

河北工业大学经济管理学院

《商务大数据分析》

课程设计报告

工业蒸汽量预测建模算法

班 级： 信管 151 班

小组成员： 151961 牟江秀

151955 韩姣

151958 季婕

151954 郭钰

指导教师： 李杰

2018 年 12 月

目录

一、赛题背景.....	3
二、数据介绍.....	3
三、实验过程.....	3
3.1 描述统计	3
3.2 分析过程	10
3.2.1 实验一——回归分析	10
3.2.2 实验二——回归分析	10
3.2.3 实验三——相关分析与回归分析	12
3.2.4 实验四——逐步回归分析	14
3.2.4 实验五——主成分分析	16
四、总结及展望.....	19

一、赛题背景

本赛题旨在利用经脱敏后的锅炉传感器采集的数据（采集频率是分钟级别），根据锅炉的工况，预测产生的蒸汽量。

赛题背景基于火力发电的基本原理是：燃料在燃烧时加热水生成蒸汽，蒸汽压力推动汽轮机旋转，然后汽轮机带动发电机旋转，产生电能。在这一系列的能量转化中，影响发电效率的核心是锅炉的燃烧效率，即燃料燃烧加热水产生高温高压蒸汽。锅炉的燃烧效率的影响因素很多，包括锅炉的可调参数，如燃烧给量，一二次风，引风，返料风，给水水量；以及锅炉的工况，比如锅炉床温、床压，炉膛温度、压力，过热器的温度等。

二、数据介绍

数据分成训练数据（train.txt）和测试数据（test.txt），其中字段”V0”-“V37”，这 38 个字段是作为特征变量，”target”作为目标变量。利用训练数据训练出模型，预测测试数据的目标变量，排名结果依据预测结果的 MSE（mean square error）。

三、实验过程

3.1 描述统计

（1）检查数据有无缺失值

利用 Excel 查找替换功能检查缺失值，训练数据无缺失值。

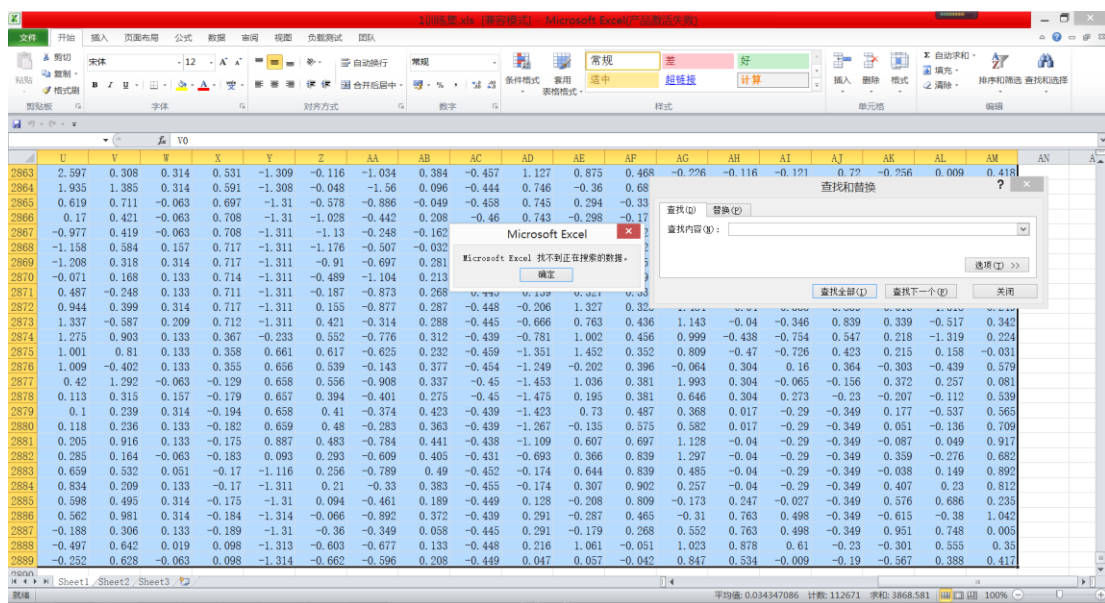


图 3-1-1 查找缺失值

(2)本次数据集所需要判断的是V0,V1,V2,V3,V4……V38与target变量之间的关系，首先是利用spss modler 18中的数据审核来查看数据的分布。

如图3-1-2所示，V0,V1,V2,V3,V4,V5,V6,V12,V13呈正态分布。

字段	样本图形	测量	最小值	最大值	平均值	标准差	偏度	唯一	有效
V0		连续	-4.335	2.121	0.123	0.928	-1.275	--	2888
V1		连续	-5.122	1.918	0.056	0.942	-1.637	--	2888
V2		连续	-3.420	2.828	0.290	0.911	-0.300	--	2888
V3		连续	-3.956	2.457	-0.068	0.970	-0.353	--	2888
V4		连续	-4.742	2.689	0.013	0.888	-1.023	--	2888
V5		连续	-2.182	0.489	-0.559	0.518	-0.749	--	2888
V6		连续	-4.576	1.895	0.183	0.918	-1.307	--	2888
V7		连续	-5.048	1.918	0.116	0.955	-1.433	--	2888
V8		连续	-4.692	2.245	0.178	0.895	-1.252	--	2888
V9		连续	-12.891	1.335	-0.169	0.954	-2.736	--	2888
V10		连续	-2.584	4.830	0.034	0.968	-0.553	--	2888
V11		连续	-3.160	1.455	-0.364	0.859	-0.901	--	2888
V12		连续	-5.165	2.657	0.023	0.894	-1.150	--	2888
V13		连续	-3.675	2.475	0.196	0.923	-0.558	--	2888
V14		连续	-2.455	2.558	0.016	1.016	0.185	--	2888

图 3-1-2 训练集数据审核结果 1

字段	样本图形	测量	最小值	最大值	平均值	标准差	偏度	唯一	有效
V15		连续	-2.903	4.314	0.096	1.033	0.451	--	2888
V16		连续	-5.981	2.861	0.114	0.983	-1.526	--	2888
V17		连续	-2.224	2.023	-0.043	0.656	-0.385	--	2888
V18		连续	-3.582	4.441	0.055	0.953	-1.019	--	2888
V19		连续	-3.704	3.431	-0.115	1.109	-0.187	--	2888
V20		连续	-3.402	3.525	-0.186	0.789	0.126	--	2888
V21		连续	-2.643	2.259	-0.057	0.781	-0.050	--	2888
V22		连续	-1.375	2.018	0.303	0.639	0.303	--	2888
V23		连续	-5.542	1.906	0.156	0.979	-3.417	--	2888
V24		连续	-1.344	2.423	-0.022	1.033	0.031	--	2888
V25		连续	-3.808	7.284	-0.052	0.916	1.539	--	2888
V26		连续	-5.131	2.980	0.072	0.890	-0.601	--	2888
V27		连续	-1.164	0.925	0.272	0.270	-1.097	--	2888
V28		连续	-2.435	4.671	0.138	0.930	1.002	--	2888
V29		连续	-2.912	4.580	0.098	1.061	0.486	--	2888

图 3-1-3 训练集数据审核结果 2

如图 3-1-3 所示，V15, V16, V20, V21, V25, V26, V27, V29 分布较为集中

V30		连续	-4.507	2.689	0.055	0.902	-2.185	--	2888
V31		连续	-5.859	2.013	0.128	0.873	-1.679	--	2888
V32		连续	-4.053	2.395	0.021	0.903	-1.597	--	2888
V33		连续	-4.627	5.465	0.008	1.007	-0.059	--	2888
V34		连续	-4.789	5.110	0.007	1.003	-0.453	--	2888
V35		连续	-5.695	2.324	0.198	0.986	-2.115	--	2888
V36		连续	-2.608	5.238	0.031	0.971	-0.593	--	2888
V37		连续	-3.630	3.000	-0.130	1.017	0.147	--	2888
target		连续	-3.044	2.538	0.126	0.984	-0.898	--	2888

