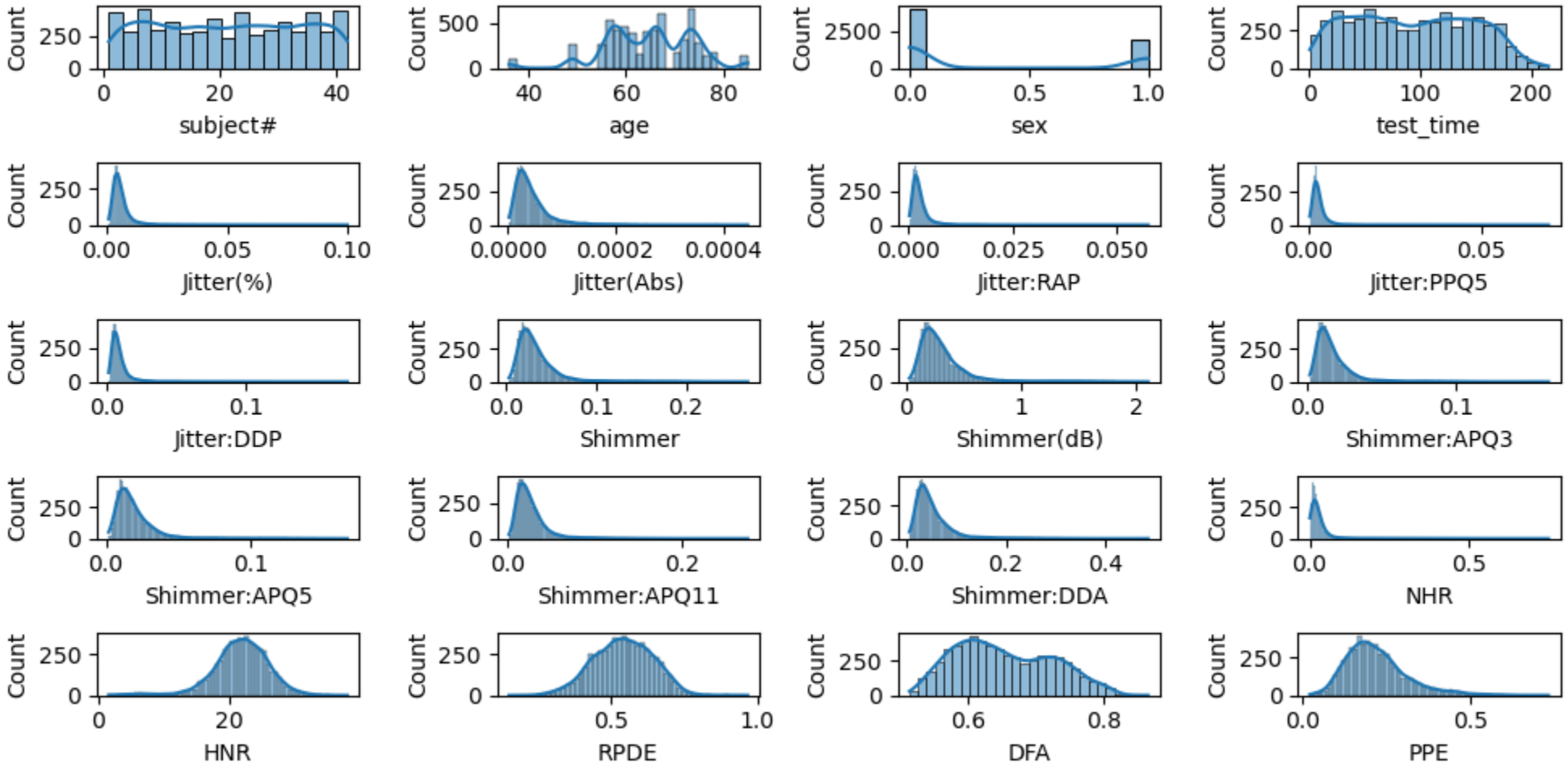
分析

受试者年龄age、性别sex和基准时间test_time和其余连续型自变量以及因变量的相关度不高

Jitter组变量(共5个)强正相关；Shimmer组变量(共6个)强正相关

Jitter组和Shimmer组有正相关关系

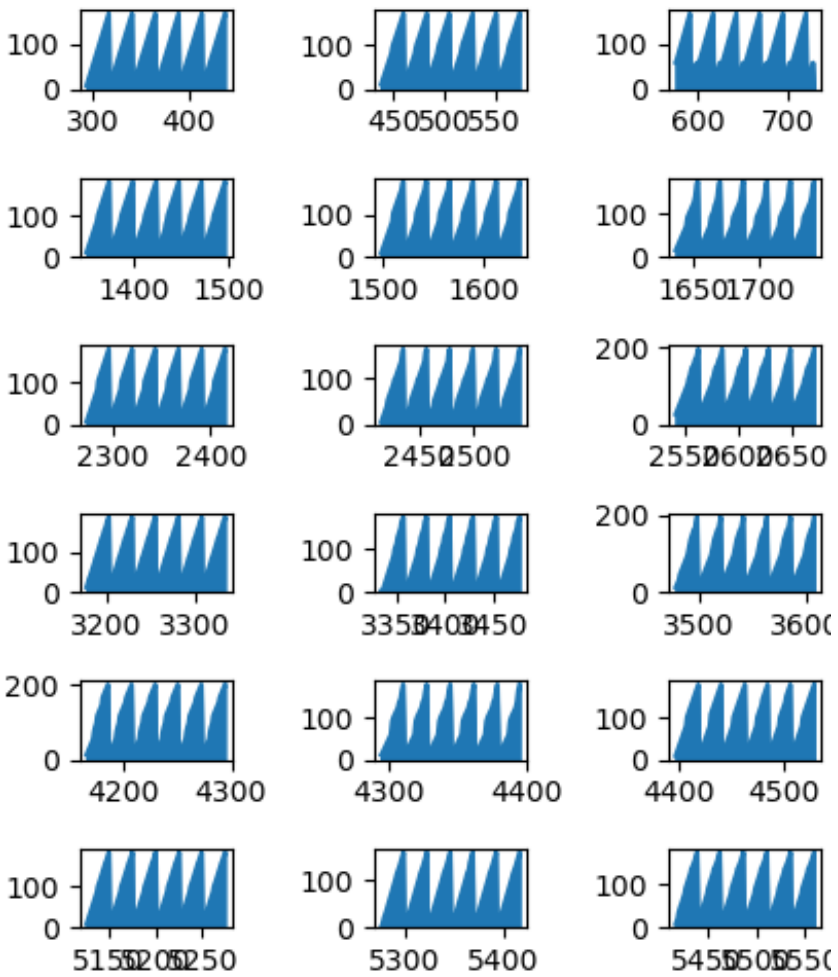
HNR和其余除test_time以外的连续型自变量负相关



分析

