

What's in a Name?

David Griffiths

The Research & Analysis Consultancy

Journal of Targeting Measurement and Analysis for Marketing Vol3 No3 1995

ABSTRACT

Age is often an important determinant affecting whether someone will, or is even eligible to, respond to a direct mailing campaign. All too often a person's age is not known and consequently this information is not available for targeting or segmentation purposes. The naming of new born children is subject to strong, time dependent factors which cause large variations in the frequency of name usage. These variations have been used to predict age. From a large and reasonably representative national sample of the population for whom both true age and forename are known, the relationship between true and predicted age is examined to show that forenames can be used to predict age with a surprising degree of accuracy.

INTRODUCTION

The element of fashion in the naming of new born children has been known for some time and has been widely reported upon. The following examples have been taken from one of the leading sources on the frequency of use of first names¹. The figures show a rate per 10,000 births of a particular name by gender. For example, 166 females per 10,000 females born in 1900 were named Ada.

Table 1	Gender	Year of Birth							
		1900	1925	1935	1955	1965	1975	1985	1990
Ada	F	166	26	18	-	2	-	-	-
Arthur	M	367	224	118	18	14	-	-	2
Alice	F	361	80	44	8	6	2	14	62
Charles	M	390	206	98	56	26	17	20	76
Daniel	M	6	10	10	16	44	286	370	476
Rebecca	F	12	4	-	10	34	174	238	390

The table shows that names like Ada and Arthur were very popular at the turn of the century but their popularity has been steadily declining. Other names like Alice and Charles have also shown a marked decline in popularity but the 1990 birth figures show that these names may be coming back into fashion. On the other hand, names like Daniel and Rebecca have become popular quite recently.

However, if this birth data is to be useful in predicting age the question is how do very common names change over time? If they are stable with little variation then the effectiveness of a first name age predictor will be severely limited. Using the same source of data (op cit) the following commonly used names were extracted.

Table 2

	Gender	Year of Birth							
		1900	1925	1935	1950	1960	1970	1980	1990
Carol	F	-	-	10	244	204	25	6	-
Deborah	F	1	1	1	20	279	170	58	12
Dorothy	F	246	350	166	48	7	-	4	-
Elizabeth	F	296	128	86	124	105	84	46	82
Janet	F	12	18	82	262	157	30	8	-
Mary	F	391	408	272	132	52	25	10	10
Susan	F	19	8	14	654	692	102	30	4
Andrew	M	13	16	8	124	338	428	312	206
David	M	52	106	270	832	625	386	340	154
John	M	726	728	750	646	305	168	104	82
Mark	M	6	6	4	12	398	467	272	96
Michael	M	12	36	280	460	421	240	262	212
Thomas	M	452	254	150	82	39	44	120	362
William	M	897	590	274	178	127	33	60	136

This shows quite marked variation in name usage over time even for popular names. Some of these names were particularly popular prior to the second world war, other names were most popular during the 50's and 60's, whilst for others the pattern to the frequency of usage is more complex. The important point to stress here is that if names are to be used to predict age then these names must show marked variations in usage over time. In fact of the 50 most widely used male and female names in current common usage, all meet this criteria. This is not to suggest that all names vary in popularity over time but that a sufficiently high proportion do and this allows age predictors to be calculated with a reasonable prospect that they will provide an accurate estimate of a person's true age.

CALCULATING AGE PREDICTORS

The published data described above was not complete because information for the war and immediate post war period was missing. Consequently the data was supplemented by calculating name usage rates per 10,000 by gender for a random sample of approximately 120,000 babies born in 1945 from the Registrar Generals Indexes of Births in England and Wales. For practical reasons, no Scottish data was used. In combination the two data sources then provided the information for calculating age predictors.

The intention was to calculate a series of probabilities that the person was aged 65 or over; 55 - 64; 45 - 54; 35 - 44; 25 - 34; 18 - 24 (the last group was derived by interpolation). But before this could happen the data had to be adjusted to allow for differential rates of mortality. After all very few of those born in 1900 are still alive today whereas the rates of survival become progressively higher the more recent the date of birth. Moreover, the data also had to be adjusted to take into account the variations in the number of births by gender over time - ie peaks and troughs in the number of male and female births. Having adjusted the data to allow for these factors the probability of a name falling into each of the age groups was calculated. For example, the probability of someone named Barry falling within the different age groups is as follows:-

Age	Probability
65+	.010
55-64	.120
45-54	.288
35-44	.251
25-34	.218
18-24	.114

Thus a person with the name of Barry has only about a 1% chance of being aged over 65 a 12% chance of being aged 55-64 and so on. Most people with this name who are alive today are middle aged.

RESULTS

The final stage was to compare the age probabilities to true age for a stratified random sample of 100,000 individuals for whom both the name and the true age were known. This was achieved by matching the sample to the directory which contains the age probabilities by name and comparing the two. As a result of the matching process it was found that age predictors could be overlaid back onto 92.5% of all individuals in the random sample. Thus some 7.5% of all individuals in the sample had a name for which

there was insufficient data to calculate age predictors - ie they had a name that was less common and for which there were no age predictions. For those cases where a match was achieved the probabilities of the name falling into the six different groups were assigned. Within each age group the probabilities of occurrence were then summed over all matched records to provide estimates as to the proportion of cases in each age group. These were then compared to the true age and the results are given in Table 3-5.

