Submitted: 22nd of November 2016
Published: 12th of December 2018
DOI: 10.26775/ODP.2018.12.12
ISSN: 2446-3884

Linguistic features in names and social status: an exploratory study of 1,890 Danish first names

Emil O. W. Kirkegaard*

Open Quantitative Sociology & Political Science

Abstract

A dataset of the relative general social status (S factor) of 1,890 first names of persons living in Denmark was obtained from a previous study. 1,100 linguistic features were generated based on n-grams augmented by regex and each name was scored on each feature. An initial check using t-tests showed strong signal in the features taken as a whole (42.5 % of p values were < .05), and that this was due mostly to low status names having rarer patterns. OLS and lasso regression were used to combine the linguistic features into a single model. The results showed strong evidence of signal in the data. As a control, the main geographic origin of each name was inferred using data from behindthename.com. I validated this by comparing social status by origin group with data from official sources, r = .72, n = 28. The main origin for each name was then entered as a covariate and models were rerun. The results indicated that subtle linguistic features still provide substantial incremental validity, though a precise numerical estimate was difficult to arrive at. I validated this conclusion by training the model only on the subset of data identified as Danish. Model out of sample predictive validity was substantial in general, r = .75 (including origin covariate), and r = .46 in the Danish subset (linguistic features only). I conclude that it is possible to train fairly accurate social status predictors from subtle linguistic patterns in names. It's possible that humans might pick up on such cues to inform social perception when limited data is available.

Keywords: first name, given name, social status, social inequality, S factor, computational linguistics, variable selection, penalized regression, lasso, n-gram, Denmark

1 Introduction

A 2015 study of 1,890 Danish first/given names found that there are large differences in social status between the first names (Kirkegaard & Tranberg, 2015). While one can relatively easily infer approximate ancestry from first names by consulting dictionaries of names, it is an open question whether first names contain subtle signs to their social status controlling for ancestry. Much evidence exists that people tend to agree which names are associated with which traits (Horne, 1986) and some research indicates that people's perceptions of first names' social status and their actual status are correlated (Joubert, 1994); see also Garwood (1976)). I did not have stereotypes for my set of first names. However, by training a machine learning model to the data one can attempt to mimic a human picking up subtle verbal cues to a name's social status. This allows one to indirectly investigate whether humans can learn to stereotype names accurately without having to memorize social status of individual names. The purpose of the study was to investigate to which degree this was possible.

2 Data

In 2014, a Danish newspaper (Ugebladet A4) bought detailed sociological data for more than 2,000 Danish first names (every name with >100 persons in Denmark). This data was placed on a website[1] where one could enter a name to see the information on each metric. A previous study used a scraper to automatically download all the data (Kirkegaard & Tranberg,

* Ulster Institute for Social Research, E-mail: emil@emilkirkegaard.dk

[1] Danish Navnehjulet, meaning The name wheel. Website found at http://www.ugebreveta4.dk/navnehjulet. Note that the link may no longer work at the time you read this paper. It is not possible to archive a working version of the website because it is uses server- side logic.

2015). The data were then analyzed and made publicly available in a standard format (csv). One analysis in the published paper examined the relationships between several important socioeconomic indicators: mean income, criminal convictions, ownership of a house and unemployment. These indicators were all found to be positively related to each other when negative outcomes were reversed, and so it was possible to speak of and score each name on a general socioeconomic factor, known as the S factor (Kirkegaard, 2014; Kirkegaard & Fuerst, 2017). The scores were based on the age-adjusted data to avoid age- related confounding. Because the age-adjustment was carried out on aggregated data, some age confounding likely remains. A few names in the dataset had data for both sexes (i.e. unisex names such as Kim). For each such pair, the data for the sex with the most persons was retained and the other excluded. There were 1,890 names in the final dataset.

2.1 *Feature creation*

There are many ways one can create features (variables) about strings such as names. A simple approach involves scoring the name for whether some letter pattern occurs or not. I used a simple n-gram approach where I created every possible 3-length letter permutation[2]. Because Danish has 29 letters[3], this resulted in 293 = 24,389 patterns. To these, I added two variants for whenever the pattern occurred in the start or the end of the name, bringing the total number of n-gram patterns to 73,167. I then pruned the features to those that occurred at least 5 times in the dataset, which reduced the set to 1,094. Finally, I added a final set of 5 linguistically informed features, namely: length, fraction vowels [a, e, I, o, u, æ, ø, a], fraction stop sounds [t, d, p, b, k, g], fraction nasals [n, m], and whether a dash was present. For instance, the name Peter would be scored as having the following n-grams: p, e, t, r, pe, et, te, er, pet, ete, ter, as well as their initial and ending variants. It would furthermore have a vowel fraction of 2/5, stop sound fraction of 2/5, nasal sound fraction of 0, and be negative for presence of a dash. All the other features would be negative. Thus, each name has 1,099 features associated with it, of which 1,995 are binary, and 4 are numeric.

3 Analyses

3.1 *Individual features*

As an initial test of signal, a t-test was run for each feature with S factor score as the dependent variable. For each t-test, the sample size of the target group, the difference score (d), and the p value were saved. Statistical theory predicts that for a large number of null hypotheses tests, a set of predictors with no predictive validity will produce a uniform distribution of p values[4]. Similarly, the d values will form a normal distribution with a mean of 0 and a standard deviation that depends on the sample sizes used (smaller samples will produce larger standard deviations because the standard errors are larger). Figures 1 and 2 show the distributions of p and d values, respectively. The numeric output of the t-tests can be found in the supplementary materials.