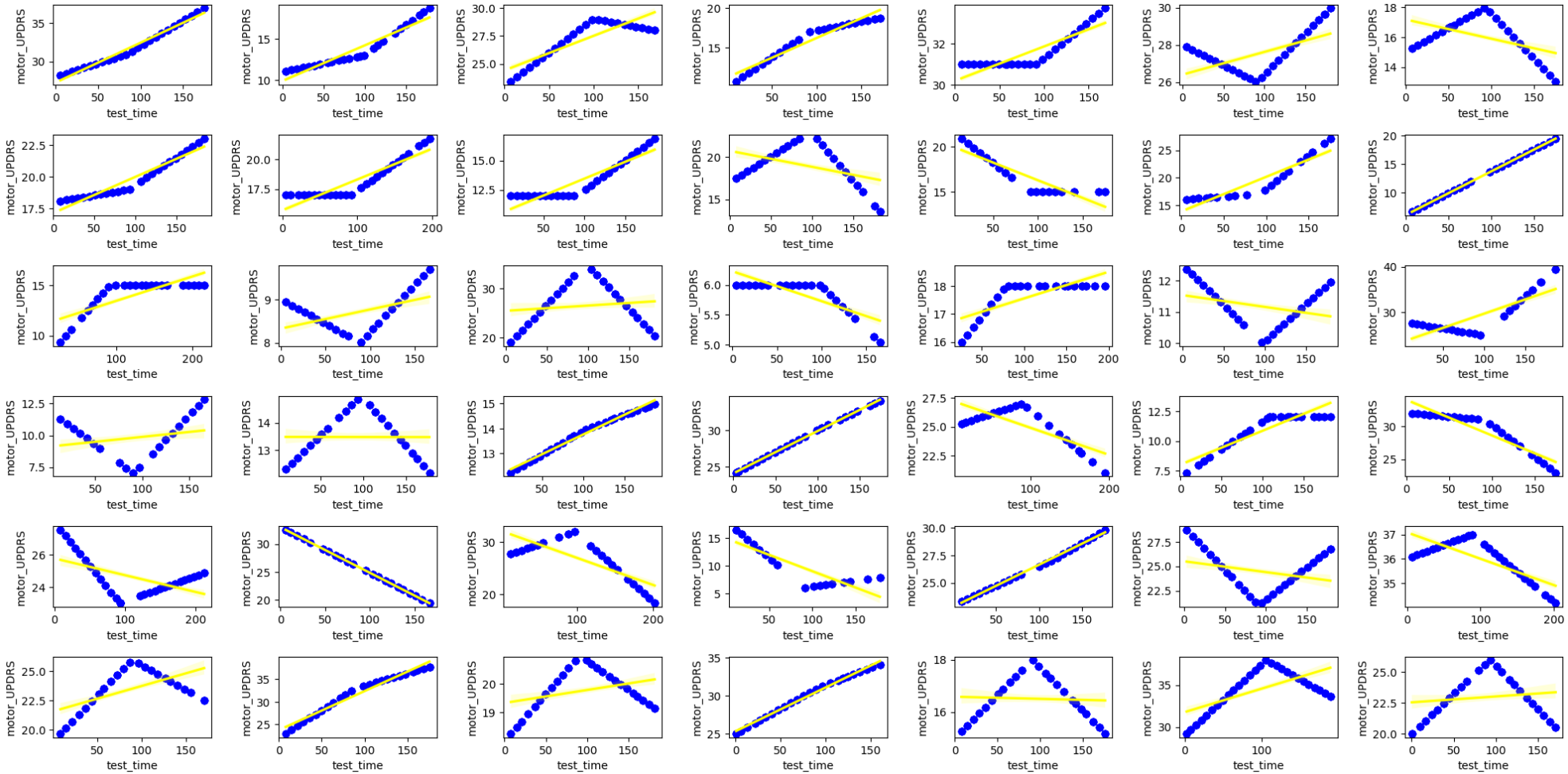
这是去除了test_time<0的记录后的各自变量分布直方图

医疗数据的长尾性



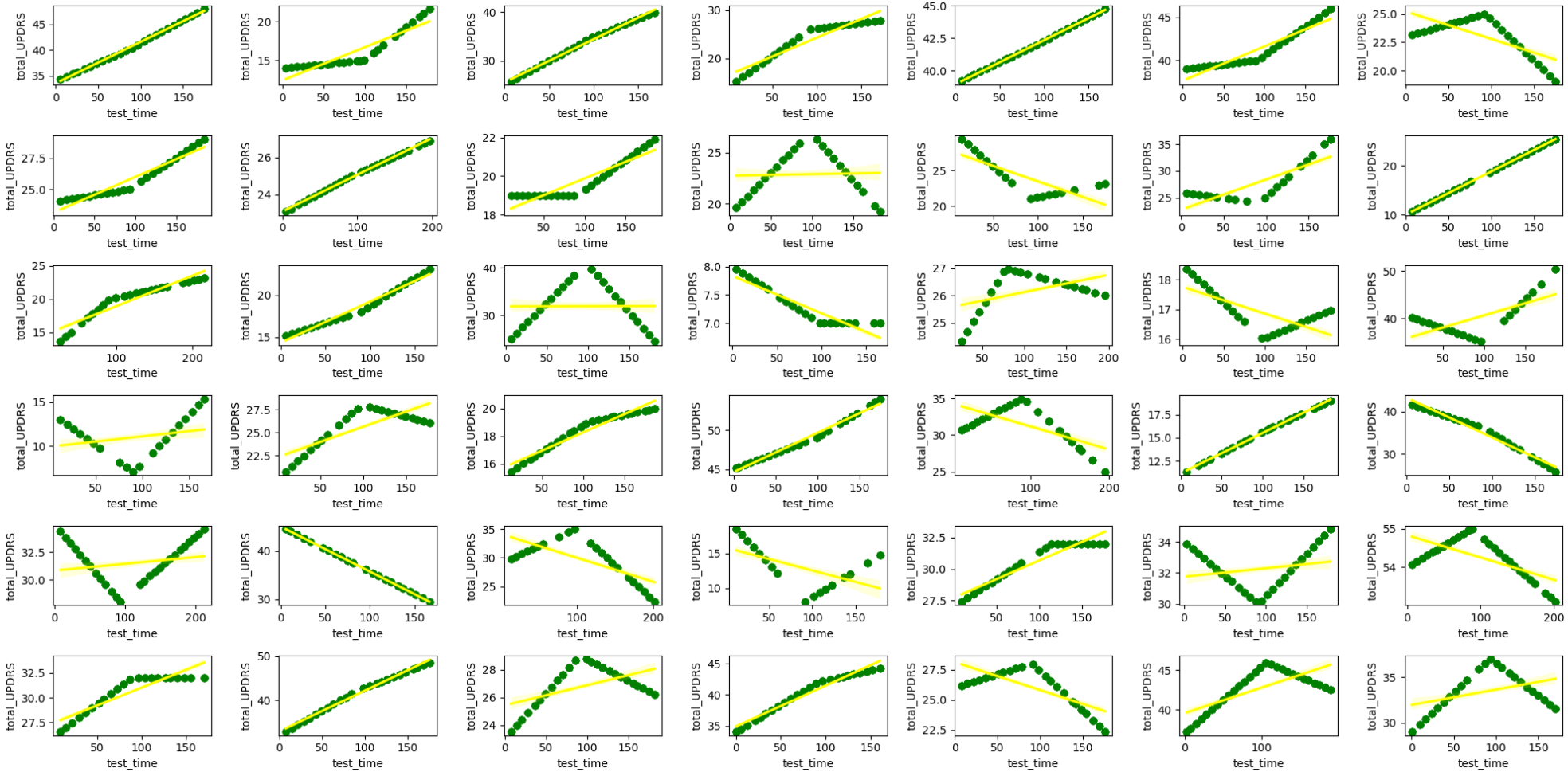
42个受试者中的15人的
test_time在数据集中的排布

这样处理便于机器学习?



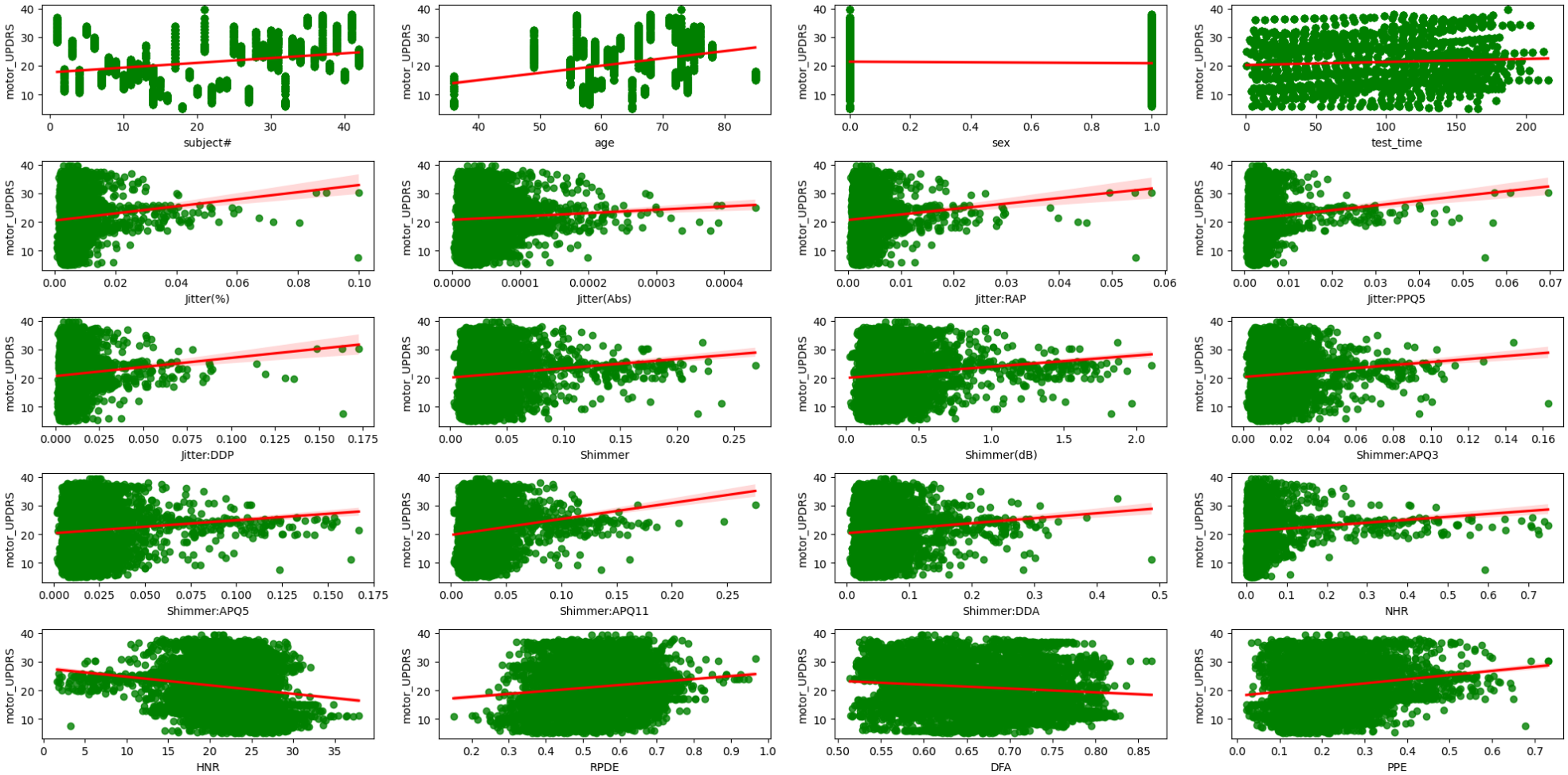
42个受试者的motor_UPDRS评分与test_time的散点分布(点)以及回归直线

