Vocational Interests

Introduction

This paper will focus on the process of regression model construction. In a Psychology study, there might be multiple response variables as well as multiple explanatory variables and the relationships among them are not immediately clear. I will address my approaches to a study of this kind.

The purpose of this study is to use regression technique to identify predictive relationships of various cognitive skills and personalities on vocational interests.

Data

Data is from http://psych.colorado.edu/~carey/Courses/PSYC7291/DataSets/Documentation/InterestDataDoc.txt. Detailed background of the study is not available. Based on the data, 250 participants were given tests that evaluate cognitive skills, personalities and vocational interests. All the variables are listed by category as follows:

General background: Gender, Education (number of years), Age

Cognitive skills: Vocabulary, Reading comprehensive (Reading), Sentence completion (Sentcomp), Mathematics, Geometry, Analytical reasoning (Analyrea)

Personalities: Social dominance (Socdom), Sociability, Stress reaction (Stress), Worry scale (Worry), Impulsivity, Thrill-seeking (Thrillsk)

Vocational interests: Carpentry, Forest Ranger, Mortician, Police, Fireman, Sales Representative, Teacher, Business Executive, Stock Broker, Artist, Social Worker, Truck Driver, Doctor, Clergyman, Lawyer, Actor, Architect, Landscaper

Analysis

Descriptive Statistics

Descriptive statistics for all test scores are summarized in Appendix A. All means are close to 0 and all standard errors are close to 1, which indicate the test scores are normalized results. Skewness and kurtosis are close to 0 for the majority of the tests too. Therefore normal distribution is assumed for all tests and no transformation is needed for regressions.

Dummy Variable

The only categorical variable in the study is gender. In the original dataset, male is coded 1 and female 2. To make the analysis results easier to interpret, they are recoded as 0 and 1.

Correlations

In order to narrow down the explanatory variables for each vocational interest, correlations of each of the vocational interests with gender, education age, cognitive skills and personalities are calculated and included in Appendix B.

Strong correlation does not always suggest predictive relationship, but weak correlation always confirms weak or no predictive relationship. So for each vocational interest, explanatory variables are narrowed down to the cognitive skills and personalities whose correlations with the vocational interest exceed certain threshold. I select the threshold to be correlation coefficient of 0.15 considering strong correlations are not common in Psychology studies. Selected explanatory variables are highlighted in the grid in Appendix B.

Correlations among explanatory variables are also calculated and included in Appendix C. Strong correlations among explanatory variables can cause collinearity issue and reduce the efficiency of regression models.

Regressions

The correlation grid suggests vocational interests of Carpentry, Forest Ranger and Landscaper cannot be explained by the tested cognitive skills and personalities. For each of the other vocational interests, linear least-square regression is fit on all the highlighted explanatory variables first. If gender is one of the explanatory variables, interactions of gender with all other explanatory variables are also included in the initial regression model. Based on the hypothesis test results, modifications to the regression model are made and conclusion is drawn for each vocational interest.

As strong correlations are rare in Psychology study, strong regressions are even rarer. For this study, I consider a regression model effective if R Square is at least 10%. For significance tests, 95% confidence level is considered.

Mortician

Initial model results:

Regression Statistics	
Multiple R	0.298001
R Square	0.088804
Adjusted R Square	0.081426
Standard Error	1.038165
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	25.9450	12.9725	12.0362	0.0000
Residual	247	266.2133	1.0778		
Total	249	292.1582			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-1.1269	0.2673	-4.2157	0.0000
age	0.0268	0.0066	4.0950	0.0001
worry	0.1792	0.0654	2.7409	0.0066

Even though both F test and t tests are significant at 95% confidence level, R Square is less than 10%. Age and worry scale do not effectively explain the variation in vocational interest in Mortician.

<u>Police</u>

Initial model results:

Regression Statistics	
Multiple R	0.433804
R Square	0.188186
Adjusted R Square	0.150665
Standard Error	0.88128
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	42.8484	3.8953	5.0155	0.0000
Residual	238	184.8436	0.7767		
Total	249	227.6920			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1.1535	0.3274	3.5235	0.0005
Gender	-1.0229	0.4801	-2.1306	**0.0341**
Age	-0.0276	0.0081	-3.4121	**0.0008**
Socdom	-0.0964	0.0940	-1.0256	0.3061
Sociability	-0.1984	0.0957	-2.0733	**0.0392**
Impulsivity	0.1079	0.0891	1.2111	0.2270
Thrillsk	0.0098	0.0891	0.1096	0.9128
gender X age	0.0174	0.0117	1.4896	0.1377
gender X socdom	0.1041	0.1399	0.7442	0.4575
gender X sociability	-0.0373	0.1366	-0.2731	0.7850
gender X impulsivity	0.0924	0.1344	0.6876	0.4924
gender X thrillsk	0.0896	0.1270	0.7051	0.4814

