Why should social scientists care about supervised learning?

Malte Schierholz[1] & Christoph Kern[2]

[1]University of Mannheim, MZES
Institute for Employment Research (IAB)
[2]University of Mannheim

Malte.Schierholz@iab.de
c.kern@uni-mannheim.de

March 21 and 22, 2018

Outline

Prediction Problems

Applied econometric perspective (Kleinberg et al. (2015) and Mullainathan et al. (2017)):

- Make use of new, high-dimensional data sources (text, images)
  - Predict future harvest or local poverty level from satellite images?
  - Predict hygiene of restaurants from restaurant reviews found online?

Prediction Problems

Applied econometric perspective (Kleinberg et al. (2015) and Mullainathan et al. (2017)):

- Make use of new, high-dimensional data sources (text, images)
  - Predict future harvest or local poverty level from satellite images?
  - Predict hygiene of restaurants from restaurant reviews found online?
- Classical statistical procedures involve prediction
  - First stage in instrumental variable estimation
  - Heterogenous treatment effects in causal inference
  - Other inference tasks may be seen to involve prediction implicitly

Prediction Problems

Applied econometric perspective (Kleinberg et al. (2015) and Mullainathan et al. (2017)):

- Make use of new, high-dimensional data sources (text, images)
  - Predict future harvest or local poverty level from satellite images?
  - Predict hygiene of restaurants from restaurant reviews found online?
- Classical statistical procedures involve prediction
  - First stage in instrumental variable estimation
  - Heterogenous treatment effects in causal inference
  - Other inference tasks may be seen to involve prediction implicitly
- More policy prediction problems

Applications

Machine learning for resource allocation

- *Labor Market Policy*: Predict unemployment spell length
  - $\rightarrow$ Decide on job search strategies
- *Social Policy*: Predict youth at risk for violence (Chandler et al. 2011)
  - $\rightarrow$ Select students for prevention initiatives
- *Education*: Predict non-graduation (Lakkaraju et al. 2015)
  - $\rightarrow$ Target intervention programs
- *Economic Development*: Predict onset of urban blight (Blancas Reyes et al. 2017)
  - $\rightarrow$ Target property inspections
- *Criminal Justice*: Identify risky defendants (Kleinberg et al. 2017)
  - $\rightarrow$ Data-driven jail-or-release decisions

Applications

ML applications cont'd

- *Predictive Policing*: Calculate probability of future crimes to be committed
  - $\rightarrow$ Police can take precautionary actions
- *Targeted Advertising*: Predict interest in a product based on personal characteristcs
  - $\rightarrow$ Advertising costs decrease
- *Telematics auto insurance*: Predict risk of car accident based on personal driving behavior
  - $\rightarrow$ Safe drivers pay less for their insurance
- *Finance*: Predict credit-worthiness
  - $\rightarrow$ Minimize risk of loan nonpayments

Applications

Original idea: Effectiveness and objectivity (compared to e.g. teachers, judges, bank employees)

Fairness of ML algorithms (Daumé 2017)?

- Train-test mismatch
  - Accuracy problems when distributions do not match
- Biased training data
  - Bias in data (by human labelers) will be reproduced by model
- Usage of gender and ethnicity as features
  - Exclusion does not help when proxy variables are available
- Feedback loops

Model Misspecification

Model Misspecification

Model Misspecification

Standard model assumption:
$f(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

What if ...

- this assumption of linearity does not hold?
- there are interactions between covariates?

$\rightarrow$ Model misspecification!

Supervised learning is about estimating flexible functions
$\rightarrow$ Less prone to misspecification error

Diagnosing Model Misspecification

But how do we know if the linear assumption is appropriate?

1. Compare predictive performance (e.g., out-of-sample R^2, MSE, ...) from linear model with a more flexible approach?
    - Large difference in predictive performance would suggest a misspecified model
2. Explain variation of residuals from linear model?
    - Run non-linear model (e.g., a tree) to predict residuals of linear model

Model Misspecification: An Example

Or maybe we should estimate and interpret $f(x)$ directly, without ever making the linear assumption?

A political science replication study taken from Hainmueller and Hazlett (2014)

Model Misspecification: An Example

Setting:

- N = 126 political instability events (internal wars / regime changes)
- What factors can predict if a genocide will happen?