数据存在自相关性



42个受试者的total_UPDRS评分与test_time的散点分布(点)以及回归直线

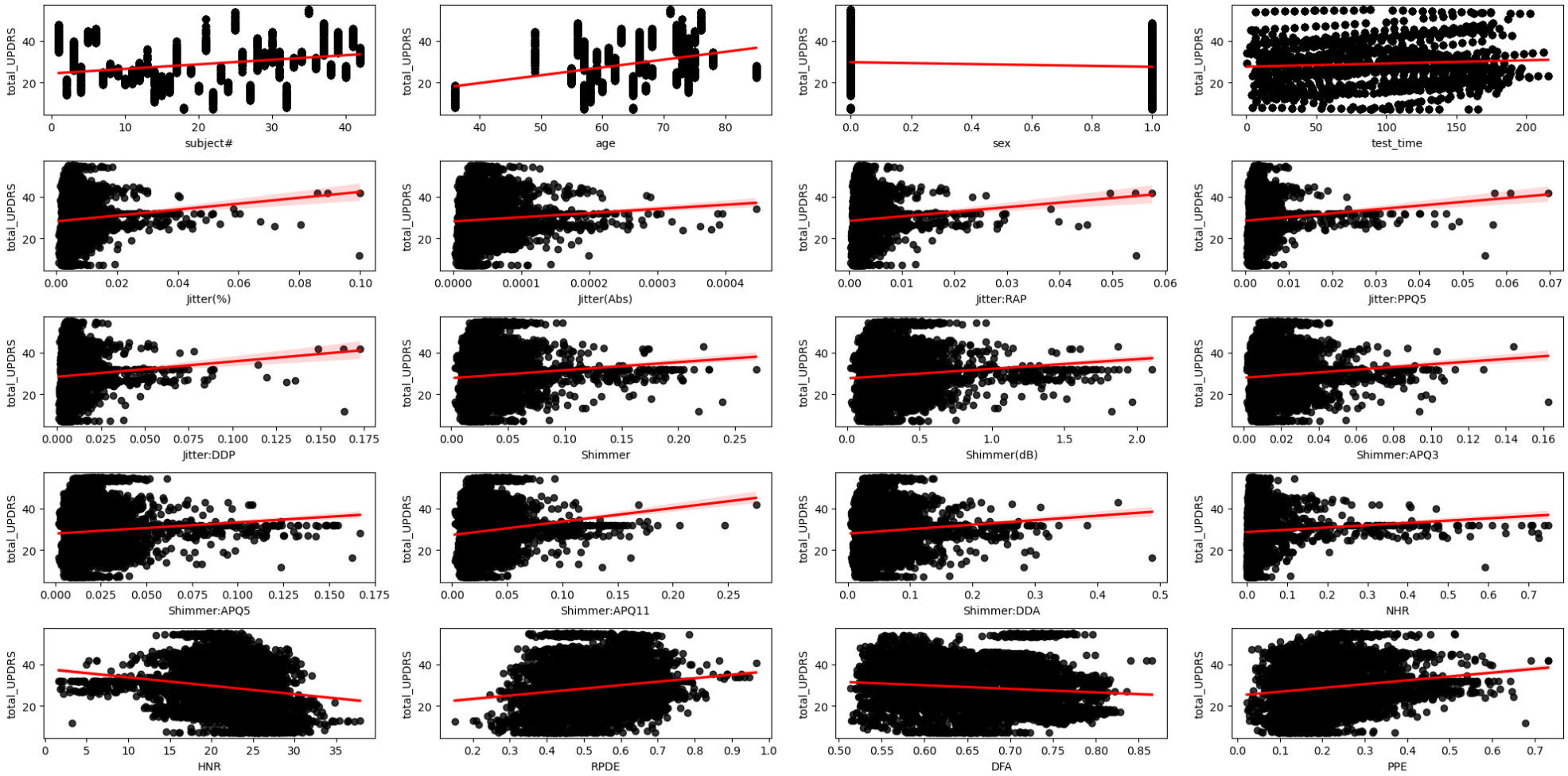
数据存在自相关性



各个自变量与motor_UPDRS的简单回归

数据异方差性

存在离群值



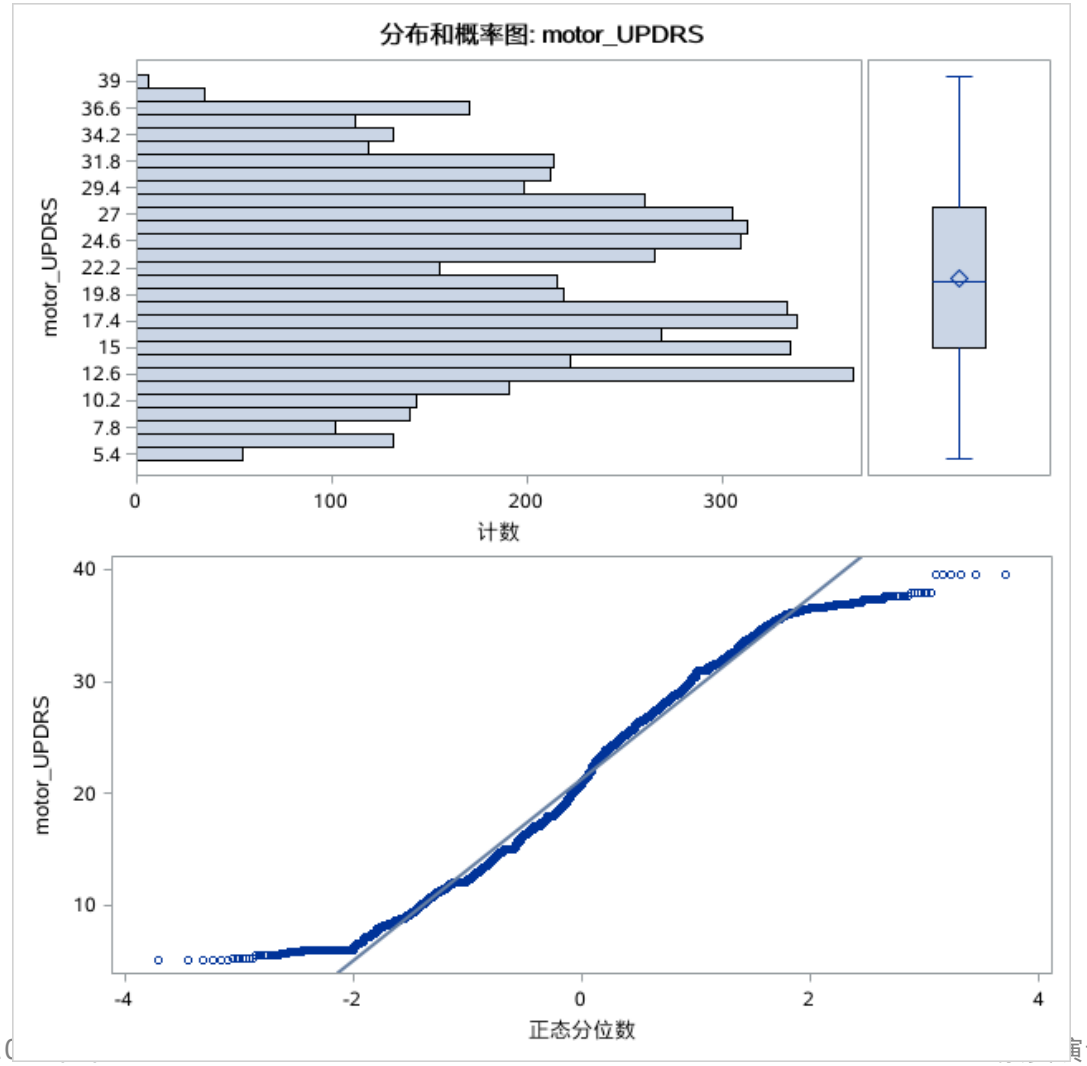
各个自变量与total_UPDRS的简单回归

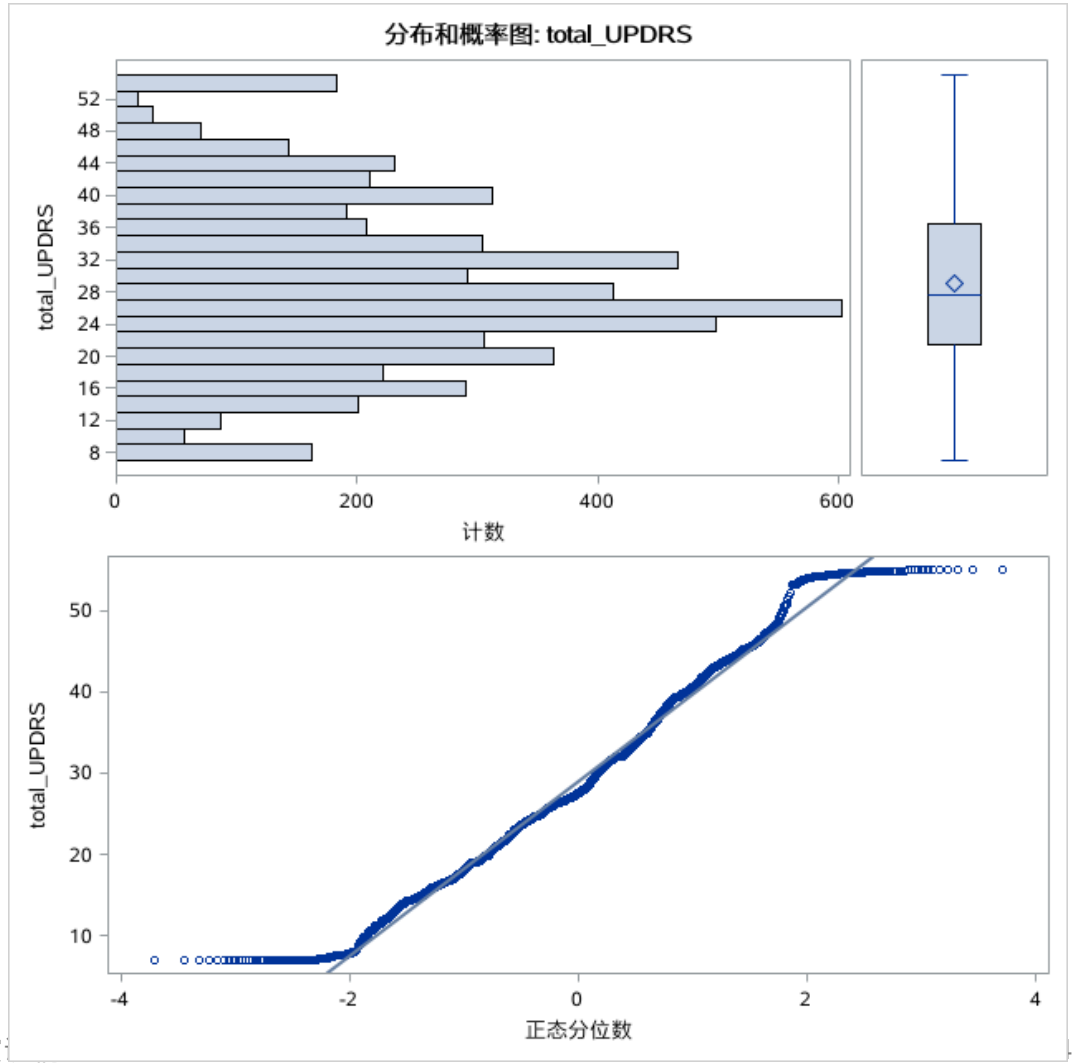
数据异方差性存在离群值

传统统计方法建模

因变量正态性检验

- 经检验，两个因变量motor_UPDRS和total_UPDRS在Kolmogorov-Smirnov检验、Cramer-von Mises检验以及Anderson-Darling检验下都以显著度5%不服从正态性假设。
- 数据具有长尾性短尾性





OLS模型解释力不佳

OLS模型-
MOTOR_UPDRS为因变量

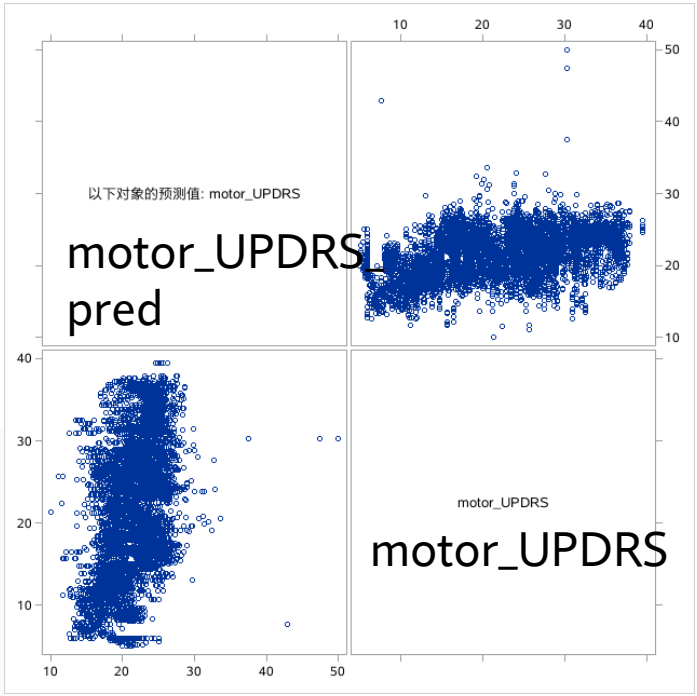
REG 过程
模型: MODEL1
因变量: motor_UPDRS

读取的观测数	5863
使用的观测数	5863

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	19	61159	3218.89009	57.58	<.0001
误差	5843	326627	55.90058		
校正合计	5862	387786			

均方根误差	7.47667	R 方	0.1577
因变量均值	21.28965	调整 R 方	0.1550
变异系数	35.11879		



OLS模型-
TOTAL_UPDRS为因变量

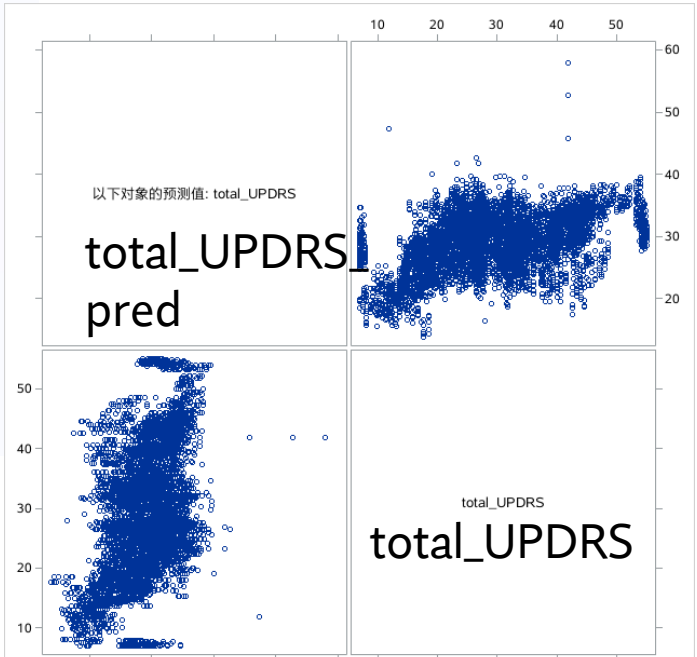
REG 过程
模型: MODEL1
因变量: total_UPDRS

读取的观测数	5863
使用的观测数	5863

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	19	118998	6263.07871	66.13	<.0001
误差	5843	553393	94.71036		
校正合计	5862	672391			

均方根误差	9.73192	R 方	0.1770
因变量均值	29.01409	调整 R 方	0.1743