The distribution of p values was right skewed (skew = 1.01), indicating signal in the dataset; 42.6 % of the p values were below .05 as opposed to 5 % expected if the null hypothesis of no signal were true.

The distribution of d values showed a long tail of negative values. This indicates that patterns tend to predict because names with low social status involve more rare patterns. This can be inferred because only rare names would tend to produce very large effect sizes[5]. Inspection of the patterns with the largest negative effect sizes confirmed this inference: the average overall number of names per pattern was 28.4, but for the top 20 and 100 patterns with the largest negative effects, this was only 6.15 and 9.68, respectively.

3.2 *Name origin*

Casual inspection of the names revealed that most of the low status names are foreign in origin. Table 1 shows the top and bottom 10 names.

It is obvious that the bottom 10 names are all non-Western, mostly/entirely Muslim. This raises the possibility that the linguistic feature associations we see are merely crude associations for Muslim or non- Scandinavian/non-Western names. To investigate whether this was the case, I collected metadata about the names by scraping the website behindthename.com, which provides origin data for first names. For each name, I saved all the origins list. The data were then recoded into two variables.

[2] I.e. sequences of letters where order matters; "abc" is another pattern than "cba".

[3] The usual 26 English letters plus the three vowel letters *æ*, *ø* and *a*.

[4] This testing approach was also recently used by Kirkegaard & Bjerrekær (2016b).

[5] Rare name features result mainly from rare patterns in Danish names and from names from other languages. Since most non-Danish languages are below Danish social status (especially the non-European ones) (Kirkegaard & Fuerst, 2014), the patterns in them will generally have a negative effect size and be rare, producing the left tail of effect sizes. Furthermore, by sampling theory, we expect larger effect sizes to come from smaller samples in general and this will be seen in both tails. In fact, every pattern with an absolute effect size above 1 (n = 93) had a sample size of 71 or fewer (see plots in supplementary materials), and the correlation between absolute effect size and sample size is -.13 [95CI: -.19 to -.07].

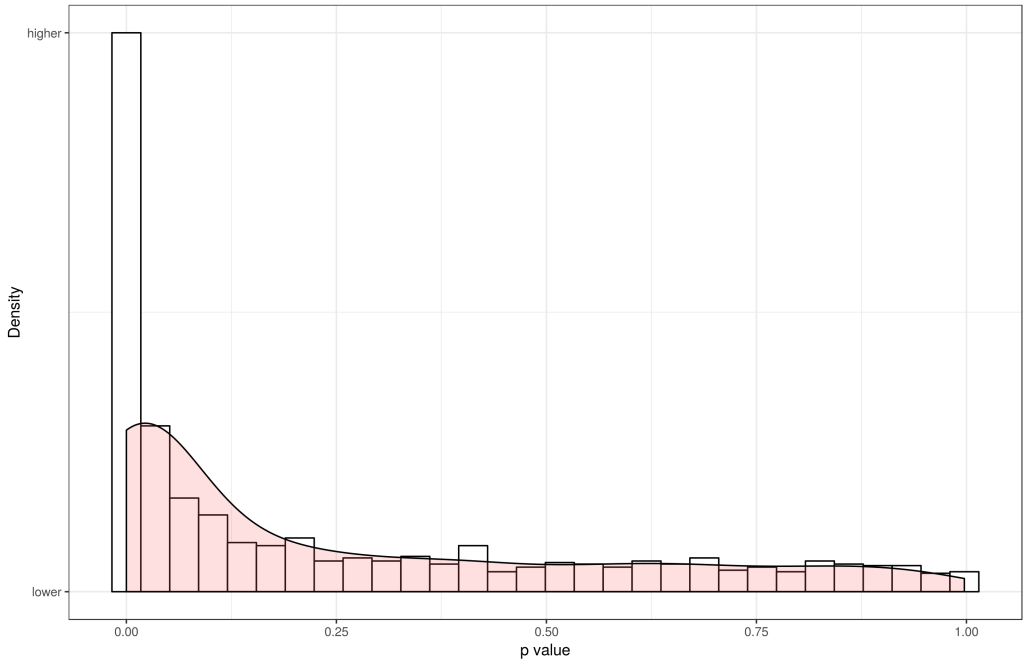

Figure 1: Density-histogram of the distribution of p values based on t-tests.