None of the interactions is significant and only coefficients of gender, age and sociability are significant at 95% confidence level. However the effect of some of the explanatory variables might be obscured in this model due to all the interactions included. The second model removes all the interactions and results below indicate in addition to gender, age and sociability, coefficient of impulsivity becomes significant as well.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.8259	0.2404	3.4348	0.0007
Gender	-0.3098	0.1123	-2.7574	**0.0063**
age	-0.0194	0.0058	-3.3525	**0.0009**
socdom	-0.0608	0.0690	-0.8817	0.3788
sociability	-0.1996	0.0673	-2.9674	**0.0033**
impulsivity	0.1566	0.0660	2.3747	**0.0183**
thrillsk	0.0578	0.0625	0.9256	0.3556

Final model on gender, age, sociability and impulsivity:

Regression Statistics	
Multiple R	0.411979
R Square	0.169726
Adjusted R Square	0.156171
Standard Error	0.878418
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	38.6453	9.6613	12.5209	0.0000
Residual	245	189.0467	0.7716		
Total	249	227.6920			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.8236	0.2379	3.4620	0.0006
Gender	-0.3155	0.1118	-2.8206	0.0052
age	-0.0192	0.0057	-3.3508	0.0009
sociability	-0.2349	0.0554	-4.2359	0.0000
impulsivity	0.1873	0.0575	3.2589	0.0013

R square is greater than 10%, F test and all the t-tests are significant at 0.95% confidence level.

Fireman

Initial model results:

Regression Statistics	
Multiple R	0.387258
R Square	0.149969
Adjusted R Square	0.125381
Standard Error	0.887199
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	33.6065	4.8009	6.0993	0.0000
Residual	242	190.4835	0.7871		
Total	249	224.0900			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1.0907	0.3277	3.3283	0.0010
Gender	-1.0466	0.4744	-2.2061	**0.0283**
age	-0.0226	0.0081	-2.7897	**0.0057**
sociability	-0.2407	0.0822	-2.9295	**0.0037**
impulsivity	0.1283	0.0827	1.5516	0.1221
gender X age	0.0152	0.0116	1.3099	0.1915
gender X sociability	0.1489	0.1131	1.3160	0.1894
gender X impulsivity	0.0265	0.1170	0.2269	0.8207

Similar to Police, none of the interactions is significant. Remove the interactions, the model results are shown below.

Regression Statistics	
Multiple R	0.373542
R Square	0.139534
Adjusted R Square	0.125485
Standard Error	0.887146
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	31.2681	7.8170	9.9323	0.0000
Residual	245	192.8219	0.7870		
Total	249	224.0900			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.7990	0.2403	3.3254	0.0010
Gender	-0.4367	0.1130	-3.8664	0.0001
age	-0.0151	0.0058	-2.6145	0.0095
sociability	-0.1622	0.0560	-2.8966	0.0041
impulsivity	0.1504	0.0580	2.5910	0.0101

R square is greater than 10%, F test and all the t-tests are significant at 0.95% confidence level.

Sales Representative

Initial model results:

Regression Statistics	
Multiple R	0.242229222
R Square	0.058674996
Adjusted R Square	0.043306425
Standard Error	0.990888856
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	14.9944	3.7486	3.8179	0.0050
Residual	245	240.5559	0.9819		
Total	249	255.5503			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.7157	0.5547	-1.2903	0.1982
education	0.0500	0.0452	1.1070	0.2694
mathematics	-0.0785	0.1196	-0.6559	0.5125
geometry	0.0472	0.0986	0.4789	0.6324
analyrea	0.2161	0.1063	2.0342	0.0430

The low R Square suggests very little variation in vocational interest in Sales Representative can be explained by these explanatory variables.