变异系数	33.54206		



连续型自变量共线性诊断

Motor_UPDRS为自变量的OLS回归模型参数估计						
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	1	30.34823	2.46317	12.32	<.0001	0
age	1	0.19244	0.01158	16.61	<.0001	1.09646
sex	1	-1.13734	0.24344	-4.67	<.0001	1.34902
test_time	1	0.01176	0.00184	6.39	<.0001	1.01054
Jitter(%)	1	294.83601	163.71514	1.80	0.0718	89.02176
Jitter(Abs)	1	-64852	7598.73459	-8.53	<.0001	7.84759
Jitter(RAP)	1	-37270	35980	-1.04	0.3003	1326587
Jitter(DDP)	1	12541	11994	1.05	0.2958	1326817
Jitter(PPQ5)	1	-329.24433	145.61196	-2.26	0.0238	31.00853
Shimmer	1	149.19500	49.96792	2.99	0.0028	174.76617
Shimmer(dB)	1	-6.83475	3.72046	-1.84	0.0663	76.99242
Shimmer(APQ3)	1	-4007.94608	36133	-0.11	0.9117	23988032
Shimmer(APQ5)	1	-123.34451	42.49779	-2.90	0.0037	52.64440
Shimmer(APQ11)	1	71.83227	19.11867	3.76	0.0002	15.33525
Shimmer(DDA)	1	1280.24191	12044	0.11	0.9154	23987906
NHR	1	-10.56000	4.79059	-2.20	0.0275	8.58750
HNR	1	-0.42441	0.05294	-8.02	<.0001	5.41629
RPDE	1	0.55992	1.40100	0.40	0.6894	2.10066
DFA	1	-23.60018	1.77409	-13.30	<.0001	1.66169
PPE	1	17.65880	2.24881	7.85	<.0001	4.43940