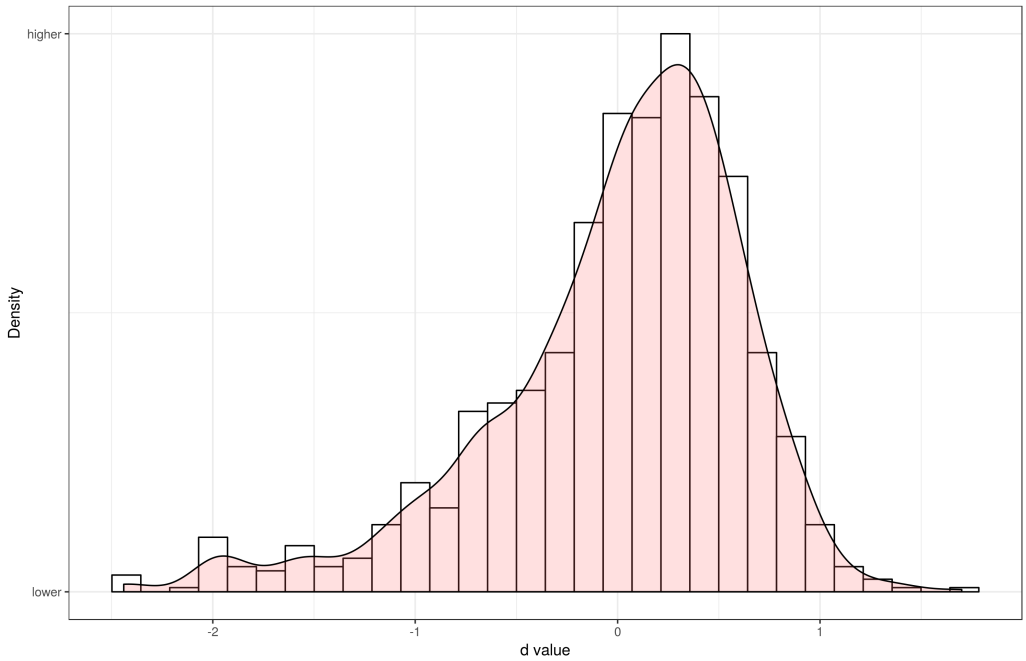

Figure 2: Density-histogram of the distribution of d values based on t-tests.

Table 1: Top and bottom 10 names by general socioeconomic factor (S) score. The number of adults in Denmark in 2012 was 5.592 million.

Rank	Name	S	Number of persons	Rank	Name	S	Number of persons
1	Lauritz	2.39	588	1881	Fadumo	-3.33	205
2	Alberte	2.31	5084	1882	Abdi	-3.15	331
3	Gustav	2.19	7279	1883	Abdullahi	-2.99	231
4	Villads	2.18	2308	1884	Fatme	-2.96	166
5	Eskild	2.16	592	1885	Mehmed	-2.88	121
6	Lauge	2.13	1037	1886	Mustapha	-2.86	173
7	Trille	2.02	174	1887	Karima	-2.81	144
8	Valdemar	2	3440	1888	Souad	-2.78	120
9	Alfred	2	2855	1889	Salah	-2.72	186
10	Jens-Ole	1.99	173	1890	Halima	-2.64	201

First, a dichotomous variable for whether Danish was one of the origins or not. Second, I used the main origin listed. Since this resulted in a large number (103) of quite rare origins, I combined similar rare origins until only 48 categories remained. For instance, "Hindi", "Hinduism", "Marathi" and "Urdu" were combined into a joint "Indian" category. A complete list of pre-recoding origins and recoding decisions can be found in the supplementary materials. Figures 3 and 4 show the distributions of S scores by Danish/non-Danish origin, and the 10 largest origins, respectively[6].

The relative ranking of the top 10 origin groups corresponds fairly well to expectations based on the origin countries' well-being (Kirkegaard, 2014) (Kirkegaard, 2014). To get a numerical comparison, I manually matched the origin groups to the national groups from a previous study of Danish immigrant groups (Kirkegaard & Fuerst, 2014). I was able to match up 28 out of the 47 main origins to a specific country, shown in Figure 5. Details of the matching can be found in the supplementary materials.

The relationship was quite strong with a number of notable outliers. This is likely to result from non-random naming of persons from the specific groups (e.g. high status, secular Turks might give their children Turkish names while most others give their children Muslim names). I was unable to match most of the Muslim countries to a main origin because most Muslim names were simply coded as general Arabic origin. Because the Muslim countries generally perform poorly (correlation between Muslim % in origin country and general social status = -.63 (Kirkegaard & Fuerst, 2014)), the non-matching resulted in decreased variance ('restriction of range') thus reducing the strength of the observed correlation.

[6] Unknown origin was among the top 10, but was excluded as plotting it is not very informative.

Satisfied with the validity of the inferred origin data, I proceeded to re-estimate the relationship between each linguistic feature and S, but this time including a control for main origin. This was done using OLS regression, one feature as a time (i.e. 2 predictors). Controlling for main origin reduced the detectable validity of the linguistic features markedly. Before, 42.6 % had p values below .05, while after only 16.3 % of them did. Still, this implies a substantial amount of signal in the features not accounted for by the inferred origin. Furthermore, the correlation between the naive bivariate effect size and the controlled effect size was .70 (CI95: .67 to .73, n = 1,095), so there was considerable relative stability of effect sizes.

3.3 OLS regression

Satisfied that the linguistic features were not entirely redundant with inferable name origins, I proceeded to estimate a multivariate model using multiple linguistic features. Since the dataset had more cases than predictors (n > p; 1,100 predictors, 1,900 cases), it was possible to fit a full model using OLS, though this would result in a lot of overfitting due to the low n/p ratio (about 1.72). I fit a number of models to the data and their summary statistics are shown in Table 2.

There was strong evidence of joint validity of the linguistic features, and while the adjusted R2 metric indicates quite a lot overfitting, there was plenty of estimated residual validity. Model 1 had 70 predictors with missing beta estimates, presumably related to strong multicolinearity among the predictors. The predictors was inspected, but there didn't seem to be any particular pattern to them. In terms of the results from the t-tests in Section 3.1, they had smaller sample sizes (mean/median 7.51/6 vs. 29.8/10) and somewhat greater absolute effect sizes (mean/median .64/.48 vs. .47/.38), though the latter might be due to inflation from the less precise estimates. When they were removed in models 2-4, no predictor had