Table 3 shows the comparison for all cases, and Tables 4 and 5 for males and females separately. Each table shows the true age and the probable age as predicted by name. Looking at Table 3, the figures show that for cases where the true age is 65 or over, some 41.4% have a predicted age (based upon the name driven inference) which is 65 or over. This prediction is in fact 2.02 times better than could have been expected given the proportion of the total population in Great Britain aged 65 or more. Thus, an index of 202 is shown. Conversely, the figures show that some 4.7% of the 65 or over group were predicted as being aged 18-24 with an index of 35. In other words significantly fewer pensioners are predicted as being aged 18-24 than could have been expected by chance. The figures show that the name based age prediction is working quite well and does show a real ability to predict age from a name. This is most marked for the older and younger age groups. When the name effectiveness of the age predictor is examined by gender (Tables 4 & 5) the same broad pattern is to be found but the figures also show small but interesting gender differences. The name age predictor is usually slightly more effective in predicting the age of females as opposed to males with the exception of those aged 65 and over where the pattern is reversed. The reasons for these differences are not obvious but this finding is borne out by an examination of the variation in usage of common names (see Table 2 above) where the common female names seem to show stronger variation than those of the men. The change in pattern for the over 65's is probably a reflection of the relative size of the groups where the female population is both larger and more diverse in age terms than the male.

Table 3: True Age Compared to Predicted Age

True Age	Predicted Age	%	Index
65+ n=19,211	65+	41.4	202
	55-64	19.6	147
	45-54	14.0	95
	35-44	11.7	65
	25-34	8.5	42
	18-24	4.7	35
55-64 n=12,242	65+	29.2	142
	55-64	21.2	160
	45-54	18.7	126
	35-44	15.6	87
	25-34	10.4	52
	18-24	4.9	37
45-54 n=13,120	65+	17.3	84
	55-64	15.4	116
	45-54	20.2	137
	35-44	21.9	122
	25-34	17.3	86
	18-24	7.9	59
35-44 n=16,608	65+	10.7	52
	55-64	9.9	74
	45-54	15.9	108
	35-44	24.3	135
	25-34	25.8	129
	18-24	13.4	99
25-34 n=18,715	65+	6.1	30
	55-64	5.3	40
	45-54	10.1	68
	35-44	18.7	104
	25-34	34.2	171
	18-24	25.5	190
18-24 n=12,353	65+	5.9	29
	55-64	4.0	30
	45-54	7.1	48
	35-44	13.6	76
	25-34	32.8	164
	18-24	36.5	272

Table 4: True Age Compared to Predicted Age for Males

True Age	Predicted Age	%	Index
65+ n=8,076	65+	36.9	216
	55-64	19.3	143
	45-54	14.8	96
	35-44	13.6	73
	25-34	9.6	46
	18-24	5.8	41
55-64 n=6,223	65+	26.3	154
	55-64	18.6	138
	45-54	17.8	116
	35-44	17.5	93
	25-34	12.6	60
	18-24	7.2	50
45-54 n=6,829	65+	17.4	102
	55-64	14.9	110
	45-54	19.1	124
	35-44	21.4	114
	25-34	17.2	82
	18-24	9.9	69
35-44 n=8,631	65+	11.5	67
	55-64	10.8	80
	45-54	15.8	103
	35-44	22.7	121
	25-34	24.1	115
	18-24	15.1	106
25-34 n=9,799	65+	7.4	43
	55-64	7.1	53
	45-54	11.8	77
	35-44	20.6	110
	25-34	30.1	143
	18-24	23.1	162
18-24 n=6,540	65+	6.8	40
	55-64	6.0	44
	45-54	9.9	64
	35-44	17.6	94
	25-34	29.0	138
	18-24	30.6	215

Table 5: True Age Compared to Predicted Age for Females

True Age	Predicted Age	%	Index
65+ n=11,135	65+	43.5	184
	55-64	19.8	151
	45-54	13.7	96
	35-44	10.8	63
	25-34	8.0	42
	18-24	4.3	34
55-64 n=6,019	65+	30.3	128
	55-64	22.1	168
	45-54	19.0	133
	35-44	14.9	86
	25-34	9.6	50
	18-24	4.0	32
45-54 n=6,291	65+	17.3	73
	55-64	15.6	119
	45-54	20.5	144
	35-44	22.1	128
	25-34	17.3	91
	18-24	7.2	57
35-44 n=7,977	65+	10.4	44
	55-64	9.6	73
	45-54	15.9	112
	35-44	24.9	144
	25-34	26.4	138
	18-24	12.8	101
25-34 n=8,916	65+	5.8	24
	55-64	4.8	37
	45-54	9.6	67
	35-44	18.2	105
	25-34	35.4	186
	18-24	26.3	208
18-24 n=5,813	65+	5.6	24
	55-64	3.5	27
	45-54	6.3	44
	35-44	12.6	73
	25-34	33.9	178
	18-24	38.1	301

FURTHER DEVELOPMENTS

The comparison between true and predicted ages shows a marked correlation, particularly for the older and younger groups. However, there is no reason to limit the prediction of age to just one simple piece of information - ones first name - other information could quite easily be incorporated into the prediction. Possible variables to consider would be:

- Length of residence.
- Census Enumeration District data on age or age of children etc.
- OPCS small area statistics on births and deaths.
- Spouses predicted age.

The inclusion of additional data and the application of suitable modelling techniques should lead to a significant improvement in the accuracy of the age predictor.

REFERENCES

1 Dunkling, L. (1993) 'The Guinness Book of Names' Guinness Publishing Ltd.