Teacher

Initial model results:

Regression Statistics	
Multiple R	0.49259
R Square	0.242645
Adjusted R Square	0.200926
Standard Error	0.91773
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	13	63.6817	4.8986	5.8162	0.0000
Residual	236	198.7661	0.8422		
Total	249	262.4478			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-1.8366	0.7960	-2.3072	0.0219
Gender	1.7629	1.0781	1.6353	0.1033
Education	0.1255	0.0647	1.9405	0.0535
Vocabulary	0.1707	0.1740	0.9810	0.3276
Reading	0.3589	0.1450	2.4753	**0.0140**
Sentcomp	-0.2852	0.1379	-2.0684	**0.0397**
Socdom	0.1666	0.1008	1.6521	0.0998
Sociability	-0.0166	0.1007	-0.1651	0.8690
gender X education	-0.1063	0.0877	-1.2120	0.2267
gender X vocab	-0.2137	0.2452	-0.8715	0.3843
gender X reading	-0.2132	0.2074	-1.0277	0.3051
gender X sentcomp	0.4510	0.2094	2.1535	0.0323
gender X socdom	0.0429	0.1466	0.2924	0.7703
gender X sociability	0.1243	0.1423	0.8735	0.3833

The results from the initial model suggest there might be interaction between gender and sentence completion, however if I remove all interactions except gender x sentcomp as below, the interaction is not significant any more.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.8619	0.5330	-1.6171	0.1072
Gender	0.4479	0.1186	3.7764	**0.0002**
Education	0.0469	0.0431	1.0890	0.2772
Vocab	0.0621	0.1231	0.5044	0.6144
Reading	0.2525	0.1026	2.4604	**0.0146**
Sentcomp	-0.1014	0.1139	-0.8907	0.3740
Socdom	0.1805	0.0725	2.4887	**0.0135**
Sociability	0.0720	0.0707	1.0180	0.3097
gender X sentcomp	0.0692	0.1210	0.5721	0.5678

The third model I looked at is one on all explanatory variables but no interactions. Coefficients for gender, reading and social dominance are significant at 95% confidence level.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.9022	0.5275	-1.7102	0.0885
Gender	0.4536	0.1180	3.8439	**0.0002**
Education	0.0502	0.0426	1.1778	0.2400
Vocabulary	0.0704	0.1221	0.5763	0.5650
Reading	0.2493	0.1023	2.4362	**0.0156**
Sentcomp	-0.0748	0.1038	-0.7207	0.4718
Socdom	0.1808	0.0724	2.4975	**0.0132**
Sociability	0.0706	0.0706	1.0010	0.3178

Final regression on gender, reading and social dominance:

Regression Statistics	
Multiple R	0.453224
R Square	0.205412
Adjusted R Square	0.195722
Standard Error	0.920714
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	53.9099	17.9700	21.1981	0.0000
Residual	246	208.5378	0.8477		
Total	249	262.4478			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.2893	0.0814	-3.5525	0.0005
Gender	0.4517	0.1177	3.8382	0.0002
Reading	0.2950	0.0591	4.9880	0.0000
Socdom	0.2274	0.0586	3.8834	0.0001

Business Executive

Initial model results:

Regression Statistics	
Multiple R	0.264984
R Square	0.070217
Adjusted R Square	0.062688
Standard Error	0.963668
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	17.3225	8.6613	9.3267	0.0001
Residual	247	229.3780	0.9287		
Total	249	246.7005			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.7965	0.2532	-3.1462	0.0019
age	0.0207	0.0062	3.3543	0.0009
socdom	0.2029	0.0619	3.2803	0.0012

Even though F test and t tests are significant at 95% confidence level, R Square is less than 10%, very little of the variation in vocational interest in Business Executive can be explained by the model.

Stock Broker

Initial model results:

Regression Statistics	
Multiple R	0.33892
R Square	0.114867
Adjusted R Square	0.100415
Standard Error	0.985076
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	30.8525	7.7131	7.9486	0.0000
Residual	245	237.7418	0.9704		
Total	249	268.5943			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-2.5529	0.5767	-4.4270	0.0000
Education	0.1473	0.0393	3.7432	**0.0002**
Age	0.0189	0.0063	2.9900	**0.0031**
Stress	0.1358	0.0764	1.7771	0.0768
Worry	0.1204	0.0706	1.7048	0.0895

The only significant explanatory variables are education and age. The second model is on education and age:

Regression Statistics	
Multiple R	0.270135
R Square	0.072973
Adjusted R Square	0.065467
Standard Error	1.004029
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	19.6001	9.8000	9.7216	0.0001
Residual	247	248.9942	1.0081		
Total	249	268.5943			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-2.5240	0.5869	-4.3006	0.0000
Education	0.1418	0.0398	3.5587	0.0004
Age	0.0198	0.0064	3.0924	0.0022

After removing both stress reaction and worry scale from the model, R Square dropped to below the 10% threshold. Notice the high correlation between stress reaction and worry scale (r = 0.47) might have caused the collinearity issue. Therefore the coefficient for either variable is significant in the initial model. Removing either would make the model more efficient than the initial model and more effective than the second.

