


对计算得出的音节数、词频、信息熵、九种单字母 (h、i、r、s、n、t、a、e、o) 出现频率和单词难度相关性最低的字母频率，五种单字母 (w、y、z、j、g) 出现频率和单词难度相关性最高的字母频率，一个单词中出现字母对的个数这六个因素对单词猜测贡献值的回归，拟采用多元线性回归分析（因为需要回归的变量中既有定量变量也有定性变量，因此选择这一模型）

首先导入数据并对数据做标记，然后对音节数、九种单字母出现频率和单词难度相关性最低的字母频率、五种单字母出现频率和单词难度相关性最高的字母频率、一个单词中是否出现多个字母这四个可以认为是离散的定性的变量做虚拟变量的处理；

考虑到验证模型的准确性这一问题，拟留出最后25个单词作为测试集，使用前334个作为训练集

A	对单词猜测贡献值	
 B	音节数	
C	词频	
D	信息熵	
E	九种单字母出现频率和...	
F	五种单字母出现频率和...	
G	单词中是否出现同一个...	
e1	E==	0.0000
e2	E==	1.0000
e3	E==	2.0000
e4	E==	3.0000
e5	E==	4.0000
e6	E==	5.0000
f1	F==	0.0000
f2	F==	1.0000
f3	F==	2.0000
f4	F==	3.0000
g1	G==	0.0000
g2	G==	1.0000
g3	G==	2.0000
g4	G==	3.0000
b1	B==	1.0000
b2	B==	2.0000
b3	B==	3.0000

对四个虚拟变量的解释：e1,e2,e3,e4,e5,e6分别表示九个单字母出现频率和单词难度相关性最低的字母在单词中出现0,1,2,3,4,5次；f1, f2, f3, f4分别表示五种单字母出现频率和单词难度相关性最高的字母在单词中分别出现0,1,2,3次；g1, g2, g3, g4表示每个单词中出现字母对的个数为0,1,2,3（g2例子为skill, g3例子为vivid, g4例子为mummy）

然后采用stata中带虚拟变量的多元线性回归分析，采用最小二乘法OLS做初步回归，得到的结果如下

Source	SS	df	MS	Number of obs	=	334
Model	.139426586	15	.009295106	F(15, 318)	=	14.78
Residual	.199998595	318	.000628926	Prob > F	=	0.0000
				R-squared	=	0.4108
				Adj R-squared	=	0.3830
Total	.339425181	333	.001019295	Root MSE	=	.02508

A	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
C	7.43e-09	7.60e-09	0.98	0.329	-7.53e-09	2.24e-08
D	-.1114745	.0422661	-2.64	0.009	-.1946311	-.0283179
b1	.0035799	.0032866	1.09	0.277	-.0028863	.0100461
b2	0 (omitted)					
b3	.0031904	.0076799	0.42	0.678	-.0119193	.0183002
e1	.0035186	.0163949	0.21	0.830	-.0287376	.0357749
e2	0 (omitted)					
e3	.0021079	.0082253	0.26	0.798	-.014075	.0182907
e4	.0036171	.0081776	0.44	0.659	-.0124719	.0197061
e5	.0081579	.0086367	0.94	0.346	-.0088345	.0251502
e6	.031557	.0099388	3.18	0.002	.0120029	.0511112
f1	.0315308	.0071565	4.41	0.000	.0174509	.0456108
f2	.0192055	.007329	2.62	0.009	.0047861	.0336249
f3	0 (omitted)					
f4	.0073149	.0145475	0.50	0.615	-.0213066	.0359364
g1	.0713942	.0237737	3.00	0.003	.0246205	.1181678
g2	.045804	.0235793	1.94	0.053	-.0005872	.0921952
g3	.0353657	.0269133	1.31	0.190	-.017585	.0883163
g4	0 (omitted)					
_cons	.2854648	.0556991	5.13	0.000	.1758795	.39505