图 3-1-4 训练集数据审核结果 3

如图 3-1-4 所示，V31, V37, target 分布较为集中，近似于正态分布。

字段	测量	高群值	低值	操作	缺失插补	方法	完成百分比	有效记录	空值	字符串空值	空白	空白值
V0	连续	43	0元	从下	从不	固定	100	2888	0	0	0	0
V1	连续	54	3元	从下	从不	固定	100	2888	0	0	0	0
V2	连续	7	0元	从下	从不	固定	100	2888	0	0	0	0
V3	连续	3	0元	从下	从不	固定	100	2888	0	0	0	0
V4	连续	46	1元	从下	从不	固定	100	2888	0	0	0	0
V5	连续	46	0元	从下	从不	固定	100	2888	0	0	0	0
V6	连续	28	1元	从下	从不	固定	100	2888	0	0	0	0
V7	连续	43	3元	从下	从不	固定	100	2888	0	0	0	0
V8	连续	37	0元	从下	从不	固定	100	2888	0	0	0	0
V9	连续	28	12元	从下	从不	固定	100	2888	0	0	0	0
V10	连续	3	0元	从下	从不	固定	100	2888	0	0	0	0
V11	连续	7	0元	从下	从不	固定	100	2888	0	0	0	0
V12	连续	40	3元	从下	从不	固定	100	2888	0	0	0	0
V13	连续	18	0元	从下	从不	固定	100	2888	0	0	0	0
V14	连续	0	0元	从下	从不	固定	100	2888	0	0	0	0
V15	连续	11	0元	从下	从不	固定	100	2888	0	0	0	0
V16	连续	43	5元	从下	从不	固定	100	2888	0	0	0	0
V17	连续	5	0元	从下	从不	固定	100	2888	0	0	0	0
V18	连续	101	0元	从下	从不	固定	100	2888	0	0	0	0
V19	连续	3	0元	从下	从不	固定	100	2888	0	0	0	0
V20	连续	22	0元	从下	从不	固定	100	2888	0	0	0	0
V21	连续	1	0元	从下	从不	固定	100	2888	0	0	0	0
V22	连续	0	0元	从下	从不	固定	100	2888	0	0	0	0
V23	连续	42	39元	从下	从不	固定	100	2888	0	0	0	0
V24	连续	0	0元	从下	从不	固定	100	2888	0	0	0	0
V25	连续	30	13元	从下	从不	固定	100	2888	0	0	0	0
V26	连续	19	5元	从下	从不	固定	100	2888	0	0	0	0
V27	连续	33	5元	从下	从不	固定	100	2888	0	0	0	0
V28	连续	32	0元	从下	从不	固定	100	2888	0	0	0	0
V29	连续	8	0元	从下	从不	固定	100	2888	0	0	0	0
V30	连续	29	39元	从下	从不	固定	100	2888	0	0	0	0
V31	连续	39	7元	从下	从不	固定	100	2888	0	0	0	0
V32	连续	68	0元	从下	从不	固定	100	2888	0	0	0	0
V33	连续	42	5元	从下	从不	固定	100	2888	0	0	0	0
V34	连续	48	4元	从下	从不	固定	100	2888	0	0	0	0
V35	连续	55	0元	从下	从不	固定	100	2888	0	0	0	0
V36	连续	4	11元	从下	从不	固定	100	2888	0	0	0	0
V37	连续	7	0元	从下	从不	固定	100	2888	0	0	0	0
target	连续	27	0元	从下	从不	固定	100	2888	0	0	0	0

图 3-1-5 训练集数据审核结果 4

综上所述，我们可以着重探究 V0, V1, V2, V3, V4, V5, V6, V12, V13, V15, V16, V20, V21, V25, V26, V27, V29, V31, V37 与 target 之间的关系。

(3) 利用 spss modeler 中的特征选择选择工具，我们可以依此判断哪些变量与我们的目标变量相关性强。由赛题背景和查阅相关资料可知，锅炉的燃烧效率的影响因素很多，包括锅炉的可调参数，如燃烧给量，一二次风，引风，返料风，给水水量；以及锅炉的工况，比如锅炉床温、床压，炉膛温度、压力，过热器的温度等。

模型 摘要 注解					
<input checked="" type="checkbox"/> 排名					
	序	字段	测量	重要性	值
<input checked="" type="checkbox"/>	1	V0	连续	重要	1.0
<input checked="" type="checkbox"/>	2	V1	连续	重要	1.0
<input checked="" type="checkbox"/>	3	V8	连续	重要	1.0
<input checked="" type="checkbox"/>	4	V27	连续	重要	1.0
<input checked="" type="checkbox"/>	5	V31	连续	重要	1.0
<input checked="" type="checkbox"/>	6	V2	连续	重要	1.0
<input checked="" type="checkbox"/>	7	V4	连续	重要	1.0
<input checked="" type="checkbox"/>	8	V12	连续	重要	1.0
<input checked="" type="checkbox"/>	9	V16	连续	重要	1.0
<input checked="" type="checkbox"/>	10	V3	连续	重要	1.0
<input checked="" type="checkbox"/>	11	V20	连续	重要	1.0
<input checked="" type="checkbox"/>	12	V10	连续	重要	1.0
<input checked="" type="checkbox"/>	13	V6	连续	重要	1.0
<input checked="" type="checkbox"/>	14	V36	连续	重要	1.0
<input checked="" type="checkbox"/>	15	V7	连续	重要	1.0
<input checked="" type="checkbox"/>	16	V23	连续	重要	1.0
<input checked="" type="checkbox"/>	17	V13	连续	重要	1.0
<input checked="" type="checkbox"/>	18	V30	连续	重要	1.0
<input checked="" type="checkbox"/>	19	V18	连续	重要	1.0
<input checked="" type="checkbox"/>	20	V15	连续	重要	1.0

选定字段数：34 可用字段总数：38

☒ > 0.95
 ☒ <= 0.95
 ☐ < 0.9

图 3-1-6 特征选择结果 1

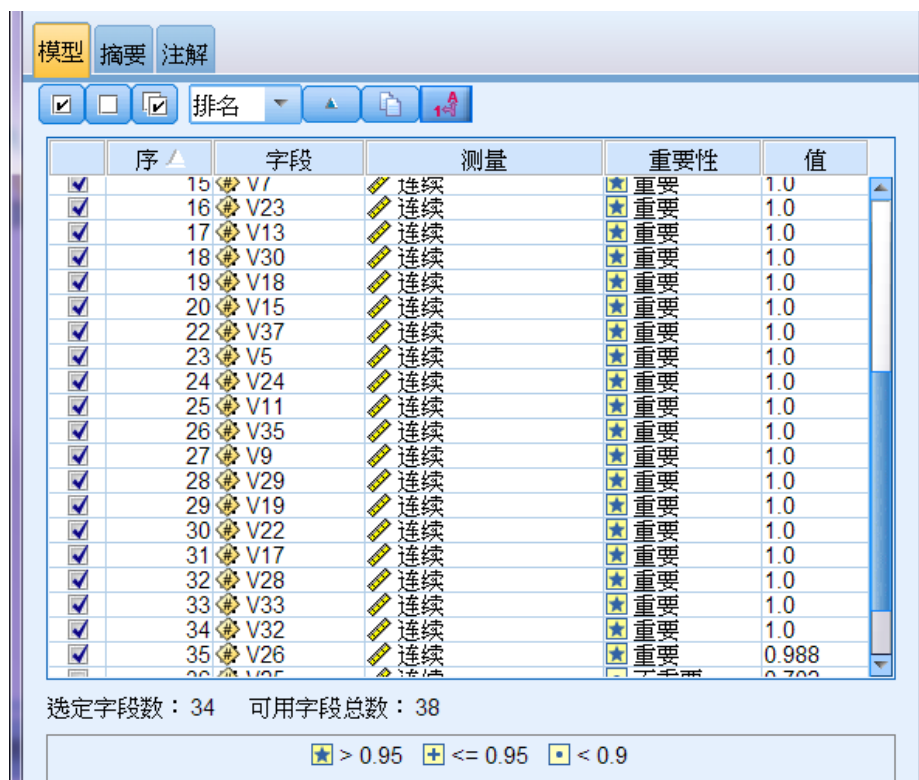


图 3-1-7 特征选择结果 2

由图 3-1-6, 图 3-1-7 的特征选择结果我们可以得知, V0, V1, V2, V3, V4……V38 与 target 变量之间的相关性大小顺序排列如下 V0, V1, V8, V27, V31, V2, V4, V12, V16, V3, V20, V10, V6, V36, V7, V23, V13, V30, V18, V15, V37, V5, V24, V11, V35, V9, V29, V19, V22, V17, V28, V33, V32。