Final model on education, age and stress:

Regression Statistics	
Multiple R	0.323058
R Square	0.104366
Adjusted R Square	0.093444
Standard Error	0.988886
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	28.0322	9.3441	9.5553	0.0000
Residual	246	240.5621	0.9779		
Total	249	268.5943			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-2.5902	0.5785	-4.4776	0.0000
Education	0.1522	0.0394	3.8624	0.0001
Age	0.0183	0.0063	2.8971	0.0041
Stress	0.1979	0.0674	2.9364	0.0036

Artist

Initial model results:

Regression Statistics	
Multiple R	0.173497
R Square	0.030101
Adjusted R Square	0.022248
Standard Error	1.008953
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	7.8036	3.9018	3.8329	0.0229
Residual	247	251.4428	1.0180		
Total	249	259.2465			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.0234	0.0641	-0.3660	0.7147
Vocabulary	0.1514	0.1101	1.3752	0.1703
Sentcomp	0.0310	0.1108	0.2798	0.7798

All statistics suggest the regression model can hardly explain vocational interest in artist.

Truck Driver

Initial model results:

Regression Statistics	
Multiple R	0.350425
R Square	0.122798
Adjusted R Square	0.097424
Standard Error	0.938415
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	29.8329	4.2618	4.8396	0.0000
Residual	242	213.1108	0.8806		
Total	249	242.9437			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.3069	0.0832	3.6866	0.0003
Gender	-0.4546	0.1205	-3.7724	**0.0002**
Vocabulary	-0.1577	0.1732	-0.9109	0.3633
Reading	-0.1467	0.1437	-1.0207	0.3084
Sentcomp	0.1472	0.1400	1.0513	0.2942
Gender X vocab	-0.0198	0.2445	-0.0809	0.9356
Gender X reading	0.0710	0.2065	0.3437	0.7314
Gender X sentcomp	-0.1922	0.2133	-0.9012	0.3684

Remove interaction terms since none is significant.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.3051	0.0829	3.6810	0.0003
Gender	-0.4618	0.1189	-3.8828	**0.0001**
Vocabulary	-0.1900	0.1207	-1.5748	0.1166
Reading	-0.1069	0.1028	-1.0390	0.2998
Sentcomp	0.0691	0.1052	0.6568	0.5119

The only significant coefficient is gender. The results for the regression model on gender only:

Regression Statistics	
Multiple R	0.250513
R Square	0.062757
Adjusted R Square	0.058978
Standard Error	0.958193
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	15.2464	15.2464	16.6058	0.0001
Residual	248	227.6973	0.9181		
Total	249	242.9437			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.2944	0.0847	3.4758	0.0006
Gender	-0.4940	0.1212	-4.0750	0.0001

R Square for the model on gender only is significantly lower than that for the initial model suggesting some effective explanatory variable has been removed. There is extremely high correlation between vocabulary and reading (r = 0.8), which might have caused the collinearity and resulted in non-significant coefficients in the initial model. Adding vocabulary back in the model, the results are shown below.

Regression Statistics	
Multiple R	0.334139
R Square	0.111649
Adjusted R Square	0.104456
Standard Error	0.934753
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27.1244	13.5622	15.5216	0.0000
Residual	247	215.8193	0.8738		
Total	249	242.9437			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.2998	0.0826	3.6277	0.0003
Gender	-0.4646	0.1185	-3.9192	0.0001
Vocabulary	-0.2193	0.0595	-3.6870	0.0003

The final model is more efficient than the initial model and more effective than the one on gender only.

Doctor

Initial model results:

Regression Statistics	
Multiple R	0.485311
R Square	0.235527
Adjusted R Square	0.206859
Standard Error	0.937654
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	65.0093	7.2233	8.2158	0.0000
Residual	240	211.0070	0.8792		
Total	249	276.0163			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-2.7018	0.5435	-4.9709	0.0000
Education	0.2165	0.0443	4.8892	**0.0000**
Vocabulary	-0.0381	0.1301	-0.2927	0.7700
Reading	0.0509	0.1072	0.4749	0.6353
Sentcomp	0.2038	0.1068	1.9094	0.0574
Mathematics	0.1768	0.1199	1.4741	0.1418
Geometry	-0.1792	0.0963	-1.8609	0.0640
Analyrea	-0.0852	0.1041	-0.8180	0.4142
Socdom	0.1174	0.0741	1.5837	0.1146
Sociability	0.0973	0.0718	1.3562	0.1763

The only explanatory variable for which the coefficient is significant at 95% confidence level is education.