建立motor_UPDRS为因变量的单因变量线性回归模型，方差膨胀系数>5的连续型自变量为Jitter组所有变量、Shimmer组所有变量、NHR以及HNR

数据存在明显共线性

Motor_UPDRS为自变量的OLS回归模型参数估计						
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	1	21.65183	0.12468	173.67	<.0001	0
age	1	1.69877	0.10225	16.61	<.0001	1.09646
sex	1	-1.13734	0.24344	-4.67	<.0001	1.34902
test_time	1	0.62731	0.09817	6.39	<.0001	1.01054
Jitter(%)	1	1.65930	0.92137	1.80	0.0718	89.02176
Jitter(Abs)	1	-2.33473	0.27356	-8.53	<.0001	7.84759
Jitter(RAP)	1	-116.50571	112.47426	-1.04	0.3003	1326587
Jitter(DDP)	1	117.61281	112.48401	1.05	0.2958	1326817
Jitter(PPQ5)	1	-1.22955	0.54378	-2.26	0.0238	31.00853
Shimmer	1	3.85458	1.29096	2.99	0.0028	174.76617
Shimmer(dB)	1	-1.57411	0.85686	-1.84	0.0663	76.99242
Shimmer(APQ3)	1	-53.05225	478.28045	-0.11	0.9117	23988039
Shimmer(APQ5)	1	-2.05643	0.70854	-2.90	0.0037	52.64440
Shimmer(APQ11)	1	1.43679	0.38241	3.76	0.0002	15.33525
Shimmer(DDA)	1	50.83883	478.27920	0.11	0.9154	23987914
NHR	1	-0.63080	0.28617	-2.20	0.0275	8.58750
HNR	1	-1.82181	0.22727	-8.02	<.0001	5.41629
RPDE	1	0.05657	0.14153	0.40	0.6894	2.10066
DFA	1	-1.67456	0.12588	-13.30	<.0001	1.66169
PPE	1	1.61568	0.20575	7.85	<.0001	4.43940

对各个连续型自变量做标准化处理后（模型解释能力不会发生改变，但自变量系数会发生改变），建立motor_UPDRS为因变量的单因变量线性回归模型，方差膨胀系数>5的连续型自变量为Jitter组所有变量、Shimmer组所有变量、NHR以及HNR

数据存在明显共线性

连续型自变量共线性诊断

Total_UPDRS为自变量的OLS回归模型参数估计						
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	1	39.64493	3.20615	12.37	<.0001	0
age	1	0.30349	0.01508	20.13	<.0001	1.09646
sex	1	-2.77469	0.31687	-8.76	<.0001	1.34902
test_time	1	0.01685	0.00240	7.03	<.0001	1.01054
Jitter(%)	1	47.94391	213.09803	0.22	0.8220	89.02176
Jitter(Abs)	1	-63980	9890.81005	-6.47	<.0001	7.84759
Jitter(RAP)	1	-39583	46833	-0.85	0.3980	1326587
Jitter(DDP)	1	13474	15612	0.86	0.3882	1326817
Jitter(PPQ5)	1	-342.30022	189.53422	-1.81	0.0710	31.00853
Shimmer	1	151.59631	65.04019	2.33	0.0198	174.76617
Shimmer(dB)	1	-8.98377	4.84269	-1.86	0.0636	76.99242
Shimmer(APQ3)	1	-16713	47032	-0.36	0.7223	23988032
Shimmer(APQ5)	1	-62.15333	55.31679	-1.12	0.2612	52.64440
Shimmer(APQ11)	1	52.60439	24.88561	2.11	0.0346	15.33525
Shimmer(DDA)	1	5497.57778	15677	0.35	0.7258	23987906
NHR	1	-15.40178	6.23561	-2.47	0.0135	8.58750
HNR	1	-0.62630	0.06892	-9.09	<.0001	5.41629
RPDE	1	4.00282	1.82359	2.20	0.0282	2.10066
DFA	1	-31.52466	2.30923	-13.65	<.0001	1.66169
PPE	1	17.63989	2.92714	6.03	<.0001	4.43940

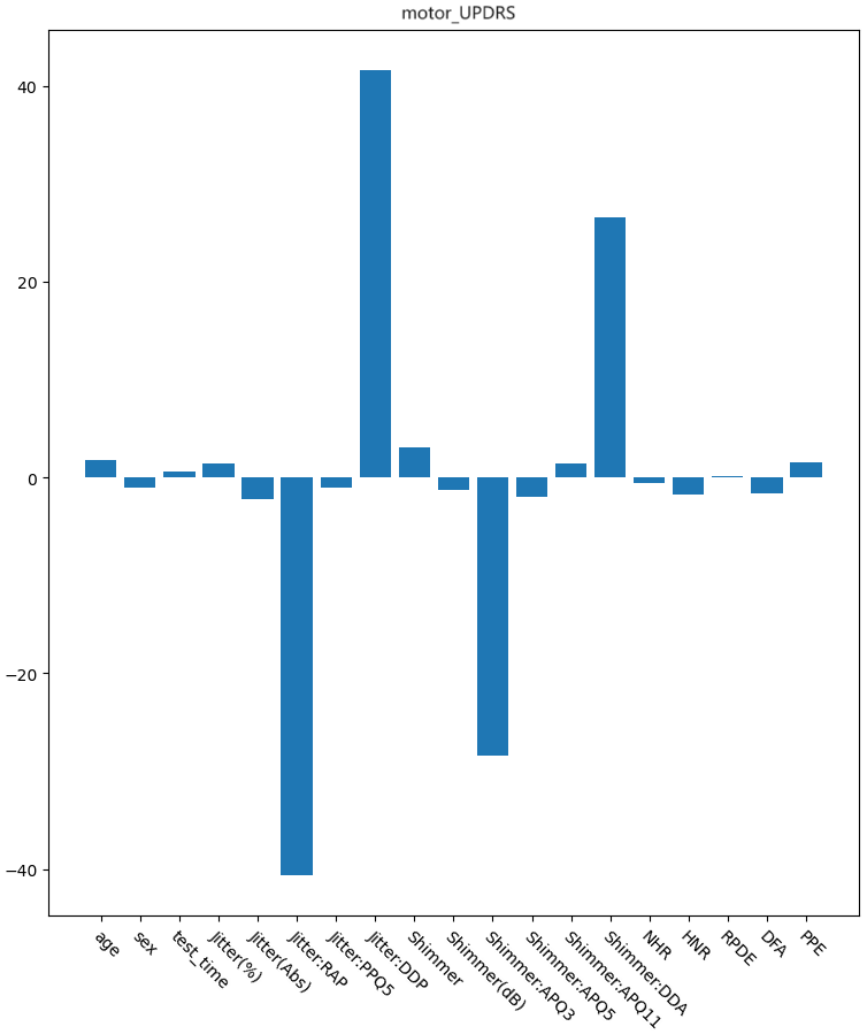
建立total_UPDRS为因变量的单因变量线性回归模型，方差膨胀系数>5的连续型自变量为Jitter组所有变量、Shimmer组所有变量、NHR以及HNR