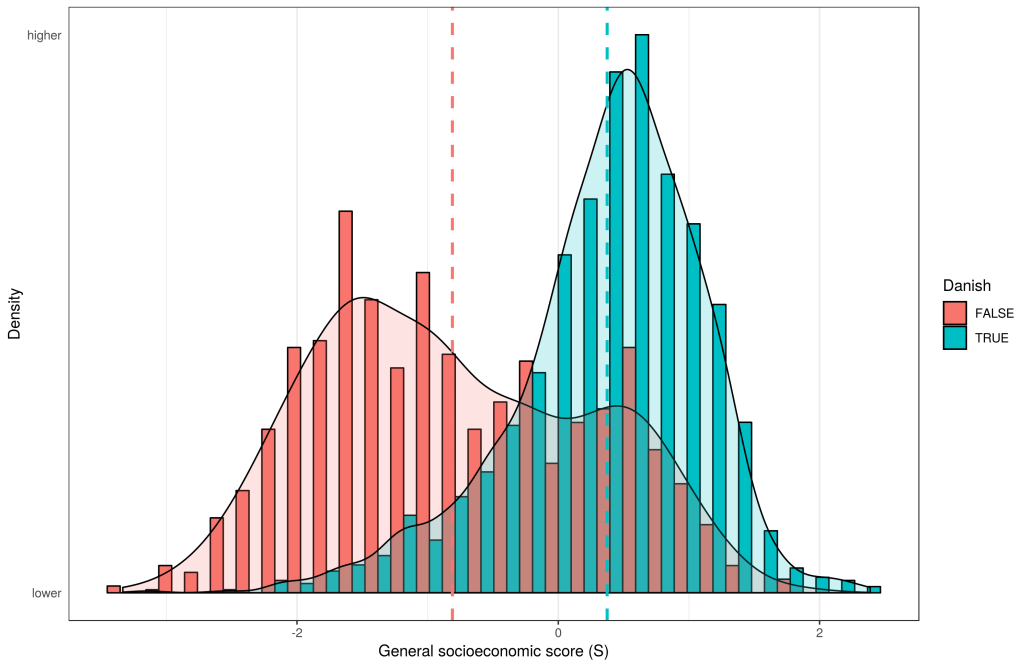

Figure 3: Distribution of Danish vs. non-Danish names on general socioeconomic factor (S). The mean/sd for Danish names was 0.37/0.72, and for non-Danish -0.81/1.04.

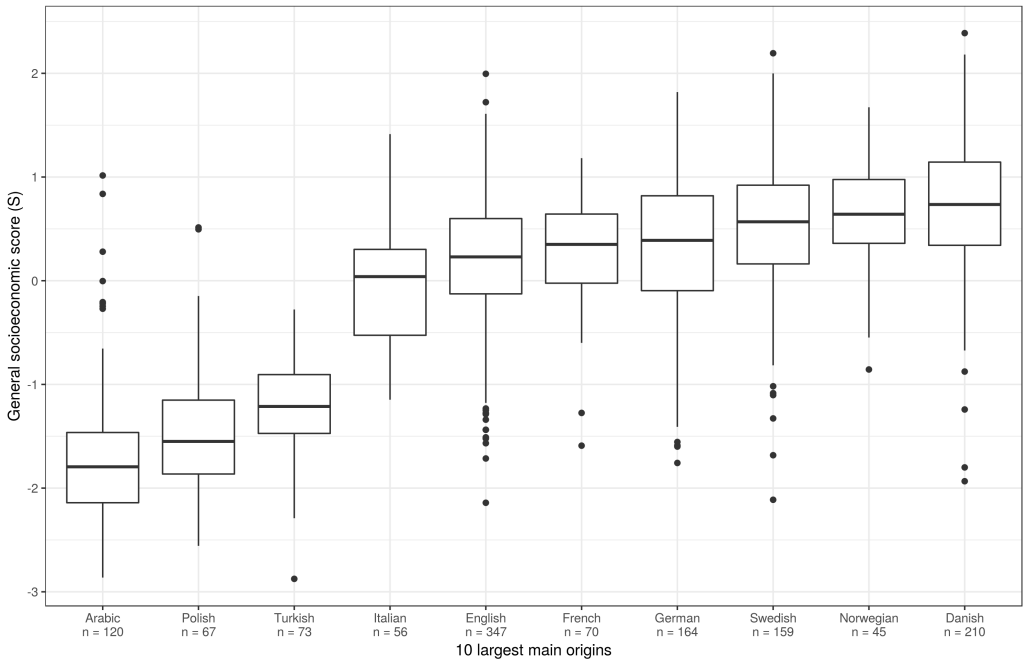

Figure 4: Distribution of top 10 most common origins' general socioeconomic factor scores (S).