Results for the model on education:

Regression Statistics	
Multiple R	0.413637
R Square	0.171096
Adjusted R Square	0.167753
Standard Error	0.960492
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	47.2252	47.2252	51.1901	0.0000
Residual	248	228.7911	0.9225		
Total	249	276.0163			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-3.3382	0.4679	-7.1345	0.0000
Education	0.2698	0.0377	7.1547	0.0000

The results suggest education alone can explain most of the variation that all the available explanatory variables can explain. Adding any of the other variable does not improve R square significantly. So the model on education alone is effective and efficient.

Clergyman

Initial model results:

Regression Statistics	
Multiple R	0.178722
R Square	0.031941
Adjusted R Square	0.028038
Standard Error	0.951617
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7.4102	7.4102	8.1828	0.0046
Residual	248	224.5827	0.9056		
Total	249	231.9929			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.0332	0.0605	-0.5487	0.5837
socdom	0.1719	0.0601	2.8606	0.0046

All statistics suggest the regression model can hardly explain vocational interest in clergyman.

Lawyer

Initial model results:

Regression Statistics	
Multiple R	0.32393
R Square	0.104931
Adjusted R Square	0.086589
Standard Error	0.944725
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	25.530	5.106	5.721	0.000
Residual	244	217.771	0.893		
Total	249	243.301			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.0180	0.0607	-0.2961	0.7674
Vocabulary	0.2089	0.1226	1.7035	0.0898
Reading	-0.0737	0.1043	-0.7068	0.4804
Sentcomp	0.1112	0.1061	1.0484	0.2955
Socdom	0.0635	0.0739	0.8584	0.3915
Sociability	0.1211	0.0722	1.6771	0.0948

R square barely meets the 10% threshold. Since none of the coefficient was significant at 95% confidence level, I looked at the model on vocabulary and sociability, the two variables with the lowest p-values and get the following results:

Regression Statistics	
Multiple R	0.311159
R Square	0.09682
Adjusted R Square	0.089507
Standard Error	0.943215
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	23.5564	11.7782	13.2390	0.0000
Residual	247	219.7448	0.8897		
Total	249	243.3012			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.0185	0.0600	-0.3091	0.7575
vocabulary	0.2427	0.0604	4.0213	0.0001
sociability	0.1577	0.0589	2.6798	0.0079

Now the coefficients for the explanatory variables are significant, but R square is slightly below the 10% threshold. The model does not do a very good job explaining the variation in vocational interest in lawyer, but I consider it acceptable.

Actor

Initial model results:

Regression Statistics	
Multiple R	0.514525
R Square	0.264736
Adjusted R Square	0.237163
Standard Error	0.907698
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	71.1972	7.9108	9.6015	0.0000
Residual	240	197.7399	0.8239		
Total	249	268.9372			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-1.8806	0.5262	-3.5743	0.0004
Education	0.1472	0.0429	3.4357	**0.0007**
Vocabulary	0.2246	0.1260	1.7828	0.0759
Reading	0.1133	0.1037	1.0921	0.2759
Sentcomp	-0.0395	0.1033	-0.3818	0.7029
Mathematics	-0.0227	0.1161	-0.1952	0.8454
Geometry	-0.0744	0.0932	-0.7982	0.4255
Analyrea	0.1046	0.1008	1.0380	0.3003
Socdom	0.1170	0.0718	1.6305	0.1043
Sociability	0.0995	0.0695	1.4319	0.1535

The only coefficient that is significant at 95% confidence level is for education. The results of the model on education are shown below:

Regression Statistics	
Multiple R	0.394289
R Square	0.155464
Adjusted R Square	0.152059
Standard Error	0.956993
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	41.8100	41.8100	45.6523	0.0000
Residual	248	227.1271	0.9158		
Total	249	268.9372			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-3.1342	0.4662	-6.7230	0.0000
Education	0.2538	0.0376	6.7567	0.0000

Another regression model on education and vocabulary is tested and the results are shown below.