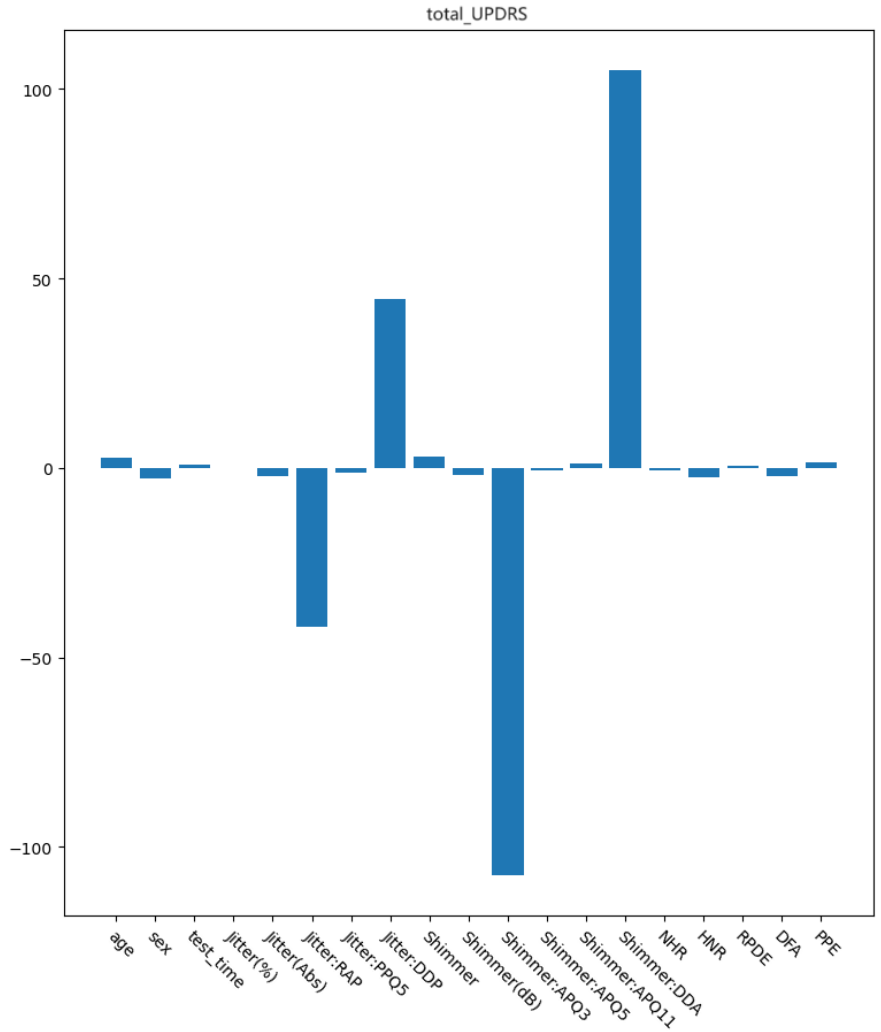
数据存在明显共线性

Total_UPDRS为自变量的OLS回归模型参数估计						
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	1	29.89765	0.16228	184.23	<.0001	0
age	1	2.67913	0.13310	20.13	<.0001	1.09646
sex	1	-2.77469	0.31687	-8.76	<.0001	1.34902
test_time	1	0.89872	0.12778	7.03	<.0001	1.01054
Jitter(%)	1	0.26982	1.19929	0.22	0.8220	89.02176
Jitter(Abs)	1	-2.30333	0.35608	-6.47	<.0001	7.84759
Jitter(RAP)	1	-123.73705	146.40089	-0.85	0.3980	1326587
Jitter(DDP)	1	126.35686	146.41359	0.86	0.3882	1326817
Jitter(PPQ5)	1	-1.27831	0.70781	-1.81	0.0710	31.00853
Shimmer	1	3.91662	1.68037	2.33	0.0198	174.76617
Shimmer(dB)	1	-2.06905	1.11532	-1.86	0.0636	76.99242
Shimmer(APQ3)	1	-221.22792	622.54854	-0.36	0.7223	23988039
Shimmer(APQ5)	1	-1.03624	0.92226	-1.12	0.2612	52.64440
Shimmer(APQ11)	1	1.05219	0.49776	2.11	0.0346	15.33525
Shimmer(DDA)	1	218.31065	622.54691	0.35	0.7258	23987914
NHR	1	-0.92003	0.37249	-2.47	0.0135	8.58750
HNR	1	-2.68839	0.29582	-9.09	<.0001	5.41629
RPDE	1	0.40438	0.18423	2.20	0.0282	2.10066
DFA	1	-2.23684	0.16385	-13.65	<.0001	1.66169
PPE	1	1.61395	0.26782	6.03	<.0001	4.43940

对各个连续型自变量做标准化处理后（模型解释能力不会发生改变，但自变量系数会发生改变），建立total_UPDRS为因变量的单因变量线性回归模型，方差膨胀系数>5的连续型自变量为Jitter组所有变量、Shimmer组所有变量、NHR以及HNR

数据存在明显共线性





各个连续型自变量标准化后的OLS模型中各个自变量的系数大小

自相关性诊断

存在显著正自相关

REG 过程
模型: MODEL1
因变量: motor_UPDRS

Durbin-Watson D	0.141
Pr < DW	<.0001
Pr > DW	1.0000
观测数	5863
第一阶自相关	0.929

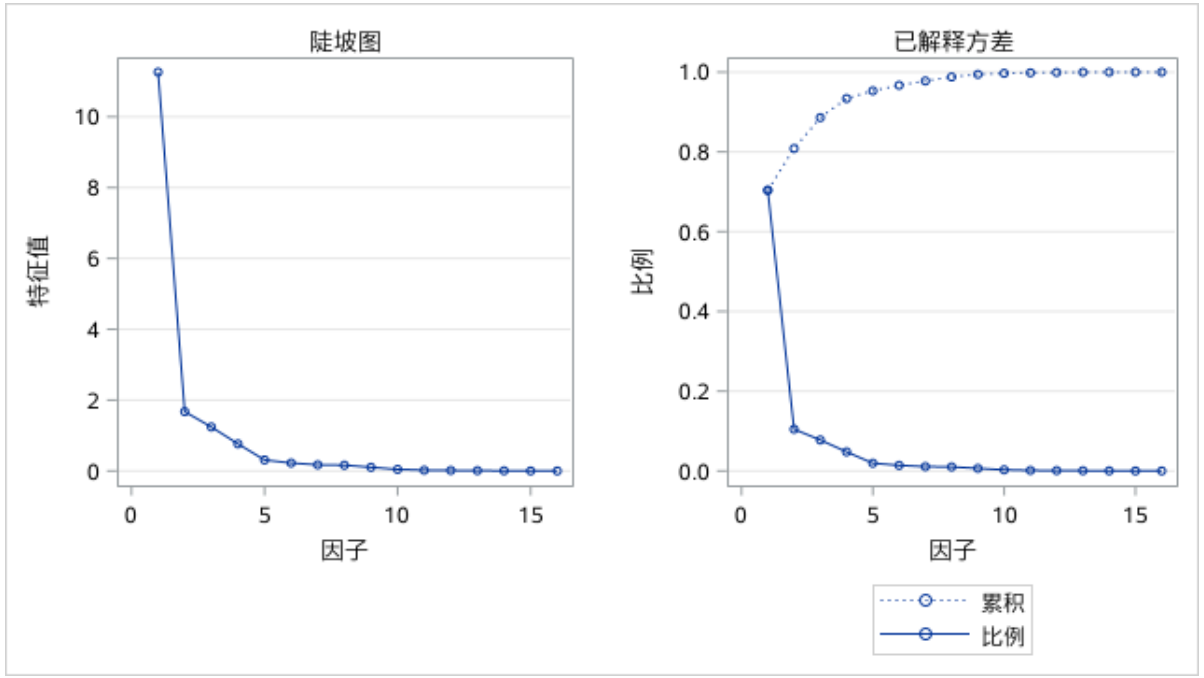
存在显著正自相关

REG 过程
模型: MODEL1
因变量: total_UPDRS

Durbin-Watson D	0.118
Pr < DW	<.0001
Pr > DW	1.0000
观测数	5863
第一阶自相关	0.941

连续型自变量因子分析

相关矩阵特征值大小



选择公共因子数为7.对初始因子按最大方差旋转法进行旋转后得旋转因子。

最终的公因子方差估计: 总计 = 15.644363

Jitter(%)	Jitter(Abs)	Jitter(RAP)	Jitter(DDP)	Jitter(PPQ5)	Shimmer	Shimmer(dB)	Shimmer(APQ3)
0.990	0.919	0.984	0.984	0.972	0.994	0.985	0.992
Shimmer(APQ5)	Shimmer(APQ11)	Shimmer(DDA)	NHR	HNR	RPDE	DFA	PPE
0.984	0.957	0.992	0.945	0.964	0.998	0.997	0.986

每个因子已解释方差

Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
6.3325680	5.3323259	1.4017181	1.1590722	0.8988385	0.2929683	0.2268718

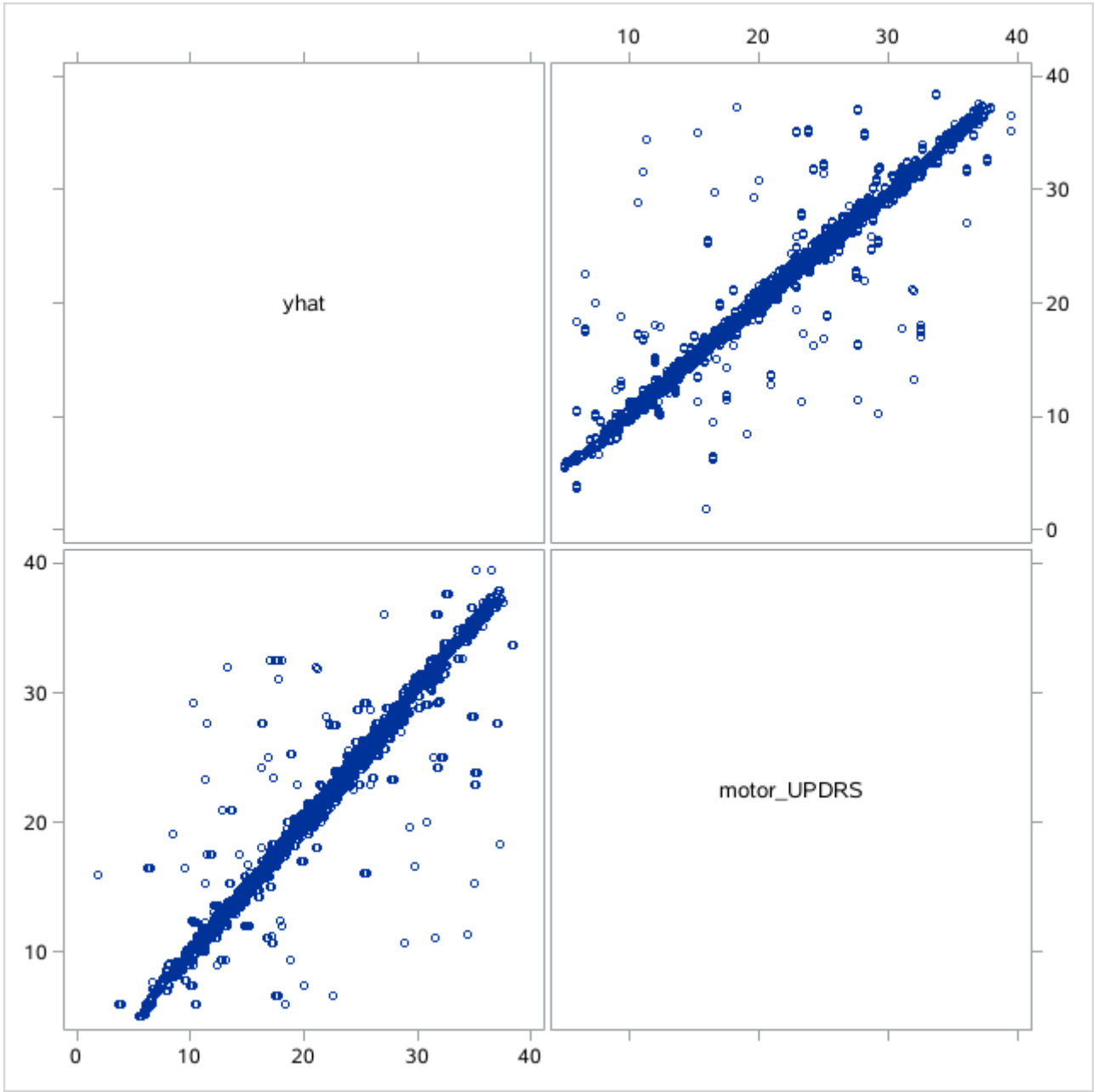
Pearson 相关系数, N = 5863 Prob > |r|, H0: Rho=0							
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Factor1	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	-0.00000 1.0000	0.00000 1.0000
Factor2	0.00000 1.0000	1.00000	-0.00000 1.0000	-0.00000 1.0000	-0.00000 1.0000	0.00000 1.0000	-0.00000 1.0000
Factor3	0.00000 1.0000	-0.00000 1.0000	1.00000	-0.00000 1.0000	-0.00000 1.0000	0.00000 1.0000	-0.00000 1.0000
Factor4	0.00000 1.0000	-0.00000 1.0000	-0.00000 1.0000	1.00000	-0.00000 1.0000	0.00000 1.0000	-0.00000 1.0000
Factor5	0.00000 1.0000	-0.00000 1.0000	-0.00000 1.0000	-0.00000 1.0000	1.00000	0.00000 1.0000	-0.00000 1.0000
Factor6	-0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000
Factor7	0.00000 1.0000	-0.00000 1.0000	-0.00000 1.0000	-0.00000 1.0000	-0.00000 1.0000	0.00000 1.0000	1.00000