(4) 变量与目标之间关系的散点图展示

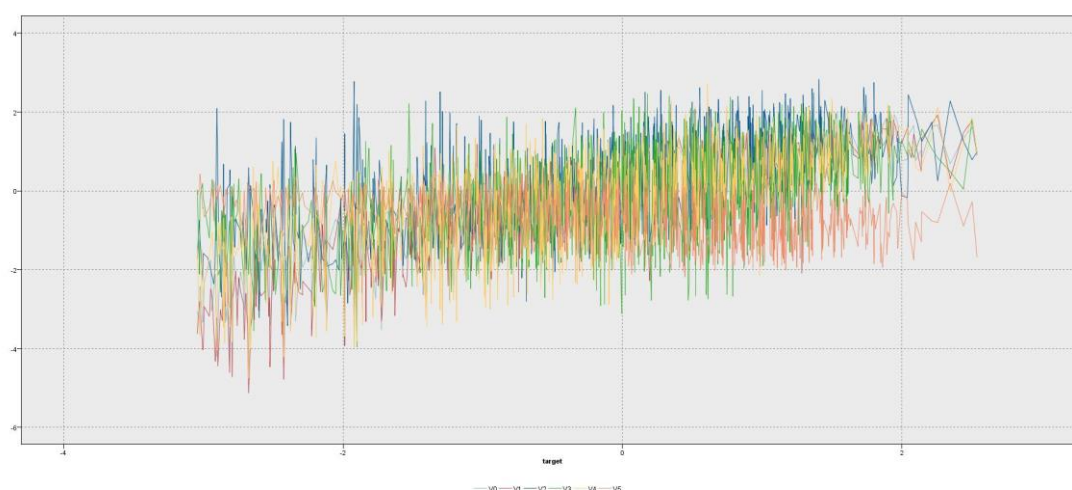


图 3-1-8 V0, V1, V2, V3, V4, V5 与 target 之间的散点图

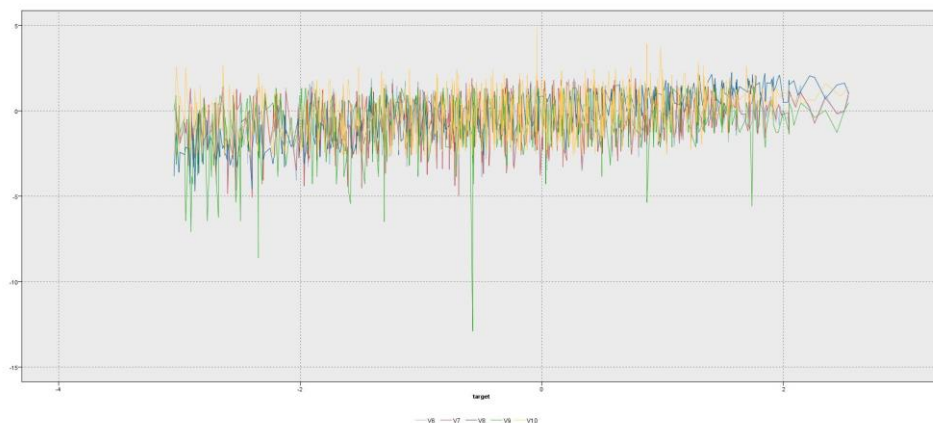


图 3-1-9 V6, V7, V8, V9, V10 与 target 之间的散点图

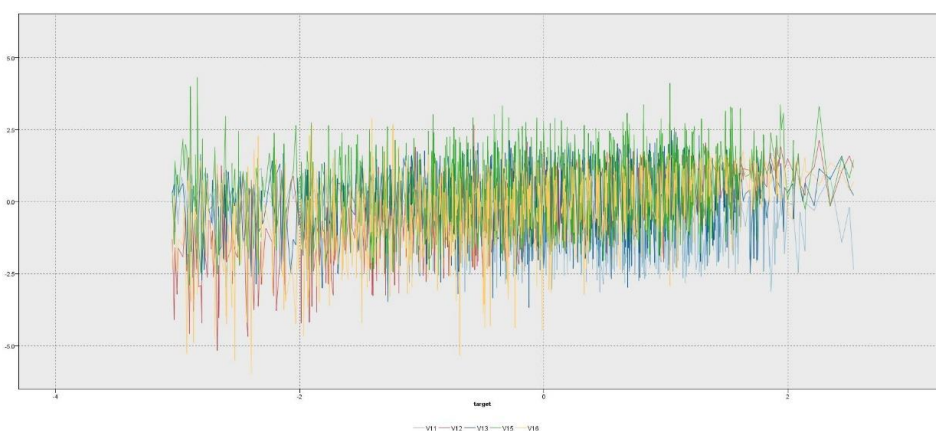


图 3-1-10 V11, V12, V13, V15, V16 与 target 之间的散点图

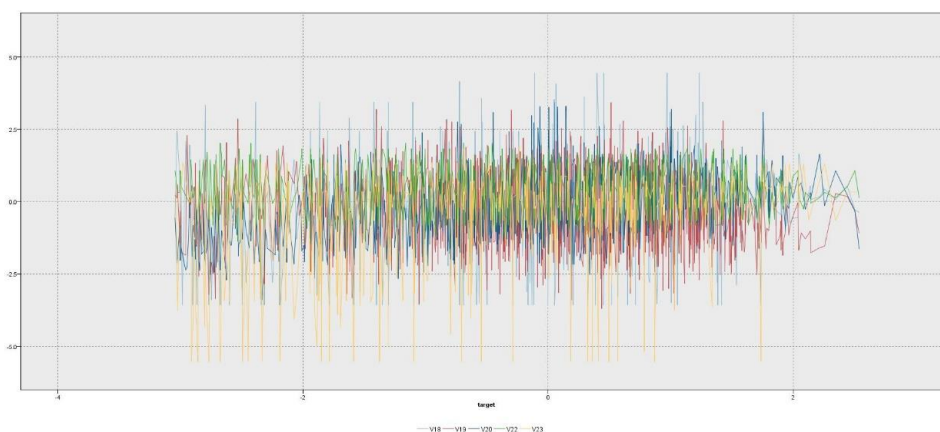


图 3-1-11 V18, V19, V20, V22, V23 与 target 之间的散点图

(5) 本部分主要是描述数据位置，离散程度，分布形态

平均数是表示一组数据集中趋势的量数，是指在一组数据中所有数据之和再除以这组数据的个数。它是反映数据集中趋势的一项指标。解答平均数应用题的关键在于确定“总数量”以及和总数量对应的总份数。在统计工作中，平均数（均值）和标准差是描述数据资料集中趋势和离散程度的两个最重要的测度值。本次数据报告中的均值为 0.128，锅炉燃烧效率的整体水平是 0.128；标准差为 0.98，总体离散程度较高；中值是 0.313，出现在中间位置的数值是 0.313；众数是 -0.02，出现次数最多的是 -0.02；全距为 5.578，最大离散是 5.578；

峰度是 0.763，偏度是-0.899 分布平坦，右偏。

统计量

target		
N	有效	2888
	缺失	0
均值		.12804
均值的标准误		.018276
中值		.31300
众数		-.020
标准差		.982169
方差		.965
偏度		-.899
偏度的标准误		.046
峰度		.763
峰度的标准误		.091
全距		5.578
极小值		-3.040
极大值		2.538
和		369.768
百分位数	25	-.35000
	50	.31300
	75	.79375