Regression Statistics	
Multiple R	0.470713
R Square	0.221571
Adjusted R Square	0.215268
Standard Error	0.920633
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	59.5887	29.7943	35.1529	0.0000
Residual	247	209.3485	0.8476		
Total	249	268.9372			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-1.9284	0.5200	-3.7082	0.0003
Education	0.1535	0.0423	3.6336	0.0003
Vocabulary	0.3130	0.0683	4.5800	0.0000

The coefficient of vocabulary is significant in this model even though it is not in the initial one. Compared to the second model, R square improves significantly suggesting the additional variable is effective and efficient in explaining the variation in the vocational interest in actor.

Architect

Initial model results:

Regression Statistics	
Multiple R	0.387622
R Square	0.150251
Adjusted R Square	0.122043
Standard Error	0.944874
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	38.0444	4.7556	5.3266	0.0000
Residual	241	215.1616	0.8928		
Total	249	253.2060			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-1.8342	0.5476	-3.3496	0.0009
Education	0.1557	0.0446	3.4910	**0.0006**
Vocabulary	0.0651	0.1311	0.4967	0.6198
Reading	-0.0179	0.1079	-0.1663	0.8680
Sentcomp	0.1144	0.1076	1.0635	0.2886
Mathematics	-0.0220	0.1208	-0.1822	0.8555
Geometry	-0.0197	0.0970	-0.2027	0.8395
Analyrea	0.0351	0.1049	0.3348	0.7381
Socdom	0.1389	0.0616	2.2566	**0.0249**

R Square exceeds the 10% threshold. The coefficients of education and social dominance are significant at 95% confidence level.

Regression Statistics	
Multiple R	0.363668
R Square	0.132254
Adjusted R Square	0.125228
Standard Error	0.943159
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	33.4876	16.7438	18.8228	0.0000
Residual	247	219.7184	0.8895		
Total	249	253.2060			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-2.3829	0.4598	-5.1822	0.0000
Education	0.2013	0.0371	5.4286	0.0000
Socdom	0.1528	0.0596	2.5624	0.0110

The model on education and social dominance explains most of the variation that all the available explanatory variables can explain. It is effective and efficient.

Conclusion

Not all vocational interests can be well explained by age, gender, education and the tested skills and personalities. The only effective models I identified are the following:

Interest in Police = 0.8236 – 0.3155 x female – 0.0192 x age – 0.2349 x sociability + 0.1873 x impulsivity

Interest in Fireman = 0.7990 – 0.4367 x female – 0.0151 x age – 0.1622 x sociability + 0.1504 x impulsivity

Interest in Teacher = -0.2893 +0.4517 x female + 0.2950 x reading + 0.2274 x socdom

Interest in Stock Broker = -2.5902 + 0.1522 x education + 0.0183 x age + 0.1979 x stress

Interest in Truck Driver = 0.2998 – 0.4646 x female - 0.2193 x vocabulary

Interest in Doctor = -3.3382 + 0.2698 x education

Interest in Lawyer = -0.0185 + 0.2427 x vocabulary + 0.1577 x sociability

Interest in Actor = -1.9284 + 0.1535 x education + 0.3130 x vocabulary

Interest in Architect = -2.3829 + 0.2013 x education + 0.1528 x socdom

The processes of determine the optimal models can be different depending on the situation. Sometimes the process is straightforward, other times it requires judgments and can involve several iterations before reaching the final conclusion.

Appendix A: Descriptive Statistics

	vocab	*reading*	*sentcomp*	*mathmtsc*	*geometry*	*analyrea*	*socdom*	*sociability*	*stress*	*worry*	*impulsivity*	*thrillsk*
Mean	0.09	0.13	0.07	0.11	0.11	0.18	0.10	0.06	-0.02	0.00	0.05	0.14
Standard Error	0.06	0.06	0.06	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.06	0.07
Median	0.04	0.19	0.11	0.10	0.09	0.20	0.07	0.04	0.01	-0.03	0.13	0.15
Mode	0.33	0.54	0.80	1.05	0.16	0.58	-0.16	1.21	0.23	0.40	0.49	-0.30
Standard Deviation	1.00	0.99	0.99	1.05	1.03	1.06	1.00	1.02	0.94	1.01	0.99	1.04
Sample Variance	1.00	0.98	0.98	1.11	1.07	1.12	1.01	1.05	0.88	1.01	0.98	1.07
Kurtosis	-0.37	-0.11	-0.15	0.30	0.60	0.22	0.05	-0.08	-0.03	0.00	0.25	-0.04
Skewness	-0.02	-0.11	-0.01	-0.36	0.09	-0.18	0.05	0.05	-0.04	-0.05	-0.16	0.07
Range	5.25	5.17	5.20	6.77	7.18	6.33	5.62	5.85	5.34	5.49	5.66	5.96
Minimum	-2.62	-2.47	-2.47	-3.71	-3.32	-2.83	-2.72	-3.09	-2.56	-3.07	-3.04	-2.96
Maximum	2.63	2.70	2.73	3.06	3.86	3.50	2.90	2.76	2.78	2.42	2.62	3.00
Sum	22.54	33.74	18.39	26.37	28.13	43.75	25.30	14.92	-4.24	0.56	13.12	35.24
Count	250	250	250	250	250	250	250	250	250	250	250	250