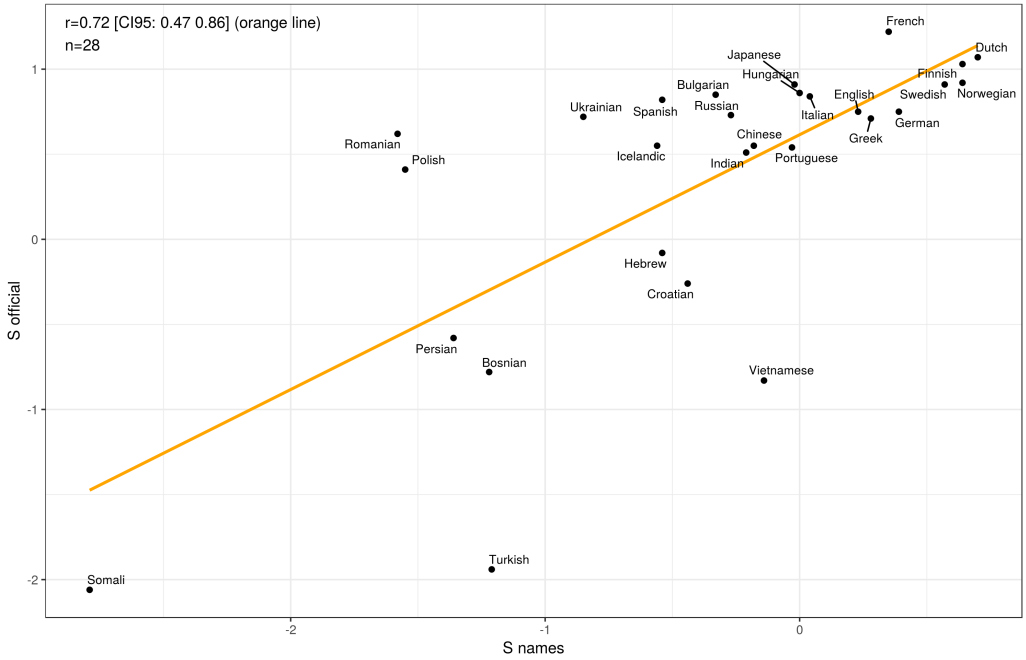

Figure 5: Scatterplot of general socioeconomic factor (S) scores estimated from first names and official statistics.

Table 2: Summary statistics of fitted OLS models.

Model #	Description	R^2	R^2 adj.	Number of predictors
1	Linguistic features only.	0.825	0.582	1099
2	Same as 1, but without predictors that had missing beta estimates.	0.825	0.616	1029
3	Same as 2, but with Danish origin control.	0.838	0.644	1030
4	Same as 2, but with main origin control.	0.865	0.686	1030
5	Danish origin only.	0.303	0.303	1
6	Main origin only.	0.538	0.526	1

missing beta estimates in the new fit. Interestingly, removing the predictors left the model with the same R2 as before (.825), showing that they were entirely redundant, but with a resulting higher R2 adj. (.616 vs. .582) because fewer degrees of freedom were used.

Models 3-4 showed that adding name origins to the model increased validity substantially, and that adding main origins had greater validity than the simple binary Danish or not origin (R2 adj. .644 vs. .686).

Models 5-6 served to check whether the origins alone with match the validity of the linguistic features. This was clearly not the same, and even the most pessimistic interpretation would leave about 16 % more variance explained by the linguistic features beyond the main origin (i.e. difference between R2 adj. of .686 vs. .526, models 4 and 6).

3.4 Penalized regression

Penalized regression is a family of regression methods that reduce the estimated betas of predictors using a modification of the error function that attaches a penalty as a function of the betas in the model (James et al., 2013). Many variations exist on the theme (Hastie et al., 2015) (Hastie, Tibshirani, & Wainwright, 2015), but here I used the elastic net approach. The elastic net combines the lasso and ridge regression approaches using a tuning parameter (alpha). When this value is 1, it reduces to the lasso, and when it is 0, it reduces to ridge regression. Intermediate values result in a combination of the two. While one could obtain a best estimate of the alpha parameter using a cross-validation approach, I instead opted for utilizing just a few values (4 values: .325, . 55, .775, 1). Using penalized regression also requires a second tuning parameter, lambda, that controls the amount of shrinkage. When this value is 0, the model reduces to the OLS equivalent. I estimated the optimal value of this parameter using cross-validation (CV) as described by James et al. (2013). Specifically, for each alpha value, I ran 5,000 double cross-validations. The method was as follows:

1. The dataset is split into training and test parts, with 50 (2.6 % of the dataset) random names chosen for the test set and the remainder for the training set.
2. 10-fold cross-validation is used on the training set to estimate lambda. The resulting model (i.e. variables with non-zero betas) at the optimal lambda is saved. The entire set of predictors is used, including those that produced missing values in the full OLS regression, as well as the main origin control.
3. Predictions for the test set are made in two ways. First, from the lasso estimated betas. Second, from the OLS estimated fit using the model selected by the lasso. This two-step approach has been suggested in the literature (Hastie et al. 2015, sec. 2.2; Meinshausen 2007)[7].
4. The R2 validity of the two sets of predictions is calculated for the test set using the usual sum of squares formula. The information of interest are then saved.

After the 5,000 runs were completed, average CV R2 was calculated for each setting combination. The results showed that using an alpha of .775 was very slightly superior to the standard lasso: .533 vs. .531 R2, respectively. Lasso estimation of betas was always superior to OLS estimation (for alpha = 1, R2 of OLS estimation was .501). To visualize the accuracy of the model, the test set predictions of each sample were averaged within case to produce a single set of predictors, shown in Figure 6[8].

The observed correlation closely matches the expected: R2 of .531 converts to r = .73. The slightly larger value seen is likely due to a small bias from ignoring the intercept (correlations ignore the intercept, but R2 in CV does not). The cross-validated lasso R2 is substantially lower than the adj. R2 obtained from a the full OLS model fit to the entire dataset (model 4, adj. R2 = .686). This likely means some combination of insufficient overfitting adjustment and that the OLS model used a slightly larger dataset for training (n = 1,980 for full vs. 1,930 used for each CV fold). If one disregards the variable selection overfitting issue and instead performs CV lasso on the full dataset, the resulting model is found to have an R2 of .636, in between the two estimates.

[7] I used cross-validated relaxed lasso on the full dataset. This indicated that the optimal value of phi (the beta shrinkage parameter) was 1, i.e. standard lasso. Despite this result, I wanted to make sure, so I used both lasso and OLS to make predictions, corresponding to phi values of 1 and 0, respectively.