Key differences between original and replication study

- automatic model selection vs. extensive human specification search for an appropriate model
- Automatic method is less susceptible to misspecification bias
- Higher predictive power: R^2 similar (32% vs. 34%) but automatic method has significant higher ROC-AUC

Model Misspecification: An Example

Table 4 Predictors of genocide onset: OLS versus KRLS

Estimator	*OLS*	*KRLS*			
		$\partial y/\partial x_{ij}$			
	β	*Average*	*1st Qu.*	*Median*	*3rd Qu.*
Prior upheaval	0.009*	0.002	−0.001	0.002	0.004
	(0.004)	(0.003)			
Prior genocide	0.263*	0.190*	0.137	0.232	0.266
	(0.119)	(0.075)			
Ideological char. of elite	0.152	0.129	0.086	0.136	0.186
	(0.084)	(0.076)			
Autocracy	0.160*	0.122	0.092	0.114	0.136
	(0.077)	(0.068)			
Ethnic char. of elite	0.120	0.052	0.012	0.046	0.078
	(0.083)	(0.077)			
Trade openness (log)	−0.172*	−0.093*	−0.142	−0.073	−0.048
	(0.057)	(0.035)			
Intercept	0.659				
	(0.217)				

Note. Replication of the "structural model of genocide" by Harff (2003). Marginal effects of predictors from OLS regression and KRLS regression with standard errors in parentheses. For KRLS, the table shows the average of the pointwise derivative as well as the quartiles of their distribution to examine the effect heterogeneity. The dependent variable is a binary indicator for genocide onsets. $N = 126$. $^*p < 0.05$.

- Similar marginal effects on most variables except for "prior upheaval"
- With log(prior upheaval) the OLS effect goes away
  $\rightarrow$ misspecification in OLS model

(taken from Hainmueller and Hazlett,2014, p. 165)

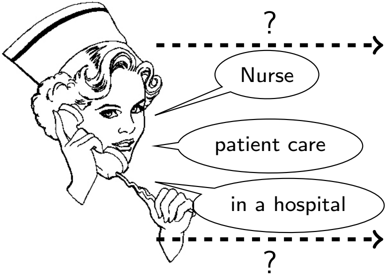
Occupation Coding

Occupation Coding

Occupation Coding

Assign verbatim answers into an official classification

Survey Answers



Classification

KldB 2010	Job Title
81302	Nurse
81313	Graduated Nurse
81323	Pediatric Nurse
⋮	⋮
81102	Doctor's Assistant
82102	Geriatric Nurse

How to find the most appropriate category efficiently?

Occupation Coding

Two approaches to automation:

- Automated coding: Computer assigns codes by itself (requires top-category)
- Computer-assisted coding: Computer suggests possible codes to a human coder who is responsible for the final decision (requires category ranking)

Occupation Coding

Different algorithms can suggest possible codes. Key idea of machine learning:

- Use coded data from the past to predict future codes
- Possible algorithms: SVM, Regression, Nearest Neighbor, Boosting, ...

Occupation Coding

Different algorithms can suggest possible codes. Key idea of machine learning:

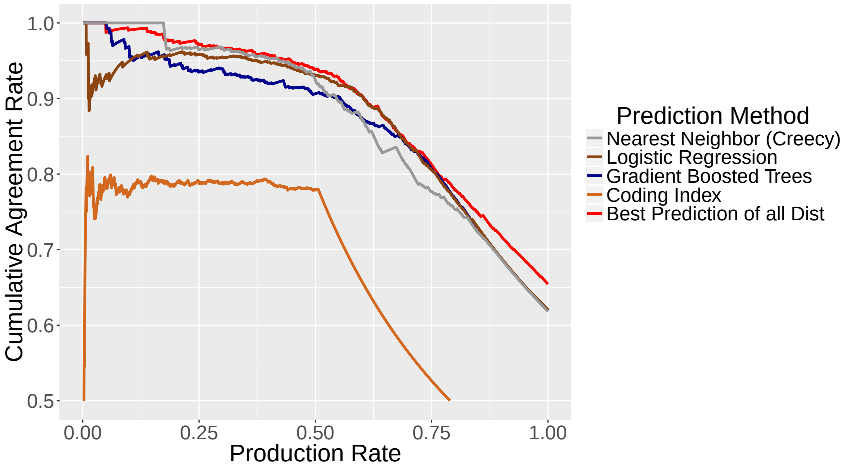
- Use coded data from the past to predict future codes
- Possible algorithms: SVM, Regression, Nearest Neighbor, Boosting, ...