对这张表的分析结果的格式已经在上个文档中阐述过了，只需要照搬格式填入这里对应的内容即可；总之这一回归在数学上的误差“看起来”不算很大，似乎有可信度就是了

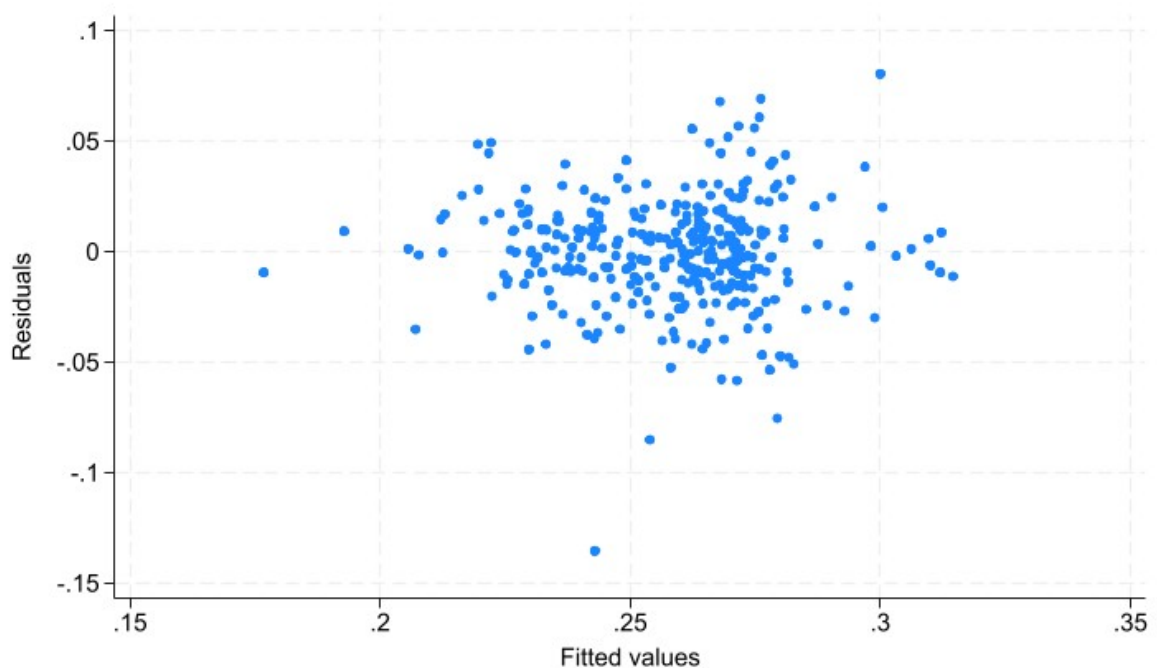
在这张表中，注意到C（对应词频，即该单词在英文文本中出现的大概频率对应的得分）的系数非常的小，因此考虑标准化回归以减小可能的误差

Source	SS	df	MS	Number of obs	=	334
Model	.139426586	15	.009295106	F(15, 318)	=	14.78
Residual	.199998595	318	.000628926	Prob > F	=	0.0000
				R-squared	=	0.4108
				Adj R-squared	=	0.3830
Total	.339425181	333	.001019295	Root MSE	=	.02508

A	Coefficient	Std. err.	t	P> t	Beta
C	7.43e-09	7.60e-09	0.98	0.329	.0441646
D	-.1114745	.0422661	-2.64	0.009	-.1475925
b1	.0035799	.0032866	1.09	0.277	.0561449
b2	0 (omitted)				0
b3	.0031904	.0076799	0.42	0.678	.0186262
e1	.0035186	.0163949	0.21	0.830	.0134033
e2	0 (omitted)				0
e3	.0021079	.0082253	0.26	0.798	.0273266
e4	.0036171	.0081776	0.44	0.659	.0558712
e5	.0081579	.0086367	0.94	0.346	.1092177
e6	.031557	.0099388	3.18	0.002	.2604955
f1	.0315308	.0071565	4.41	0.000	.4679782
f2	.0192055	.007329	2.62	0.009	.2700411
f3	0 (omitted)				0
f4	.0073149	.0145475	0.50	0.615	.0249602
g1	.0713942	.0237737	3.00	0.003	1.013542
g2	.045804	.0235793	1.94	0.053	.6374914
g3	.0353657	.0269133	1.31	0.190	.1206768
g4	0 (omitted)				0
_cons	.2854648	.0556991	5.13	0.000	.