[8] One might think that this averaging within case might average out some error and inflate the validity. To investigate this, instead another dataset was created by sampling a random predicted value for each case. This produced nearly the same result (r's .75 vs. . 74).

An alternative approach to cross validation to estimate out of sample validity is to utilize bootstrapping as described by Harrell (2015, sec. 5.3.4). This has the advantage that one can train on a dataset of the full size and thus avoid loss of model accuracy due to decreased sample size. Unfortunately, using this method for the present dataset presented with both programmatic and statistical issues due to the fact that the main set of predictors are binary variables often with only a few cases. This means that resampled samples usually do not contain any cases for many of the predictors, making it impossible to fit the model. See further discussion at Cross- Validated[9].

3.4.1 *Danish subset*

As a robustness check, I subset the data to the names tagged as Danish origin, n = 1,299. This allowed me to drop the origin controls, and furthermore reduced the number of possible n-grams because many of them were no longer found in at least 5 cases (805 remaining). Running the t-test initial test still showed substantial evidence of signal with 25.9 % of p values < .05. The distribution of effect sizes was now close to normal (Figure 7), including that the inclusion of foreign names caused most of the left-tail seen in Figure 2.

The lasso analysis was run on the Danish subset as before. This produced an optimal R2 of .172 using alpha = 0.1 and lasso beta estimates. However, as before, the differences between the alpha parameters were marginal with the standard lasso obtaining R2 = .170. Figure 8 shows the out of sample predictions for the standard lasso (compare with Figure 5).

If instead the lasso fit on the full dataset is used, the correlation is .58.

4 Discussion and conclusion

The present study examined whether simple linguistic features (regex-enhanced n-grams) of first names could be used to predict relative social status in Denmark. There are multiple findings of interest. First, there was very strong evidence of signal in the features taken one by one, with 42.6 % of features having p < .05 in a t-test as opposed to 5 % if the null hypothesis of no signal were true.

Second, it was examined whether one could reliably infer origins of the names based on information from the popular name site http://behindthename.com. It was found that origin data was highly predictive

[9] Direct link https://stats.stackexchange.com/questions/213837/k-fold-cross-validation-nominal-predictor-level-appears-in-the-test-data-but-no, archived http://archive.is/03161.

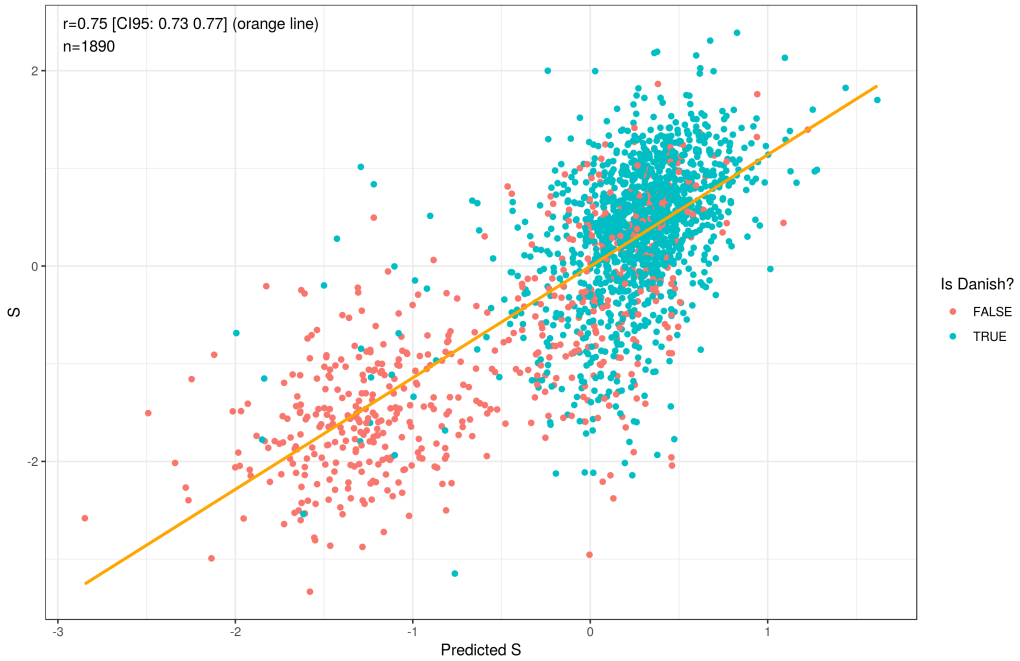

Figure 6: Predicted vs. actual social status (S factor). Danish status shown for illustration purposes only.