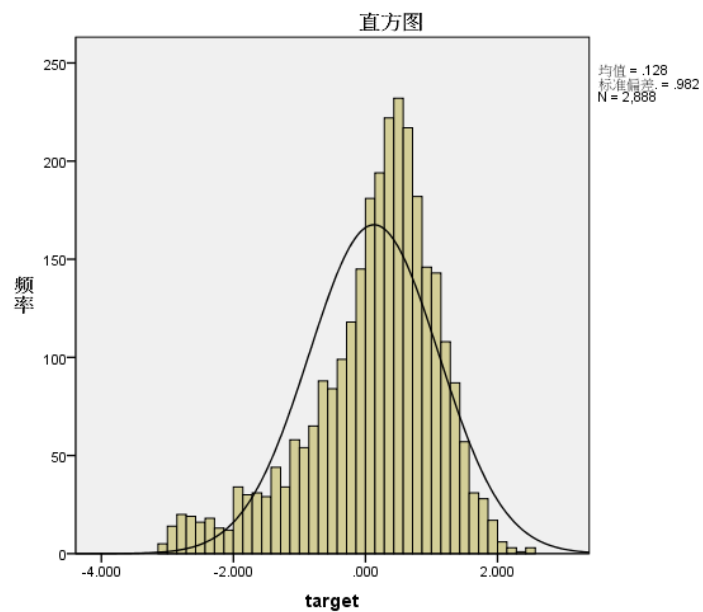


图 3-1-12 直方图

3.2 分析过程

3.2.1 实验一——回归分析

基于对数据的观察和描述统计分析，由于给出的训练集数据之间没有明确的关系，所以我们首先对全变量进行了回归分析，观察目标变量与其他所有变量之间的线性关系。

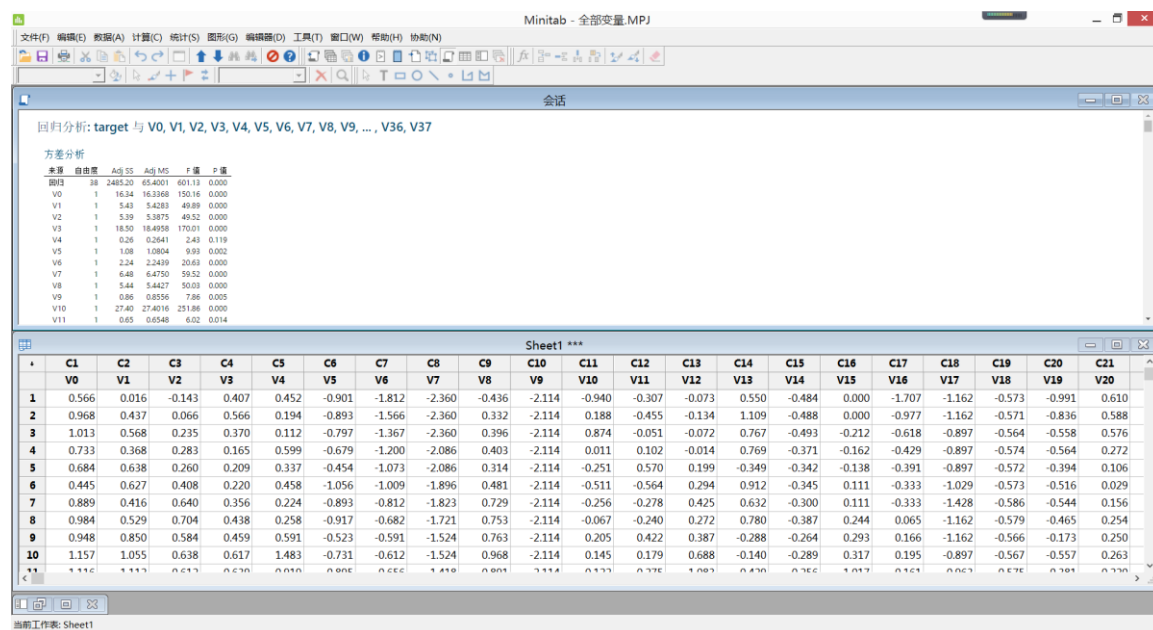


图 3-2-1 全变量回归

回归方程

$$\begin{aligned} \text{target} = & -0.2953 + 0.3400 V0 + 0.1823 V1 + 0.1562 V2 + 0.12181 V3 + 0.0408 V4 \\ & - 0.1156 V5 + 0.1577 V6 - 0.1866 V7 - 0.2055 V8 + 0.0289 V9 + 0.3338 V10 \\ & + 0.0503 V11 + 0.0954 V12 - 0.0123 V13 + 0.0560 V14 + 0.0074 V15 \\ & + 0.0396 V16 + 0.0968 V17 + 0.01286 V18 + 0.02086 V19 + 0.0120 V20 \\ & - 0.0091 V21 + 0.0164 V22 + 0.0174 V23 - 0.04237 V24 - 0.0312 V25 \\ & + 0.02703 V26 + 1.1285 V27 - 0.01032 V28 - 0.0435 V29 + 0.0176 V30 \\ & - 0.0003 V31 - 0.0089 V32 + 0.01900 V33 - 0.00647 V34 - 0.01814 V35 \\ & - 0.2613 V36 - 0.0552 V37 \end{aligned}$$

图 3-2-2 全变量回归方程

利用得出的回归方程预测测试集中的目标变量，提交的结果均方误差（mean square error）为 3.0896，结果不是很理想，有待优化。

3.2.2 实验二——回归分析

第一次实验中回归方程的变量较多，所以我们基于回归分析中的 P 值将不显著的变量进行删减（删除 P>0.05 的变量），重新进行回归分析。

系数

项	系数	系数标准误	T 值	P 值	方差膨胀因子
常量	-0.2953	0.0230	-12.82	0.000	
V0	0.3400	0.0277	12.25	0.000	17.59
V1	0.1823	0.0258	7.06	0.000	15.67
V2	0.1562	0.0222	7.04	0.000	10.85
V3	0.12181	0.00934	13.04	0.000	2.18
V4	0.0408	0.0262	1.56	0.119	14.34
V5	-0.1156	0.0367	-3.15	0.002	9.58
V6	0.1577	0.0347	4.54	0.000	26.95
V7	-0.1866	0.0242	-7.71	0.000	14.17
V8	-0.2055	0.0291	-7.07	0.000	17.96
V9	0.0289	0.0103	2.80	0.005	2.57
V10	0.3338	0.0210	15.87	0.000	11.01
V11	0.0503	0.0205	2.45	0.014	8.23
V12	0.0954	0.0232	4.12	0.000	11.37
V13	-0.0123	0.0126	-0.97	0.332	3.61
V14	0.0560	0.0110	5.09	0.000	3.32
V15	0.0074	0.0221	0.34	0.736	13.78
V16	0.0396	0.0283	1.40	0.163	20.57
V17	0.0968	0.0134	7.24	0.000	2.04
V18	0.01286	0.00976	1.32	0.188	2.30
V19	0.02086	0.00909	2.30	0.022	2.70
V20	0.0120	0.0106	1.13	0.257	1.84
V21	-0.0091	0.0113	-0.81	0.420	2.06
V22	0.0164	0.0170	0.97	0.333	3.12
V23	0.0174	0.0112	1.56	0.120	3.17
V24	-0.04237	0.00956	-4.43	0.000	2.59
V25	-0.0312	0.0140	-2.23	0.026	4.35
V26	0.02703	0.00899	3.01	0.003	1.70
V27	1.1285	0.0709	15.91	0.000	9.76
V28	-0.01032	0.00744	-1.39	0.166	1.27
V29	-0.0435	0.0233	-1.87	0.062	16.18
V30	0.0176	0.0111	1.60	0.111	2.64
V31	-0.0003	0.0235	-0.01	0.989	11.13
V32	-0.0089	0.0103	-0.87	0.384	2.27
V33	0.01900	0.00977	1.95	0.052	2.57
V34	-0.00647	0.00931	-0.70	0.487	2.32
V35	-0.01814	0.00964	-1.88	0.060	2.40
V36	-0.2613	0.0204	-12.79	0.000	10.44
V37	-0.0552	0.0158	-3.49	0.000	6.85

图 3-2-3 第一次回归分析结果

变量 V4、V13、V15、V16、V18、V20、V21、V22、V23、V28、V29、V30、V31、V32、V33、V34、V35 的 P 值均大于 0.05，故删除这些变量，再进行回归分析。