这是标准化后的回归系数，标准化的过程是对原始数据减去均值后除以标准差得到的新变量值

这是残差与拟合值的散点图，可以看到大多数数据的残差在可控范围之内，也符合正态分布



接下来是异方差检验，采用BP检验和怀特检验，原假设为数据不存在异方差，BP检验和怀特检验的结果分别如下：

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity

Assumption: i.i.d. error terms

Variables: All independent variables

H0: Constant variance

chi2(15) = 10.55

Prob > chi2 = 0.7838

. estat imtest,white

White's test

H0: Homoskedasticity

Ha: Unrestricted heteroskedasticity

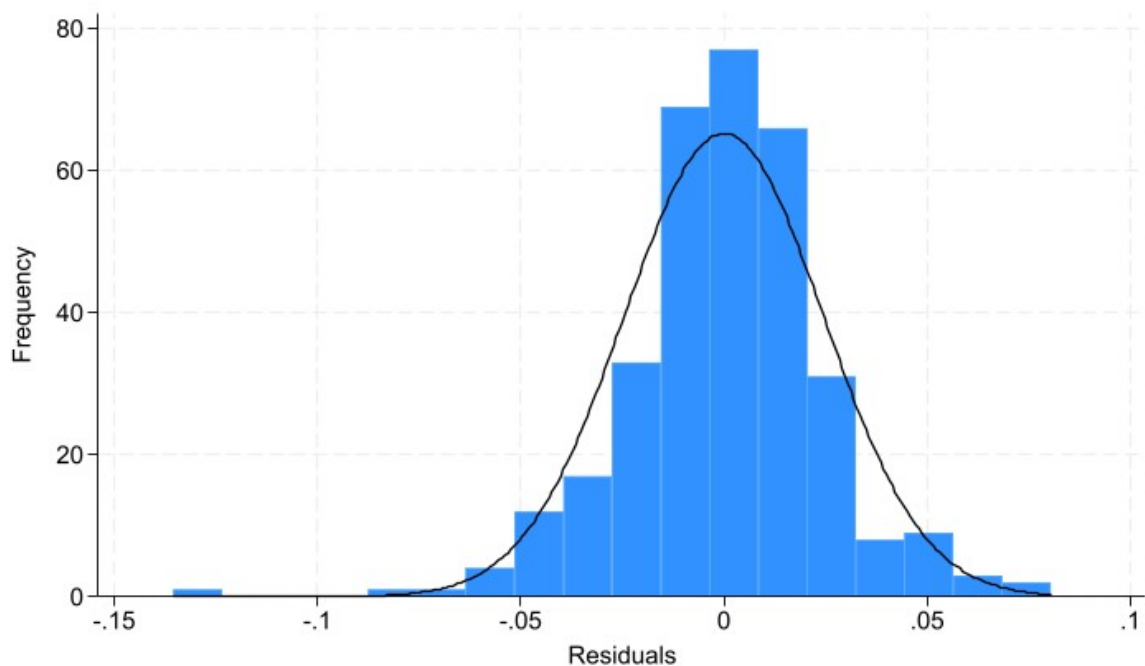
chi2(67) = 55.15

Prob > chi2 = 0.8492

Cameron & Trivedi's decomposition of IM-test

Source	chi2	df	p
Heteroskedasticity	55.15	67	0.8492
Skewness	10.47	15	0.7892
Kurtosis	1.92	1	0.1664
Total	67.53	83	0.8910

第一张为BP检验，第二张为怀特检验，可以看到二者的P值均远大于0.05，说明在95%的置信区间下数据不存在异方差，不需要使用“OLS+稳健的标准误”方法。



这是对334个数据的对猜测值的贡献这一变量预测值和实际值的残差图，可以看到残差符合正态分布，从这一角度模型是可信的

接下来考虑数据内部的多重共线性，通过方差膨胀因子（VIF，Variance Inflation Factor）评估

. estat vif

Variable	VIF	1/VIF
g1	61.47	0.016267
g2	58.12	0.017205
e4	8.61	0.116135
e5	7.22	0.138587
e3	6.14	0.162960
f1	6.09	0.164240
f2	5.73	0.174486
g3	4.55	0.219704
e6	3.63	0.275282
e1	2.10	0.475073
D	1.69	0.591688
b1	1.43	0.697405
f4	1.33	0.751962
C	1.10	0.906768
b3	1.08	0.921725
Mean VIF	11.35	

可以注意到除了g1和g2（一个单词中出现一个字母对和两个字母对）以外，其它元素都没有明显的多重共线性，而且多重共线性对多个因素下模型本身的预测能力影响也不大；因此可以忽略多重共线性