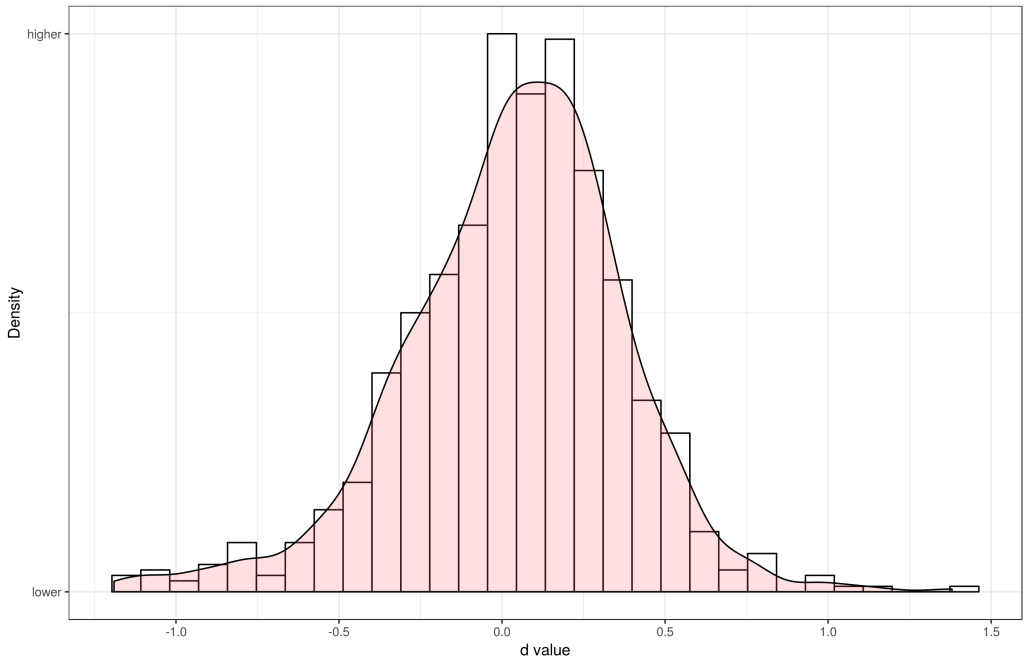

Figure 7: Density-histogram of the distribution of d values based on t-tests for Danish subset.

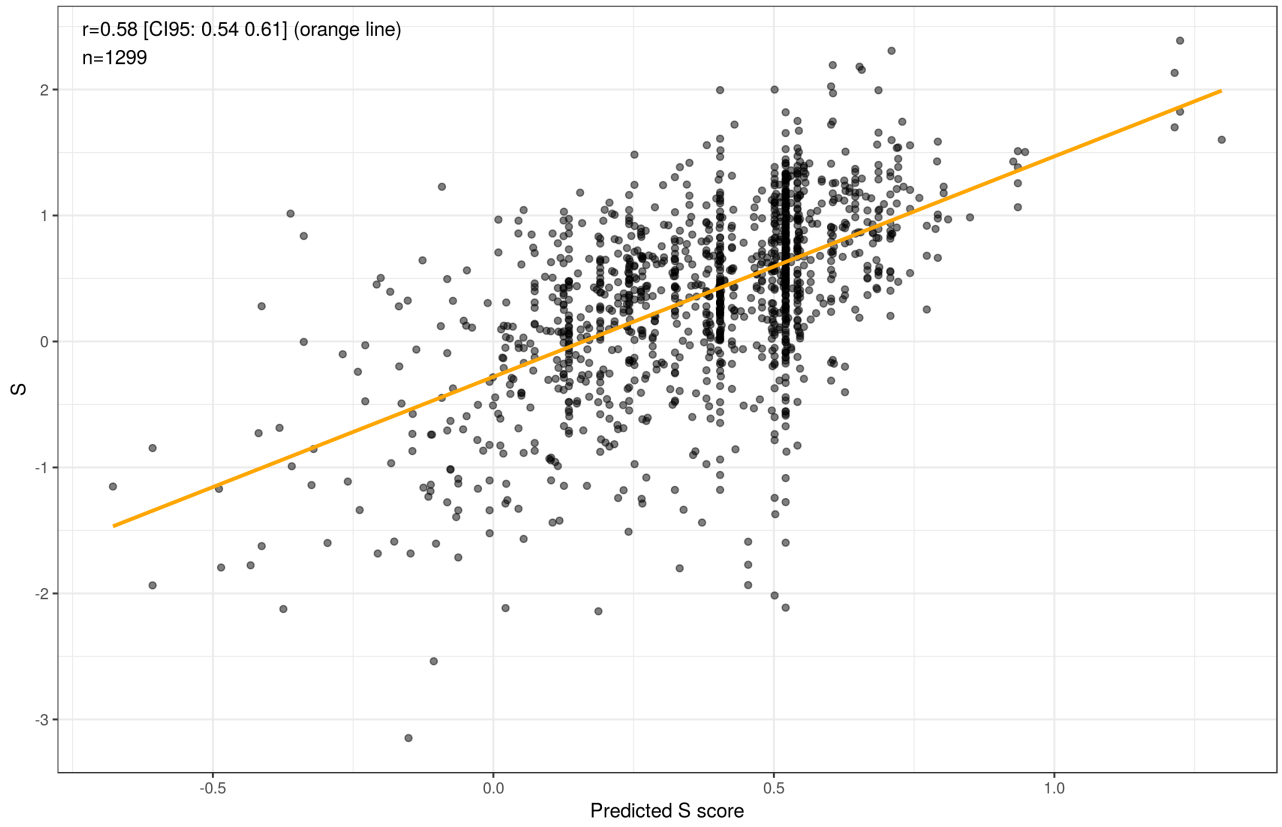

Figure 8: Scatterplot of general socioeconomic factor (S) scores estimated from first names and official statistics. Danish subset.

of social status, and that the predicted social status matched up well (r = .72, Figure 5) with official statistics-based results from a previous study (Kirkegaard & Fuerst, 2014).

Third, it was then shown using OLS models that linguistic features provided enhanced predictive value beyond inferred origin of the names. In the case where one considers just incremental adj. R2 validity on top of main origin, the linguistic features provided an increase of 16 % variance explained (from 52.6 % to 68.6, cf. Table 2).

Fourth, a penalized regression approach was used on the name data to avoid issues with overfitting related to the large number of predictor variables compared to the dataset. Double cross-validation was used to estimate the optimal amount of shrinkage and then validate the resulting model in a test set. The results indicated that lasso did not result in superior predictive validity beyond OLS on the full dataset if one uses the adjusted R2 metric: R2 .686 vs. .533. This might be due to the slight decrease in the training dataset for the cross validation runs (n = 50 decrease i.e. 2.6 %), or more likely because the R2 adjustment metric insufficiently corrects for overfitting.