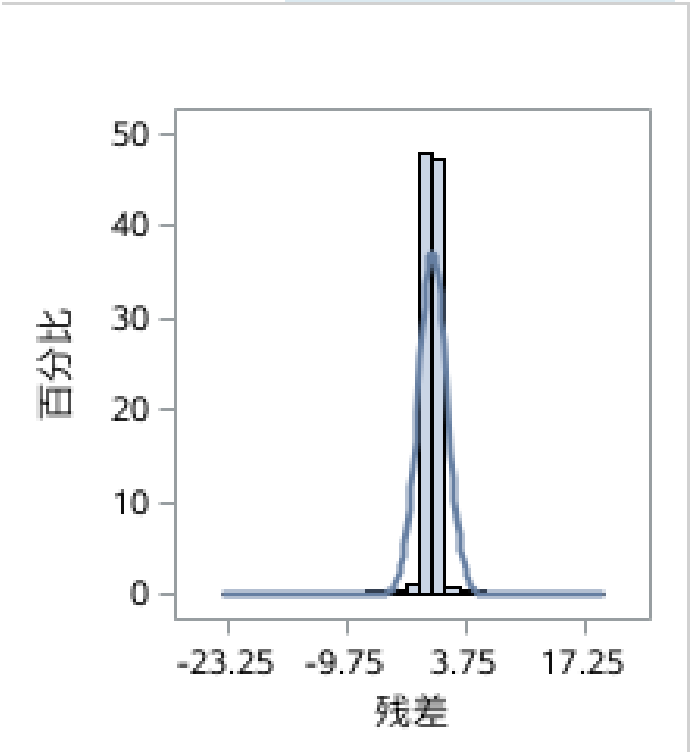
旋转因子模式							
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Shimmer(DDA)	0.91400	0.33330	0.12522	0.03780	0.07888	-0.06155	-0.13444
Shimmer(APQ3)	0.91399	0.33330	0.12522	0.03780	0.07888	-0.06155	-0.13443
Shimmer	0.90166	0.37673	0.15205	0.03198	0.11068	-0.05392	0.01219
Shimmer(APQ5)	0.90140	0.36241	0.13843	0.03339	0.09508	-0.06711	0.07910
Shimmer(dB)	0.88908	0.38264	0.15550	0.02135	0.14446	-0.04651	0.02736
Shimmer(APQ11)	0.86457	0.28685	0.18870	0.09112	0.20248	0.04783	0.20085
HNR	-0.61292	-0.34674	-0.39178	-0.17338	-0.31469	0.42929	0.02904
Jitter(RAP)	0.34255	0.91343	0.10369	0.08465	0.10956	-0.01499	-0.04958
Jitter(DDP)	0.34255	0.91343	0.10368	0.08466	0.10957	-0.01501	-0.04957
Jitter(%)	0.36515	0.89475	0.14007	0.08819	0.16723	-0.02288	0.02554
Jitter(PPQ5)	0.41552	0.86571	0.09074	0.04858	0.10447	-0.06499	0.15565
Jitter(Abs)	0.31520	0.73223	0.29692	0.20168	0.30912	-0.07122	-0.23243
NHR	0.54763	0.68261	0.14906	-0.15065	0.03720	-0.27327	0.24164
RPDE	0.26032	0.19095	0.93151	0.07699	0.13617	-0.03380	0.00645
DFA	0.03765	0.11222	0.07488	0.98327	0.10197	-0.01429	-0.00670
PPE	0.33697	0.46694	0.28590	0.23382	0.71794	-0.05430	0.00982

对Motor_UPDRS做自相关回归：

最大似然估计			
SSE	15243.9431	DFE	5850
MSE	2.60580	均方根误差	1.61425
SBC	22356.6364	AIC	22269.8429
MAE	0.63691811	AICC	22269.9052
MAPE	3.58149817	HQC	22300.0188
对数似然	-11121.921	转换回归 R 方	0.0807
Durbin-Watson	1.9976	总 R 方	0.9607
		观测	5863

参数估计					
变量	自由度	估计	标准误差	t 值	近似 Pr > |t|
Intercept	1	8.9012	1.5089	5.90	<.0001
age	1	0.1762	0.0178	9.89	<.0001
sex	1	-0.1626	0.4034	-0.40	0.6870
test_time	1	0.0114	0.000581	19.55	<.0001
Factor1	1	0.004258	0.0244	0.17	0.8616
Factor2	1	-0.0399	0.0184	-2.17	0.0303
Factor3	1	-0.0543	0.0227	-2.39	0.0167
Factor4	1	-0.1918	0.0459	-4.18	<.0001
Factor5	1	0.002556	0.0206	0.12	0.9013
Factor6	1	-0.0175	0.0208	-0.84	0.4015
Factor7	1	0.0183	0.0194	0.94	0.3459
AR1	1	-0.9941	0.0131	-75.93	<.0001
AR2	1	0.0161	0.0131	1.23	0.2178