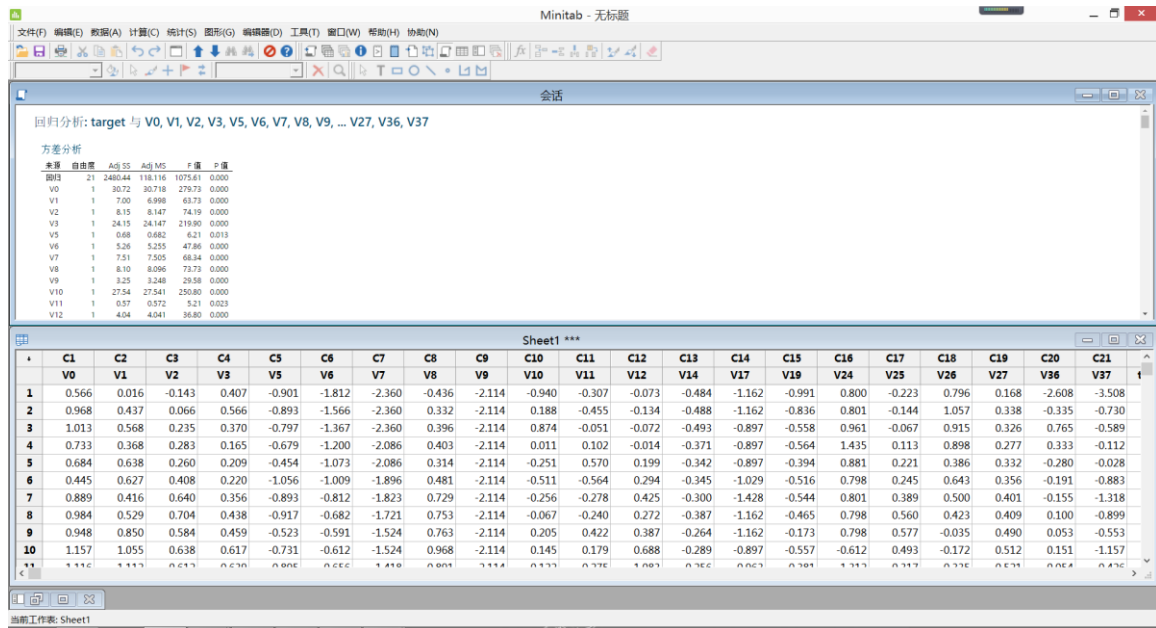


图 3-2-4 第二次回归

回归方程

$$\begin{aligned} \text{target} = & -0.2871 + 0.3737 V0 + 0.1902 V1 + 0.1617 V2 + 0.12505 V3 - 0.0866 V5 \\ & + 0.1840 V6 - 0.1876 V7 - 0.2139 V8 + 0.04614 V9 + 0.3234 V10 \\ & + 0.0444 V11 + 0.0978 V12 + 0.05957 V14 + 0.0980 V17 + 0.01192 V19 \\ & - 0.03492 V24 - 0.0096 V25 + 0.01751 V26 + 1.1353 V27 - 0.2440 V36 \\ & - 0.0558 V37 \end{aligned}$$

图 3-2-5 第二次回归方程

利用此次得出的回归方程预测测试集中的目标变量，提交的结果均方误差（mean square error）为 3.0831，结果依旧不是很理想，无较大进步。

3.2.3 实验三——相关分析与回归分析

由于第一、二次实验单纯根据回归分析所得的结果不理想，仔细分析，我们的目标是根据训练集构建一个预测模型，得出目标变量，所以应该依据目标变量与其他变量之间的关系来构建模型。我们首先想到了相关分析。

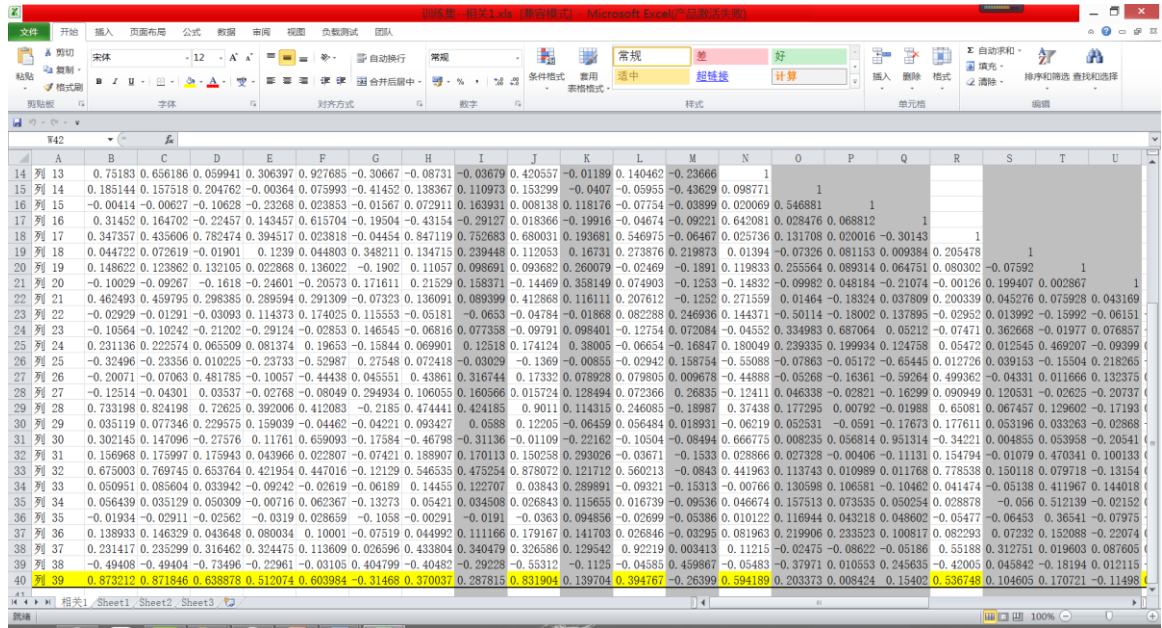


图 3-2-6 相关分析

将相关系数小于 0.3 的变量删除。下一步进行回归分析。

系数

项	系数	系数标准误差	T 值	P 值	方差膨胀因子
常量	-0.2355	0.0195	-12.09	0.000	
V0	0.3551	0.0266	13.36	0.000	14.74
V1	0.2748	0.0242	11.35	0.000	12.59
V2	0.1029	0.0188	5.47	0.000	7.12
V3	0.11113	0.00805	13.81	0.000	1.48
V4	0.0427	0.0246	1.73	0.083	11.61
V5	-0.0460	0.0159	-2.89	0.004	1.64
V6	0.0102	0.0153	0.67	0.502	4.75
V8	-0.2271	0.0290	-7.83	0.000	16.33
V10	0.3384	0.0215	15.76	0.000	10.47
V12	0.0607	0.0218	2.79	0.005	9.20
V16	0.0675	0.0249	2.71	0.007	14.50
V20	-0.00713	0.00957	-0.75	0.456	1.38
V27	0.9939	0.0709	14.02	0.000	8.89
V31	-0.0190	0.0236	-0.80	0.421	10.29
V36	-0.2353	0.0202	-11.62	0.000	9.35
V37	-0.0738	0.0138	-5.34	0.000	4.78

图 3-2-7 回归分析

变量 V4、V6、V20、V31 的 P 值均大于 0.05，故删除这些变量，再进行一次回归分析，得出回归方程。

回归方程

$$\begin{aligned} \text{target} = & -0.2380 + 0.3790 V0 + 0.2623 V1 + 0.0997 V2 + 0.11107 V3 - 0.0443 V5 \\ & - 0.2406 V8 + 0.3348 V10 + 0.0838 V12 + 0.0697 V16 + 1.0150 V27 \\ & - 0.2352 V36 - 0.0697 V37 \end{aligned}$$

图 3-2-8 回归方程

利用此次得出的回归方程预测测试集中的目标变量，提交的结果均方误差 (mean square

error) 为 2.5639, 结果有进步, 所以我们在相关分析上进行了改善, 保留与目标变量相关性更大的变量进行下一步分析。

系数

项	系数	系数标准误	T 值	P 值	方差膨胀因子
常量	-0.1067	0.0172	-6.20	0.000	
V0	0.4295	0.0269	15.95	0.000	13.78
V1	0.2777	0.0248	11.20	0.000	12.02
V2	0.1602	0.0180	8.91	0.000	5.92
V3	0.11460	0.00819	14.00	0.000	1.39
V4	-0.0091	0.0254	-0.36	0.720	11.27
V8	-0.0981	0.0279	-3.51	0.000	13.80
V12	0.0621	0.0219	2.84	0.005	8.43
V16	0.0666	0.0165	4.03	0.000	5.83
V27	0.4688	0.0664	7.06	0.000	7.11
V31	0.0225	0.0236	0.95	0.341	9.35
V37	-0.0302	0.0135	-2.24	0.025	4.17