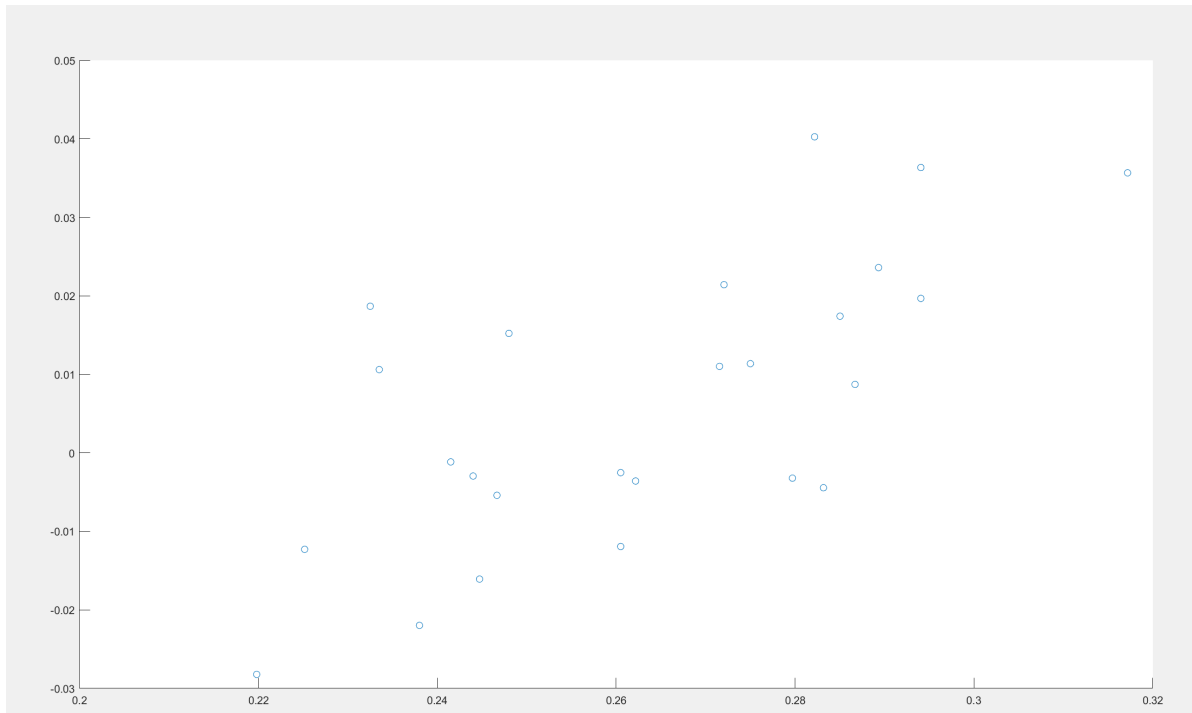
因此我们得到的多元线性回归方程为：

$$A(\text{单词猜测的贡献值}) = C \times 7.43e-09 + D \times -0.1114745 + b_1 \times 0.0035799 + b_3 \times 0.0031904 + e_1 \times 0.0035186 + e_3 \times 0.0021709 + e_4 \times 0.0036171 + e_5 \times 0.0081579 + e_6 \times 0.031557 + f_1 \times 0.0315308 + f_2 \times 0.0192055 + f_4 \times 0.0073149 + g_1 \times 0.0713942 + g_2 \times 0.045804 + g_3 \times 0.0353657 + 0.2854648$$

接下来用测试组验证训练组，得到的结果如下：

单词	实际贡献值	回归得到的贡献值	残差
light	0.271549	0.260529	0.01102
wrung	0.2415	0.242651	-0.00115
could	0.279703	0.282918	-0.00321
perky	0.244	0.246951	-0.00295
mount	0.2867	0.277979	0.008721
whack	0.2605	0.263016	-0.00252
sugar	0.272054	0.250632	0.021422
knoll	0.225167	0.237455	-0.01229
crimp	0.275	0.263632	0.011368
wince	0.244719	0.260796	-0.01608
prick	0.285017	0.267605	0.017412
robot	0.282178	0.241923	0.040255
point	0.317172	0.281502	0.03567
proxy	0.219802	0.248013	-0.02821
shire	0.283168	0.287611	-0.00444
solar	0.289333	0.265732	0.023601
panic	0.294059	0.257714	0.036345
tangy	0.248	0.232779	0.015221
abbey	0.233502	0.222893	0.010608
favor	0.238	0.259972	-0.02197
drink	0.294059	0.274382	0.019677
query	0.246667	0.252067	-0.0054
gorge	0.2325	0.213826	0.018674
crank	0.2605	0.272422	-0.01192
slump	0.262167	0.265748	-0.00358

得到的测试组真实值和模型模拟的值的残差散点图为：



可以看到测试组的残差散点图的情况还是较为不错的，总体没有出现太大的偏差，大部分数据的残差都在正负0.01以内

在该结果下，单词“EERIE”对单词猜测的贡献值的计算如下：

EERIE中， $b_1 = 1$ ， b_2 到 b_3 都为0， $e_6 = 5$ ， e_1 到 e_5 都为零， f_1 为1， f_2 到 f_4 都为零， g_4 等于1， g_1 到 g_3 都为零

（本来准备选35~40个单词做训练集，结果算错了.....）

尽管由这个“单词猜测结果的贡献值”并不能直接得到猜测1~6次的比例，但是借此可以做一个难度的参考，接下来猜测难度的矩阵打算采用BP神经网络模型