One worrying finding is that the cross-validated R2 is actually smaller than the OLS adj. R2 for the main origin only model (model 6, Table 2). This could be taken to indicate that there was no true signal in the linguistic features, and the apparent validity was only due to overfitting. However, this interpretation is contradicted by the one at a time feature regression in Section 3.2, which showed that 15 % of features still showed $p < .05$ when main origin was a covariate. This value should have been 5 % if there was no signal. Thus, the overall pattern is findings is best interpreted as showing that linguistic features provide incremental validity beyond name origin, though the exact extent is somewhat unclear due to the overfitting of the full sample OLS and the inability of the adjusted R2 metric to properly correct for this.

In a robustness check, I reran the analyses on a subset of names that were identified as Danish in origin (n = 1,299). The findings on this dataset were weaker, but still quite substantial, with the standard lasso obtaining a validated R2 of .170 (r = .46 in Figure 8). Thus, linguistic patterns offer quite a bit of information about the relative social status of Danish first names.

The study was based on data from a previous study which showed that there is substantial first name-linked variation in social status in Denmark. The present study shows that this extends to subtle patterns in the names. Given the pervasive evidence

that humans are able to fairly accurately estimate average trait levels of groups (stereotype accuracy) (Jussim, 2012; Jussim et al., 2015; Kirkegaard & Bjerrekær, 2016a), these findings suggest that humans may also be able to estimate social status based on first names and perhaps even subtle linguistics patterns in names. There is some indirect evidence of this (Gebauer et al., 2012). If humans are able to do so, this would enable them to use realistic priors in situations where uncertainty is present and first names are known (e.g. Airbnb renting, and online peer to peer trading a la ebay). This has implications for e.g. studies attempting to hypotheses using names (Bertrand & Mullainathan, 2003; Edelman et al., 2017; Fryer & Levitt, 2004) (Bertrand & Mullainathan, 2003; Edelman, Luca, & Svirsky, 2017; Fryer & Levitt, 2004).

The study was primarily limited by the number of first names available and the fact that the large number of predictors lead to substantial issues with overfitting.

Supplementary material and acknowledgments

Supplementary materials including code, high quality figures and data can be found at https://osf.io/adgwd/.

The peer review thread is located at https://openpsych.net/forum/showthread.php?tid=295.

Thanks to reviewers: Dr. g (pseudonym) and Jose Luis Ricon.

Special thanks to gwern for helpful comments.

References

Bertrand, M., & Mullainathan, S. (2003). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *National Bureau of Economic Research*. (Working Paper No. 9873) doi: 10.3386/w9873

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, *9*(2), 1–-22. doi: 10.1257/app.20160213

Fryer, R. G., & Levitt, S. D. (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics*, *119*(3), 767–-805.

Garwood, S. G. (1976). First-name stereotypes as a factor in self-concept and school achievement. *Journal of Educational Psychology*, *68*(4), 482–-487. doi: 10.1037/0022-0663.68.4.482

Gebauer, J. E., Leary, M. R., & Neberich, W. (2012). Unfortunate first names: Effects of name-based relational devaluation and interpersonal neglect. *Social Psychological and Personality Science*, *3*(5), 590–-596. doi: 10.1177/1948550611431644

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). New York: Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Boca Raton: CRC Press, Taylor and Francis Group.

Horne, M. D. (1986). Potential significance of first names: A review. psychological reports. , *59*(2), 839-–845. doi: 10.2466/pr0.1986.59.2.839

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. New York: Springer.

Joubert, C. E. (1994). Relation of name frequency to the perception of social class in given names. *Perceptual and Motor Skills*, *79*(1), 623-–626. doi: 10.2466/pms.1994.79.1.623

Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. Oxford University Press.

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in)accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science*. doi: 10.1177/0963721415605257

Kirkegaard, E. O. W. (2014). The international general socioeconomic factor: Factor analyzing international rankings. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2014/09/the-international-general-socioeconomic-factor-factor-analyzing-international-rankings/ doi: 10.26775/ODP.2014.09.08

Kirkegaard, E. O. W., & Bjerrekær, J. D. (2016a). Country of origin and use of social benefits: A large, preregistered study of stereotype accuracy in denmark. *Open Differential Psychology*. Retrieved from https://openpsych.net/paper/49 doi: 10.26775/ODP.2016.11.11

Kirkegaard, E. O. W., & Bjerrekær, J. D. (2016b). The okcupid dataset: A very large public dataset of dating site users. *Open Differential Psychology*. doi: 10.26775/ODP.2016.11.03

Kirkegaard, E. O. W., & Fuerst, J. (2014). Educational attainment, income, use of social benefits, crime rate and the general socioeconomic factor

among 70 immigrant groups in denmark. *Open Differential Psychology*. Retrieved from https://openpsych.net/paper/21 doi: 10.26775/ODP.2014.05.12a

Kirkegaard, E. O. W., & Fuerst, J. (2017). Admixture in argentina. mankind quarterly. , *57*.

Kirkegaard, E. O. W., & Tranberg, B. (2015). What is a good name? The S factor in Denmark at the name-level. *The Winnower*. Retrieved from https://thewinnower.com/papers/what-is-a-good-name-the-s-factor-in-denmark-at-the-name-level

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis*, *52*(1), 374-–393. doi: 10.1016/j.csda.2006.12.019