图 3-2-9 回归分析

删除 P 值大于 0.05 的变量, 即删去 V4、V31。得出回归方程如下:

回归方程

$$\begin{aligned} \text{target} = & -0.1043 + 0.4215 V0 + 0.2844 V1 + 0.1612 V2 + 0.11408 V3 - 0.0865 V8 \\ & + 0.0610 V12 + 0.0763 V16 + 0.4603 V27 - 0.0288 V37 \end{aligned}$$

图 3-2-10 回归方程

根据这次得出的回归方程对训练集数据的目标变量进行预测, 提交的结果均方误差 (mean square error) 为 0.6032, 相较前几次分析结果有很大的进步。

3.2.4 实验四——逐步回归分析

利用逐步回归分析方法, 选取 $\alpha = 0.05$, 筛选出重要性变量, 得出残差正态概率图及直方图, 判断回归方程的可靠性。

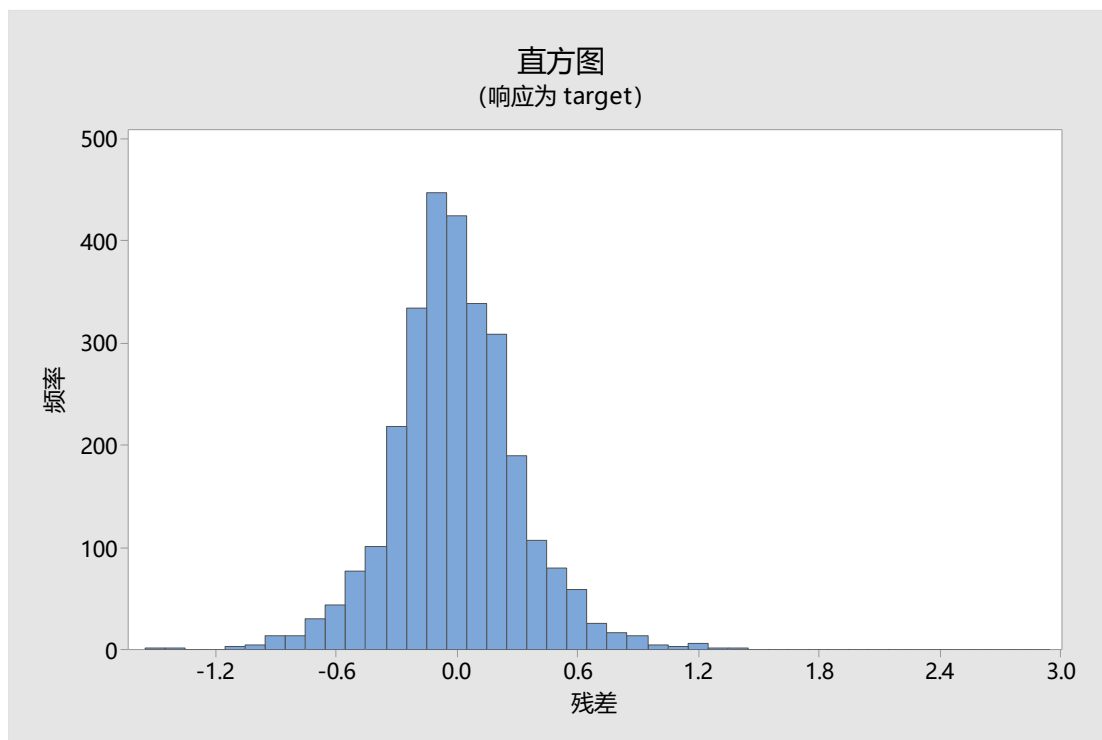


图 3-2-11 残差直方图

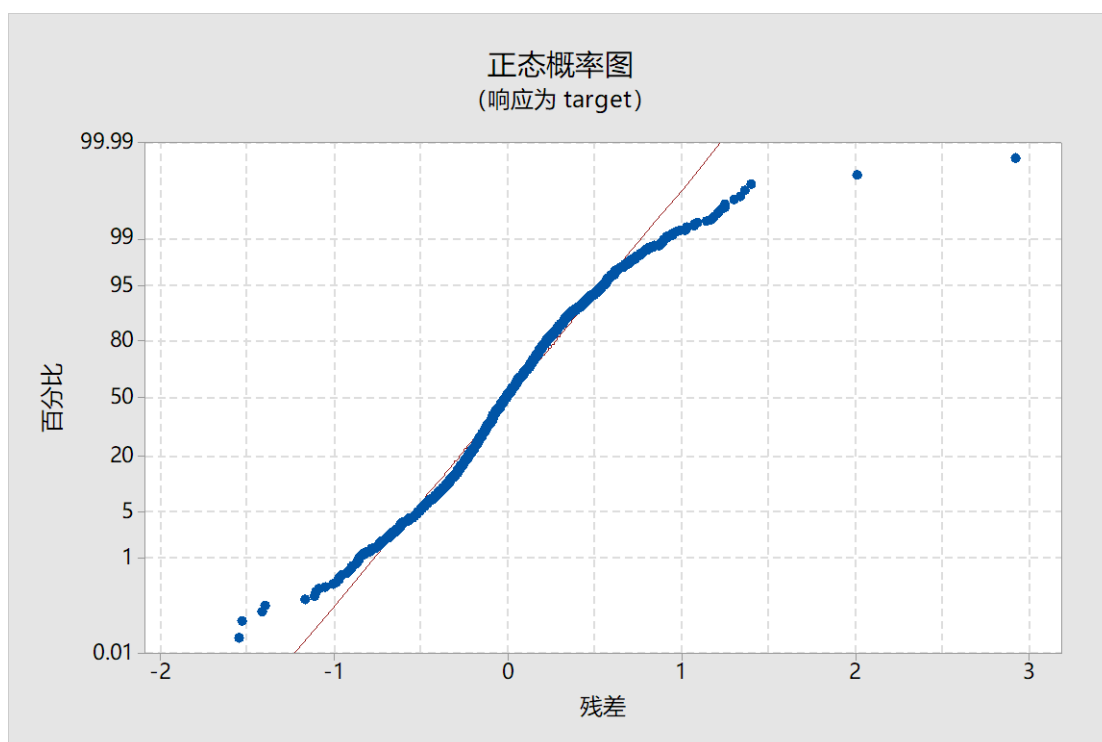


图 3-2-12 正态概率图

由直方图可以看出，残差直方图呈现正态分布；由残差正态概率图可以看出，散点基本呈直线分布，可以认为残差服从正态分布，所以回归方程可靠。

回归方程

$$\begin{aligned} \text{target} = & -0.2508 + 0.3710 V0 + 0.1805 V1 + 0.1633 V2 + 0.12945 V3 + 0.1642 V6 \\ & - 0.1839 V7 - 0.2096 V8 + 0.04023 V9 + 0.3326 V10 + 0.1221 V12 \\ & + 0.06131 V14 + 0.1003 V17 + 0.02815 V18 - 0.04166 V24 + 1.1350 V27 \\ & - 0.0293 V29 - 0.2513 V36 - 0.0486 V37 \end{aligned}$$