Practical challenges (what makes my research difficult):

- Prediction problem is high-dimensional (1200 categories in target variable, 25.000 variables (words) in predictor matrix), but my training data is comparatively small (90.000 observations)
  - $\rightarrow$ specialized algorithm needed
- Even human coders often disagree about the correct code
  - $\rightarrow$ ask respondents during the interview for more details

(see Schierholz et al. 2018 for details)

Occupation Coding



Performance of automated coding

(ALWA data, $N_{train} = 29,740$, $N_{test} = 3,189$)

Sampling Theory and Statistical Learning

Sampling Theory and Statistical Learning

Relates to Generalized Regression (GREG) estimators (not discussed here)

The following is based on Breidt and Opsomer (2017).

Sampling Theory and Statistical Learning

Setting:

- Finite population U (size $= N$), variable $y_i, i = 1, ..., N$, not observed
- Goal: calculate population total $t_y = \sum_{i \in U} y_i$
  - Only chosen for simplicity, others targets are possible
- Draw sample S at random with inclusion probabilities π_i known

Sampling Theory and Statistical Learning

Setting:

- Finite population U (size $= N$), variable $y_i, i = 1, ..., N$, not observed
- Goal: calculate population total $t_y = \sum_{i \in U} y_i$
  - Only chosen for simplicity, others targets are possible
- Draw sample S at random with inclusion probabilities π_i known

Horwitz-Thompson estimator

$$\hat{t}_{y,HT} = \sum_{i \in S} \frac{y_i}{\pi_i}$$

is unbiased.

It still can be improved if we have auxiliary variables $x_i, i = 1, ..., N$ available for the complete population U (as in administrative data)

Sampling Theory and Statistical Learning

Imagine a "method" $f(\cdot)$ for predicting y_i from x_i

- May be known in advance or estimated from survey data

Sampling Theory and Statistical Learning

Imagine a "method" $f(\cdot)$ for predicting y_i from x_i

- May be known in advance or estimated from survey data

Consider the difference estimator

$$\hat{t}_{y,Diff} = \underbrace{\sum_{i \in U} f(x_i)}_{\substack{t_f \\ \text{Predicted total in population}}} + \underbrace{\sum_{i \in S} \frac{y_i - f(x_i)}{\pi_i}}_{\substack{\hat{t}_{y,HT} - \hat{t}_f \\ \text{Estimated difference}}}$$

Key results:

- $\hat{t}_{y,Diff}$ is unbiased regardless of the quality of f
- $\mathrm{Var}(\hat{t}_{y,Diff})$ becomes smaller, the better f predicts y
- Most promising for small area estimation

Propensity Score Estimation

Propensity Score Estimation

Propensity Score Estimation

Potential outcomes (Rosenbaum & Rubin 1983)

- Framework for analyzing causal problems
- Defines treatment effects in terms of counterfactuals
    - Unit i, treatment w, potential outcome $Y_i(w)$
    - Causal effect: $Y_i(w') - Y_i(w)$
- However, only one outcome for each i can be observed

Observational data

- Estimate treatment effects between groups ($\neq$ within individuals)
- Needs identifying assumptions (e.g. CIA)

Propensity Score Estimation

→ Propensity scores: Account for differences between treated and untreaded by matching or balancing based on treatment propensity

- Reduces high dimensional problem (many covariates) to single score

1. Estimate propensity to be treated given covariates
   - Logistic regression
     - Many covariates, functional form unclear
   - Boosting (e.g. McCaffrey 2004, Lee et al. 2010, Ridgeway et al. 2017)
     - Adaptive, starts with baseline log-odds, pseudo residuals $y_i - \hat{p}_i$
2. Use propensity score to find matches or to reweight observations

Propensity Score Estimation

Tuning GBMs for propensity score estimation

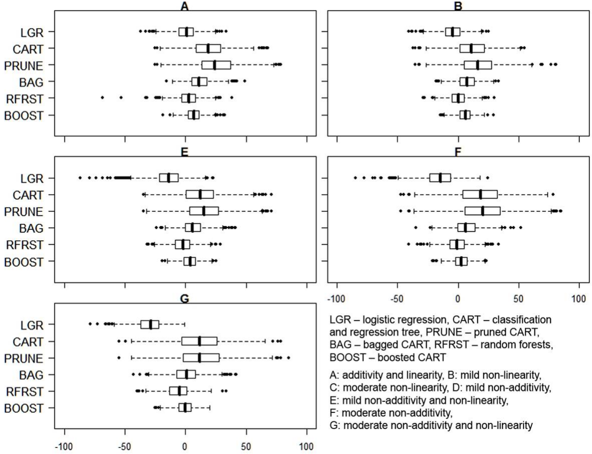
- Objective is not pure prediction
- Goal is to find scores that equalize groups