Appendix A: Descriptive Statistics

	Carpentry	*Forest_Ranger*	*Mortician*	*Police*	*Fireman*	*Sales_Representative*	*Teacher*	*Business_Executive*	*Stock_Broker*
Mean	0.02	0.09	-0.07	-0.09	-0.01	-0.07	-0.01	0.04	0.00
Standard Error	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.06	0.07
Median	0.05	0.11	-0.08	-0.08	0.00	-0.05	0.08	0.05	-0.02
Mode	-0.15	1.10	-0.26	0.22	0.43	-0.42	0.36	0.24	-0.31
Standard Deviation	1.03	1.06	1.08	0.96	0.95	1.01	1.03	1.00	1.04
Sample Variance	1.07	1.11	1.17	0.91	0.90	1.03	1.05	0.99	1.08
Kurtosis	0.45	-0.10	-0.31	0.00	0.16	0.23	0.83	-0.03	-0.41
Skewness	-0.21	0.09	0.08	0.12	0.07	-0.15	-0.30	0.05	0.22
Range	6.13	5.47	5.94	5.12	5.66	6.16	6.90	5.40	5.45
Minimum	-3.49	-2.49	-2.94	-2.42	-2.86	-3.62	-4.42	-2.51	-2.33
Maximum	2.64	2.98	3.00	2.70	2.80	2.54	2.48	2.89	3.12
Sum	5.16	21.25	-16.35	-23.62	-3.70	-16.45	-1.52	11.05	0.88
Count	250	250	250	250	250	250	250	250	250

Appendix A: Descriptive Statistics

	Artist	*Social_Worker*	*Truck_Driver*	*Doctor*	*Clergyman*	*Lawyer*	*Actor*	*Architect*	*Landscaper*
Mean	-0.01	0.04	0.05	-0.02	-0.02	0.01	-0.01	0.11	0.07
Standard Error	0.06	0.06	0.06	0.07	0.06	0.06	0.07	0.06	0.07
Median	0.07	0.03	0.06	0.00	0.01	0.01	0.03	0.05	0.05
Mode	0.07	-0.35	0.11	0.35	-0.28	0.09	-0.23	0.00	0.22
Standard Deviation	1.02	1.00	0.99	1.05	0.97	0.99	1.04	1.01	1.04
Sample Variance	1.04	1.00	0.98	1.11	0.93	0.98	1.08	1.02	1.08
Kurtosis	-0.14	0.38	-0.16	-0.15	-0.17	0.48	0.11	-0.34	0.02
Skewness	-0.21	-0.06	-0.10	0.20	-0.11	0.21	-0.25	0.15	0.03
Range	5.39	6.20	5.49	5.59	4.93	5.91	5.86	4.69	5.61
Minimum	-2.56	-2.97	-2.78	-2.58	-2.76	-2.75	-3.30	-2.11	-2.82
Maximum	2.83	3.23	2.71	3.01	2.17	3.16	2.56	2.58	2.79
Sum	-1.88	8.86	13.32	-4.72	-3.95	3.19	-2.74	27.30	17.38
Count	250	250	250	250	250	250	250	250	250