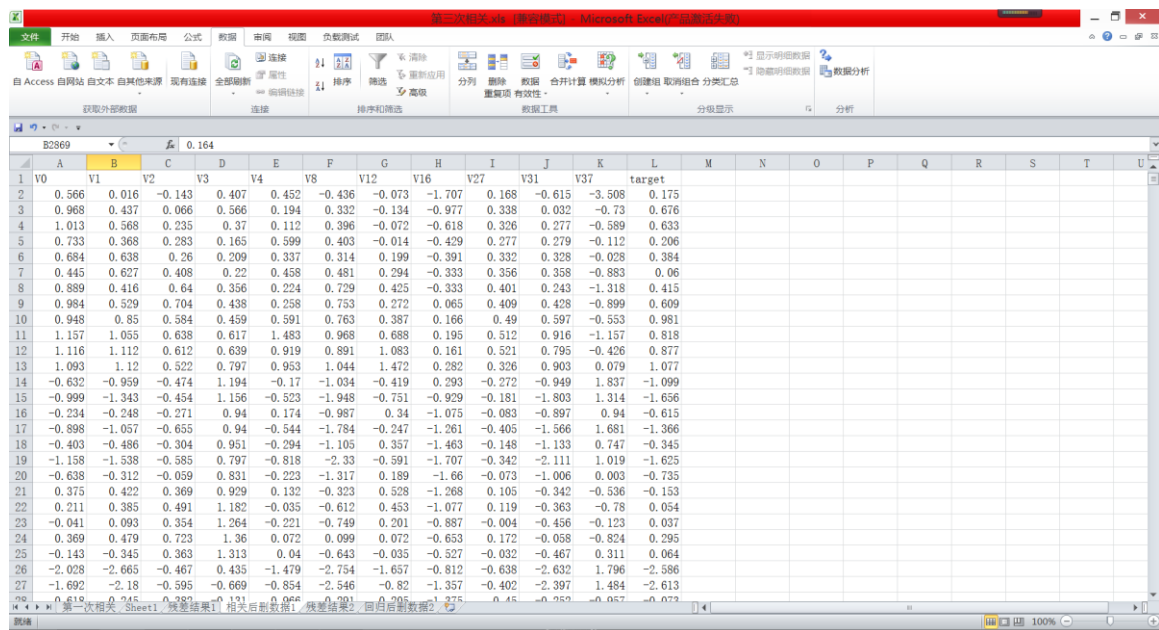
图 3-2-13 逐步回归的方程

根据这次得出的回归方程对训练集数据的目标变量进行预测，提交的结果均方误差（mean square error）为 2.8437，相较前几次分析结果并没有很大的进步，再考虑其他方法。

3.2.4 实验五——主成分分析

在先前实验中，我们选择利用多次回归的方式来剔除不显著变量，用较少的变量来构建出回归方程。本次实验中为保留原始变量的特性，我们选择利用主成分分析法，利用降维（线性变换）的思想，在损失很少信息的前提下把多个指标转化为几个不相关的综合指标（主成分），即每个主成分都是原始变量的线性组合，且各个主成分之间互不相关，使得主成分比原始变量具有某些更优越的性能（主成分必须保留原始变量 90% 以上的信息），从而达到简化系统结构，抓住问题实质的目的。

第一步利用相关分析将相关系数小于 0.5 的变量删除；



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	V0	V1	V2	V3	V4	V8	V12	V16	V27	V31	V37	target									
2	0.566	0.016	-0.143	0.407	0.452	-0.436	-0.073	-1.707	0.168	-0.615	-3.508	0.175									
3	0.968	0.437	0.066	0.566	0.194	0.332	-0.134	-0.977	0.338	0.032	-0.73	0.676									
4	1.013	0.568	0.235	0.37	0.112	0.396	-0.072	-0.618	0.326	0.277	-0.589	0.633									
5	0.733	0.368	0.283	0.165	0.599	0.403	-0.014	-0.429	0.277	0.279	-0.112	0.206									
6	0.684	0.638	0.26	0.209	0.337	0.314	0.199	-0.391	0.332	0.328	-0.028	0.384									
7	0.445	0.627	0.408	0.22	0.458	0.481	0.294	-0.333	0.356	0.358	-0.883	0.06									
8	0.889	0.416	0.64	0.356	0.224	0.729	0.425	-0.333	0.401	0.243	-1.318	0.415									
9	0.984	0.529	0.704	0.438	0.258	0.753	0.272	0.065	0.409	0.428	-0.899	0.609									
10	0.948	0.85	0.584	0.459	0.591	0.763	0.387	0.166	0.49	0.597	-0.553	0.981									
11	1.157	1.055	0.638	0.617	1.483	0.968	0.688	0.195	0.512	0.916	-1.157	0.818									
12	1.116	1.112	0.612	0.639	0.919	0.891	1.083	0.161	0.521	0.795	-0.426	0.877									
13	1.093	1.12	0.522	0.797	0.953	1.044	1.472	0.282	0.326	0.903	0.079	1.077									
14	-0.632	-0.959	-0.474	1.194	-0.17	-1.034	-0.419	0.293	-0.272	-0.949	1.837	-1.099									
15	-0.999	-1.343	-0.454	1.156	-0.523	-1.948	-0.751	-0.929	-0.181	-1.803	1.314	-1.656									
16	-0.234	-0.248	-0.271	0.94	0.174	-0.987	0.34	-1.075	-0.083	-0.897	0.94	-0.615									
17	-0.898	-1.057	-0.655	0.94	-0.544	-1.784	-0.247	-1.261	-0.405	-1.566	1.681	-1.366									
18	-0.403	-0.486	-0.304	0.951	-0.294	-1.105	0.357	-1.463	-0.148	-1.133	0.747	-0.345									
19	-1.158	-1.538	-0.585	0.797	-0.818	-2.33	-0.591	-1.707	-0.342	-2.111	1.019	-1.625									
20	-0.638	-0.312	-0.059	0.831	-0.223	-1.317	0.189	-1.66	-0.073	-1.006	0.003	-0.735									
21	0.375	0.422	0.369	0.929	0.132	-0.323	0.528	-1.268	0.105	-0.342	-0.536	-0.153									
22	0.211	0.385	0.491	1.182	-0.035	-0.612	0.453	-1.077	0.119	-0.363	-0.78	0.054									
23	-0.041	0.093	0.354	1.264	-0.221	-0.749	0.201	-0.887	-0.004	-0.456	-0.123	0.037									
24	0.369	0.479	0.723	1.36	0.072	0.099	0.072	-0.653	0.172	-0.058	-0.824	0.295									
25	-0.143	-0.345	0.363	1.313	0.04	-0.643	-0.035	-0.527	-0.032	-0.467	0.311	0.064									
26	-2.028	-2.665	-0.467	0.435	-1.479	-2.754	-1.657	-0.812	-0.638	-2.632	1.796	-2.586									
27	-1.692	-2.18	-0.595	-0.669	-0.854	-2.546	-0.82	-1.357	-0.402	-2.397	1.484	-2.613									
28	0.618	0.245	0.282	0.131	0.064	0.201	0.205	0.45	-0.282	-0.057	-0.073										

图 3-2-14 相关分析后数据

第二步利用主成分分析法，从相关分析后得到的数据集中构建主分量。

首先利用特征值的大小确定主分量数。保留具有最大特征值的主分量。我们选取特征值大于 1 的主分量。

为了直观地比较特征值的大小，我们利用碎石图，来实现基于特征值的大小确定分量数。

主成分分析: V0, V1, V2, V3, V4, V8, V12, V16, V27, V31, V37

相关矩阵的特征分析

特征值	6.4664	2.1658	0.8512	0.7003	0.3035	0.1486	0.1055	0.0963	0.0743	0.0508
比率	0.588	0.197	0.077	0.064	0.028	0.014	0.010	0.009	0.007	0.005
累积	0.588	0.785	0.862	0.926	0.953	0.967	0.976	0.985	0.992	0.997
特征值	0.0373									
比率	0.003									
累积	1.000									

特征向量

变量	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
V0	0.348	-0.235	-0.184	0.109	0.031	-0.269	-0.202	0.549	0.011	-0.071	-0.597
V1	0.360	-0.146	-0.171	-0.060	0.364	-0.247	-0.069	-0.128	-0.563	-0.375	0.379
V2	0.290	0.384	-0.024	0.143	-0.450	0.300	-0.631	-0.130	-0.181	-0.089	-0.005
V3	0.209	0.005	0.695	0.642	0.241	-0.026	0.035	-0.039	0.024	0.001	0.000
V4	0.236	-0.515	0.031	0.012	-0.277	0.273	-0.018	0.274	0.398	-0.231	0.489
V8	0.370	0.077	-0.050	-0.187	0.335	-0.058	-0.264	0.048	0.214	0.719	0.262
V12	0.231	-0.511	0.022	0.034	-0.409	-0.009	0.211	-0.414	-0.318	0.404	-0.189
V16	0.266	0.373	0.330	-0.308	-0.356	-0.168	0.413	0.423	-0.240	0.061	0.145
V27	0.358	0.113	-0.130	-0.084	0.311	0.728	0.370	-0.016	-0.036	-0.073	-0.256
V31	0.347	0.068	0.221	-0.404	-0.013	-0.275	-0.011	-0.468	0.465	-0.321	-0.222
V37	-0.237	-0.295	0.525	-0.501	0.161	0.254	-0.366	0.129	-0.263	-0.007	-0.143