→ Usage of covariate balance as evaluation criterion (e.g. `twang` package)

1. Fix interaction depth and shrinkage
2. Consider a large (maximum) number of trees
3. Stop GBM interations when balance is optimized
    - e.g., average standardized absolute mean difference

Propensity Score Estimation

Figure: Percent bias by method for different simulation scenarios



(Lee et al. 2010)

Predicting Panel Drop-outs

Predicting Panel Drop-outs

Predicting Panel Drop-outs

The challenge: Nonresponse in panel studies

- Panel attrition reduces sample sizes and can introduce bias due to systematic dropout patterns
- Standard approach: Construct weights based on regression models

However, machine learning techniques can also be utilized...

- Modeling nonresponse and constructing weights (e.g. Buskirk & Kolenikov 2015)
- Predicting panel nonresponse (Kern 2017, Klausch 2017)

→ Potential of moving from post- to "pre-correction" of panel dropouts through ML?

Predicting Panel Drop-outs

German Socio-Economic Panel Study (2013–2014)
Sample: Respondents 2013 (mode $\neq$ by mail)

Table: Description of variables

(a) Outcome

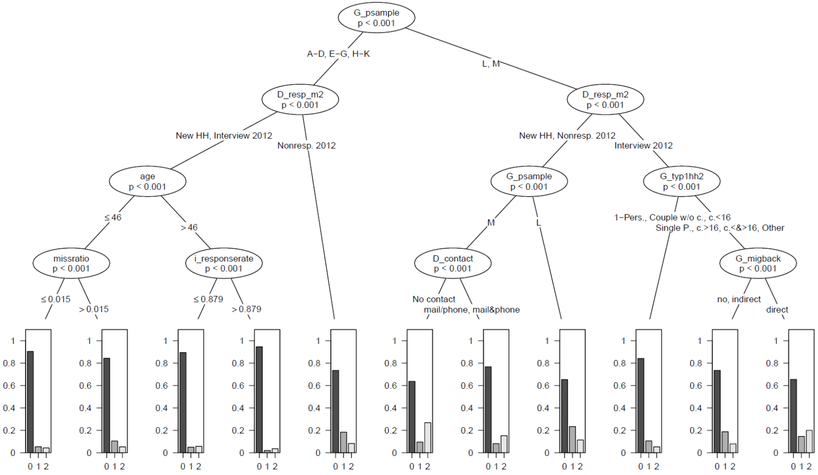
	Variable	Categories	Year
y_1	G_Response	Interview/temp. Ref./final. Ref.	2014
y_2	D_Response	Interview/temp. or final. Ref.	2014

(b) Features

Variables	Year
SOEP years	1984-2013
Interviewer Contacts	2013
Mode	2013
Refusal in HH	2013
Contact information	2013
Response	2012
Missing ratio (items)	2013
Interviewer: Gender, age, exp., RR, mean int. length	2013
SOEP Sample	1984-2013
Inverse Staying Probability	2013
Demographic variables	2013

Predicting Panel Drop-outs

Figure: Small `ctree` (training set; y_1)



Predicting Panel Drop-outs

Table: Confusion matrices (test set; y_1)

(a) Random Forest

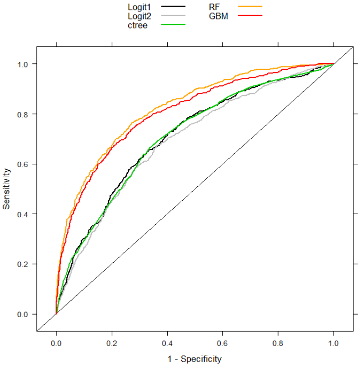
Prediction	Reference Interview	temp. Ref.	final Ref.	Sum
Interview	4224	295	278	4797
temp. Ref.	24	41	5	70
final Ref.	9	3	33	45
Sum	4257	339	316	4912
Sensitivity	0.992	0.121	0.104	
Precision	0.881	0.586	0.733	
Null Rate	0.8667			
Accuracy	0.875			
Kappa	0.178			