Appendix B: Correlations between explanatory and response variables

	Female	*edu*	*age*	*vocab*	*reading*	*sentcomp*	*mathmtsc*	*geometry*	*analyrea*	*socdom*	*sociability*	*stress*	*worry*	*impulsivity*	*thrillsk*
Carpentry	-0.069	-0.005	0.010	-0.094	-0.125	-0.065	0.049	0.003	0.035	-0.027	0.003	0.045	0.014	0.068	0.065
Forest_Ranger	-0.051	0.028	-0.095	-0.023	-0.051	-0.071	-0.025	-0.031	-0.009	0.027	-0.027	-0.067	-0.108	0.089	0.024
Mortician	0.115	0.134	***0.247***	0.003	0.020	0.023	-0.004	-0.063	0.006	0.118	0.068	0.120	***0.164***	-0.063	0.003
Police	***-0.191***	-0.141	***-0.204***	-0.146	-0.146	-0.144	-0.024	-0.035	-0.054	***-0.164***	***-0.221***	0.025	0.002	***0.213***	***0.168***
Fireman	***-0.249***	-0.073	***-0.170***	-0.061	-0.108	-0.039	0.023	0.002	-0.033	-0.110	***-0.160***	-0.058	-0.096	***0.171***	0.137
Sales_Representative	-0.047	***0.162***	0.094	0.131	0.111	0.138	***0.180***	***0.179***	***0.230***	0.102	0.114	0.122	0.123	0.058	0.037
Teacher	***0.270***	***0.217***	0.078	***0.288***	***0.311***	***0.234***	0.070	0.075	0.120	***0.258***	***0.223***	0.059	0.073	-0.080	-0.062
Business_Executive	0.052	0.108	***0.172***	0.064	0.064	0.091	0.030	-0.037	0.081	***0.167***	0.086	0.076	0.057	-0.041	-0.037
Stock_Broker	0.084	***0.193***	***0.160***	0.114	0.126	0.132	0.096	0.042	0.141	0.112	0.083	***0.171***	***0.178***	-0.044	0.013
Artist	0.007	0.074	0.060	***0.173***	0.119	***0.151***	-0.010	0.010	0.016	0.073	0.047	0.090	-0.015	-0.089	-0.045
Social_Worker	***0.358***	***0.213***	***0.165***	***0.349***	***0.353***	***0.295***	0.032	0.026	0.071	***0.204***	***0.235***	0.072	0.087	-0.122	-0.056
Truck_Driver	***-0.251***	-0.071	-0.078	***-0.237***	***-0.231***	***-0.184***	-0.094	-0.085	-0.125	-0.097	-0.060	-0.144	-0.107	0.010	-0.040
Doctor	0.022	***0.414***	0.113	***0.312***	***0.296***	***0.324***	***0.269***	***0.176***	***0.233***	***0.172***	***0.182***	0.072	0.103	-0.103	-0.084
Clergyman	0.064	0.037	0.120	0.106	0.043	0.023	-0.124	-0.095	-0.070	***0.179***	0.109	0.052	0.026	0.003	-0.054
Lawyer	0.010	0.138	0.064	***0.266***	***0.189***	***0.247***	-0.037	-0.031	-0.056	***0.169***	***0.194***	0.063	0.018	0.017	-0.021
Actor	0.029	***0.394***	0.091	***0.424***	***0.387***	***0.349***	***0.329***	***0.277***	***0.353***	***0.210***	***0.216***	0.020	0.126	-0.113	-0.012
Architect	0.044	***0.330***	0.114	***0.283***	***0.235***	***0.275***	***0.222***	***0.205***	***0.234***	***0.169***	0.090	0.086	0.068	-0.024	-0.041
Landscaper	-0.044	-0.034	-0.095	-0.048	-0.056	-0.045	-0.051	-0.007	-0.024	0.066	0.060	-0.069	-0.099	0.104	0.010

Appendix C: Correlations between skill tests

	vocabulary	*reading*	*sentcomp*	*mathmtsc*	*geometry*	*analyrea*	*socdom*	*sociability*	*stress*	*worry*	*impulsivity*	*thrillsk*
vocabulary	1.000											
reading	0.803	1.000										
sentcomp	0.813	0.725	1.000									
mathmtsc	0.708	0.660	0.618	1.000								
geometry	0.633	0.526	0.575	0.774	1.000							
analyrea	0.673	0.636	0.618	0.817	0.715	1.000						
socdom	0.116	0.035	0.088	0.015	0.122	0.104	1.000					
sociability	0.126	0.085	0.097	0.057	0.127	0.119	0.583	1.000				
stress	-0.026	-0.031	-0.024	-0.064	0.014	-0.040	-0.002	-0.093	1.000			
worry	-0.057	-0.042	-0.023	0.004	0.057	0.012	-0.010	-0.043	0.470	1.000		
impulsivity	-0.177	-0.244	-0.204	-0.155	-0.130	-0.084	0.044	0.064	-0.043	-0.092	1.000	
thrillsk	-0.038	-0.099	-0.015	-0.029	-0.043	0.024	-0.005	0.014	-0.064	-0.123	0.505	1.000