图 3-2-15 相关矩阵的特征分析

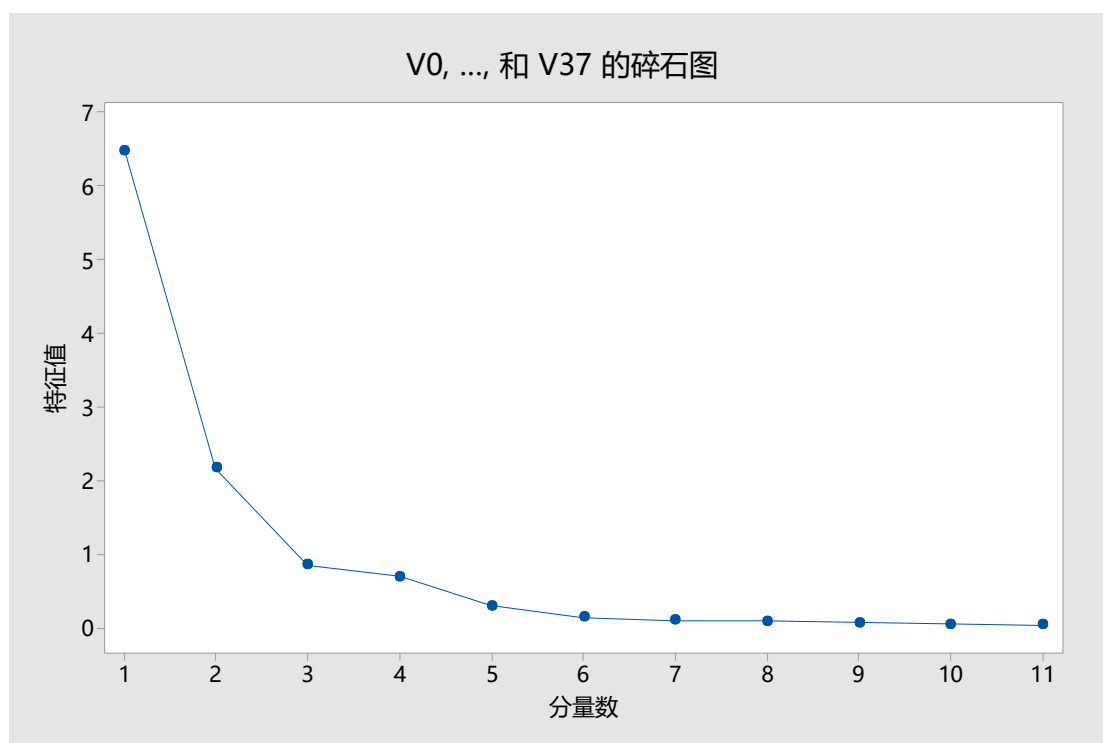


图 3-2-16 碎石图

在这些结果中，前两个主分量的特征值大于 1。这两个分量解释 78.5% 的数据变异。此碎石图显示特征值在第三个主分量之后开始形成直线。如果 78.5% 的已解释变异量在这些数据中已经足够，则应使用前两个主分量。

主分量是指原始变量的线性组合，这些变量说明数据中的方差。提取的最大分量数始终等于变量数。特征向量包括与每个变量相对应的系数，可用于计算主分量分值。这些系数表明分量中每个变量的相对权重。

系数的绝对值越大，对应变量的计算分量时就越重要。系数的绝对值多大才视为重要具有主观性。我们选择 V0、V1、V8、V27 和 V31 作为第一分量的变量，选择 V4 和 V12 作为第二分量的变量。

主成分分析: V0, V1, V2, V3, V4, V8, V12, V16, V27, V31, V37

相关矩阵的特征分析

特征值	6.4664	2.1658	0.8512	0.7003	0.3035	0.1486	0.1055	0.0963	0.0743	0.0508
比率	0.588	0.197	0.077	0.064	0.028	0.014	0.010	0.009	0.007	0.005
累积	0.588	0.785	0.862	0.926	0.953	0.967	0.976	0.985	0.992	0.997
特征值	0.0373									
比率	0.003									
累积	1.000									

特征向量

变量	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
V0	0.348	-0.235	-0.184	0.109	0.031	-0.269	-0.202	0.549	0.011	-0.071	-0.597
V1	0.360	-0.146	-0.171	-0.060	0.364	-0.247	-0.069	-0.128	-0.563	-0.375	0.379
V2	0.290	0.384	-0.024	0.143	-0.450	0.300	-0.631	-0.130	-0.181	-0.089	-0.005
V3	0.209	0.005	0.695	0.642	0.241	-0.026	0.035	-0.039	0.024	0.001	0.000
V4	0.236	-0.515	0.031	0.012	-0.277	0.273	-0.018	0.274	0.398	-0.231	0.489
V8	0.370	0.077	-0.050	-0.187	0.335	-0.058	-0.264	0.048	0.214	0.719	0.262
V12	0.231	-0.511	0.022	0.034	-0.409	-0.009	0.211	-0.414	-0.318	0.404	-0.189
V16	0.266	0.373	0.330	-0.308	-0.356	-0.168	0.413	0.423	-0.240	0.061	0.145
V27	0.358	0.113	-0.130	-0.084	0.311	0.728	0.370	-0.016	-0.036	-0.073	-0.256
V31	0.347	0.068	0.221	-0.404	-0.013	-0.275	-0.011	-0.468	0.465	-0.321	-0.222
V37	-0.237	-0.295	0.525	-0.501	0.161	0.254	-0.366	0.129	-0.263	-0.007	-0.143

图 3-2-17 相关矩阵的特征分析

在这些结果中，第一个主分量与 V0、V1、V8、V27 和 V31 具有较大的正关联，因为变量具体意义未知，所以只能将这 5 个变量归为第一变量。第二个分量与 V4 和 V12 具有较大的负关联，因此，将这 2 个变量归为第二变量。

所得方程为：

$$\text{target} = 0.348 * V0 + 0.36 * V1 + 0.37 * V8 + 0.358 * V27 + 0.347 * V31 - 0.515 * V4 - 0.511 * V12。$$

根据主成分分析得出的方程来预测测试集的目标变量，提交的结果的均方误差为 1.8457，效果没有第三次利用相关分析和回归的方法好，可能是因为我们没有熟练掌握这些方法。

四、总结及展望

在参加此次天池大数据竞赛过程中，我们小组成员经历了不少挫折和失败，根据大赛所给数据进行一步步的摸索，逐渐找到了合适的方向。经过对赛题所给数据的测试和实验，我们使用过的方法有相关分析、线性回归分析、逐步回归、主成分分析，根据各种方法得出的结果，我们小组成员一致认为以相关分析和线性回归相结合的方法来建立预测模型是较好的方法。

因为时间和技术等各方面原因，我们对目前提交的结果还不够满意，预想中会得到好结果的方法却不如最初的简单方法好。我们认为很大的原因还是由于自身对主成分分析、因子分析等方法掌握不够熟练，应该深入学习这些方法，综合自己所掌握的方法再进行分析建模。

我们这次主要用到了相关和回归分析，只考虑了一些简单的情况，还有进一步完善的空间。在比赛的论坛中，我们注意到很多参赛者都选择用代码来实现建模过程，效果比较理想，十分值得学习和借鉴。

通过这次参赛，我们也意识到数据分析往往不是一蹴而就的事情，有很多人认为搞大数据很难，其实只要找对方法，坚持下去，那就并不难，这是一个不断学习的过程，我的老师曾经对我们说过，机器学习就是一层窗户纸，捅破了，你就知道原来这么简单。