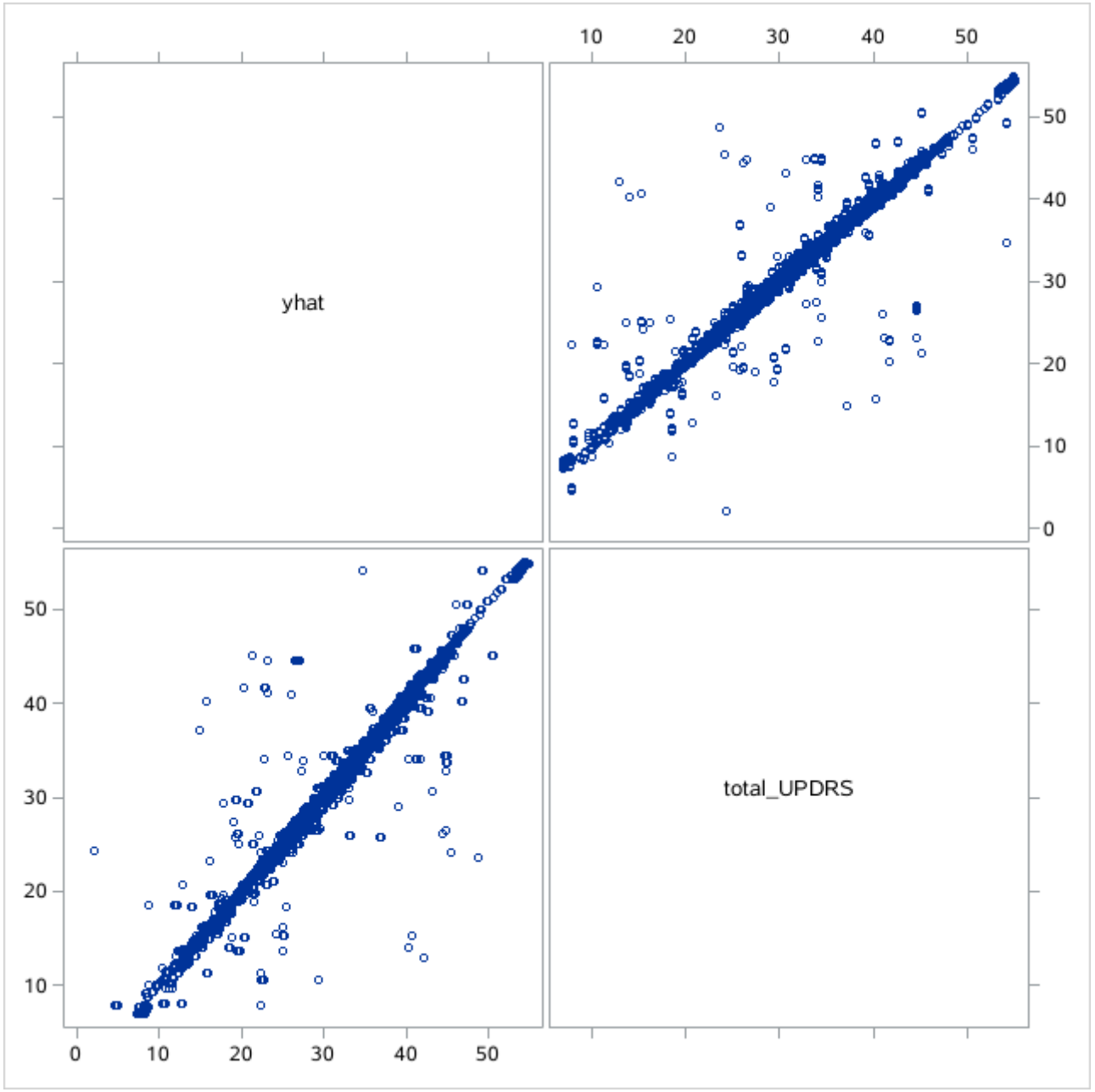

yhat
motor_UPDRS
10
20
30
40
0

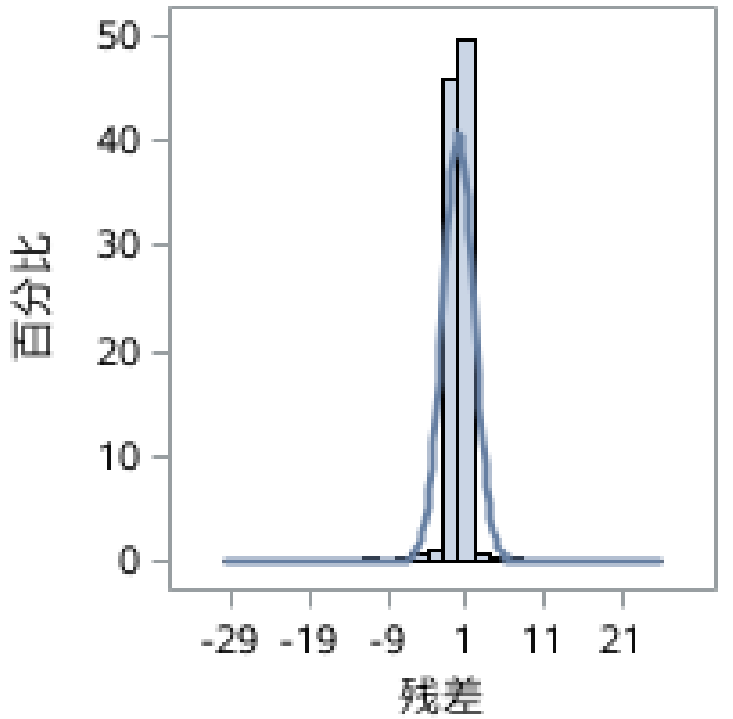

50
40
30
20
10
0
百分比
-23.25
-9.75
3.75
17.25
残差

对Total_UPDRS做自相关回归：

最大似然估计			
SSE	22616.6972	DFE	5850
MSE	3.86610	均方根误差	1.96624
SBC	24669.7542	AIC	24582.9608
MAE	0.75809535	AICC	24583.023
MAPE	3.15783934	HQC	24613.1366
对数似然	-12278.48	转换回归 R 方	0.1245
Durbin-Watson	1.9991	总 R 方	0.9664
		观测	5863

参数估计					
变量	自由度	估计	标准误差	t 值	近似 Pr > |t|
Intercept	1	11.6805	1.9449	6.01	<.0001
age	1	0.2480	0.0217	11.41	<.0001
sex	1	-1.1734	0.4928	-2.38	0.0173
test_time	1	0.0180	0.000707	25.52	<.0001
Factor1	1	-0.003056	0.0297	-0.10	0.9181
Factor2	1	-0.0650	0.0224	-2.90	0.0038
Factor3	1	-0.0697	0.0276	-2.53	0.0116
Factor4	1	-0.2559	0.0559	-4.58	<.0001
Factor5	1	0.001543	0.0251	0.06	0.9510
Factor6	1	-0.0130	0.0253	-0.51	0.6066
Factor7	1	0.0218	0.0237	0.92	0.3560
AR1	1	-0.9955	0.0131	-76.01	<.0001
AR2	1	0.0148	0.0131	1.13	0.2576


yhat
total_UPDRS
10
20
30
40
50
0


百分比
50
40
30
20
10
0
-29
-19
-9
1
11
21
残差

机器学习建模

随机森林将多个决策树结合起来，为回归任务提供准确的预测

选择数据集的80%数据用于训练，20%数据用于测试。后同

```
from sklearn.ensemble import RandomForestRegressor
rdf = RandomForestRegressor()
rdf.fit(X_train,y_train)
showResults(y_test,rdf.predict(X_test))
```

```
R2 score:  0.9755004821148352
Mean squared error:  2.2051865736924983
Mean absolute error:  0.6268520034100593
Predictions:
```

motor_UPDRS	total_UPDRS	motor_UPDRS_pred	total_UPDRS_pred
18	26.968	17.9812	26.9323
11.088	13.088	10.9872	13.0013
22.178	25.904	21.687	25.8656
34.012	42.81	33.9046	42.4597
17.334	22.953	16.797	22.3716

随机森林对输入数据进行子样本替换，而 Extra Trees 则使用整个原始样本，并在建树过程中通过选择随机分割而不是最优分割来增加随机性。

```
from sklearn.ensemble import ExtraTreesRegressor
extra_reg = ExtraTreesRegressor()
extra_reg.fit(X_train,y_train)
showResults(y_test,extra_reg.predict(X_test))
```

```
R2 score:  0.9803894778750649
Mean squared error:  1.763386499573825
Mean absolute error:  0.6896817127024716
Predictions:
```

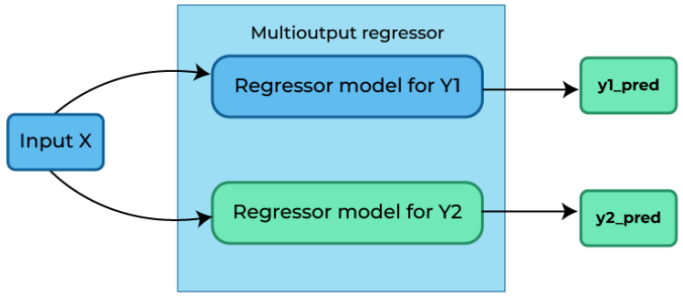
motor_UPDRS	total_UPDRS	motor_UPDRS_pred	total_UPDRS_pred
18	26.968	17.6719	26.5115
11.088	13.088	11.1589	13.1594
22.178	25.904	21.8516	26.0088
34.012	42.81	34.1386	42.8932
17.334	22.953	17.3433	23.1515

K近邻回归器根据邻近特征估算目标值，使其适用于多输出回归，尤其是在特征存在局部模式的情况下，该回归器效果显著

```
from sklearn.neighbors import KNeighborsRegressor
knn = KNeighborsRegressor()
knn.fit(X_train,y_train)
showResults(y_test,knn.predict(X_test))
```

```
R2 score:  0.5245197752322595
Mean squared error:  42.2538651662943
Mean absolute error:  4.767536947996586
Predictions:
```

motor_UPDRS	total_UPDRS	motor_UPDRS_pred	total_UPDRS_pred
18	26.968	14.4348	19.696
11.088	13.088	11.0874	13.0874
22.178	25.904	25.389	30.511
34.012	42.81	31.431	39.5766
17.334	22.953	15.2628	21.7472



SVR 最初是为单输出回归任务设计的，而 MultiOutputRegressor 就像是 SVR 的包装器，它扩展了 SVR 以有效处理多输出回归。

```
from sklearn.multioutput import MultiOutputRegressor
from sklearn.svm import SVR

svm_multi = MultiOutputRegressor(SVR(kernel="rbf", C=100, gamma=0.1, epsilon=0.1))
svm_multi.fit(X_train,y_train)
showResults(y_test,svm_multi.predict(X_test))
```

```
R2 score:  0.7988785418738004
Mean squared error:  17.79332672889405
Mean absolute error:  2.069184533999553
Predictions:
```

motor_UPDRS	total_UPDRS	motor_UPDRS_pred	total_UPDRS_pred
18	26.968	17.8662	26.8435
11.088	13.088	11.2619	13.454
22.178	25.904	23.45	27.6616
34.012	42.81	34.1	42.5883
17.334	22.953	12.7476	19.4979

模型名称	Random Forest	Extra Tree	K-nearest Neighbors	SVR
MSE	2.205	1.763	42.254	17.793
R^2	0.976	0.980	0.525	0.799

该数据集采用树模型+集成学习，能够有较好的预测能力

模型评价与总结

总结和展望

传统统计方法建模

本次大作业通过运用课程所学OLS模型、验证OLS基本假设、因子分析等内容，以实际数据分析巩固了对知识点的理解

传统统计方法V.S.机器学习

传统统计方法比机器学习方法具有更高的解释性，并且如果抓住了主要矛盾，传统统计模型的解释力可以很强

改进方向1

使用传统统计方法建模时，需要科学地对异常值进行检测和剔除

改进方向2

深入了解机器学习模型，学习超参数调参中的贝叶斯优化方法

谢谢