(b) Gradient Boosting

Prediction	Reference Interview	temp. Ref.	final Ref.	Sum
Interview	4184	276	239	4699
temp. Ref.	43	57	11	111
final Ref.	30	6	66	102
Sum	4257	339	316	4912
Sensitivity	0.983	0.168	0.209	
Precision	0.890	0.514	0.647	
Null Rate	0.8667			
Accuracy	0.8768			
Kappa	0.267			

Predicting Panel Drop-outs

Figure: ROC curves (test set; y_2)



	Accuracy	Kappa	Sens.	Spec.	AUC
Logit1	0.867	0.007	0.005	1.000	0.707
Logit2	0.855	0.102	0.101	0.971	0.691
ctree	0.866	0.010	0.008	0.998	0.708
RF	0.881	0.239	0.171	0.991	0.822
GBM	0.878	0.315	0.270	0.972	0.803

Predicting Panel Drop-outs

Preliminary conclusion

- Model specification
  - Conditional Inference Trees enable exploration of subgroups with high dropout risks
- Prediction
  - Ensemble methods (RF, GBM) outperform parametric models and single trees
  - However, accuracy of (current) RF and GBM only slightly above no information rate

$\rightarrow$ Potentially improved prediction with extended set of predictors in longitudinal setup

The Common Task Framework

Speech Recognition, Natural Language Processing and the Common Task Framework

The following discussion is based on Liberman (2015) on Donoho (2017)

Speech Recognition and the Common Task Framework

Speech Recognition (e.g. Amazon Echo) is essentially a prediction problem:

Speech (digital audio signal) $\rightarrow$ written words

- Difficult problem without obvious solution
- How to ensure that algorithms improve (and money is not spent for nothing)?

Speech Recognition and the Common Task Framework

Ingredients of the Common Task Framework:

- Training dataset with features and class label is publicly available
- Well-defined metric for evaluation
- Competing groups with the common task to infer a prediction rule from the data
- Automatic evaluation of prediction rules at the end of the competition on separate test data that is not published

Speech Recognition and the Common Task Framework

Ingredients of the Common Task Framework:

- Training dataset with features and class label is publicly available
- Well-defined metric for evaluation
- Competing groups with the common task to infer a prediction rule from the data
- Automatic evaluation of prediction rules at the end of the competition on separate test data that is not published

A culture developed around the CTF:

- Famous example: Netflix competition ($1 million for the winning team)
- `https://www.kaggle.com/` hosts several competitions on real data
- New algorithms are tested on published data with published algorithms at conferences

Speech Recognition and the Common Task Framework

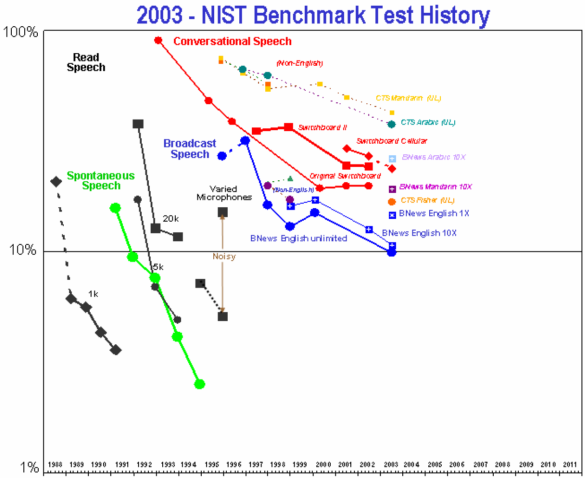


Figure 1 NIST Benchmark Test History

Speech recognition algorithms improved indeed. (Word Error Rates shown)

Speech Recognition and the Common Task Framework

Liberman (2015) summarizes the general experience with CTF:

> *1. Error rates decline by a fixed percentage each year, to an asymptote depending on task and data quality*
> *2. Progress usually comes from many small improvements; a change of 1% can be a reason to break out the champagne.*
> *3. Shared data plays a crucial role - and is re-used in unexpected ways.*

Most Machine learning products we know today benefited from the CTF framework!

References

Blancas Reyes, Eduardo, Helsby, Jennifer, Rasch, Katharina, van der Boor, Paul, Ghani, Rayid, Haynes, Lauren, Cunningham, Edward P. (2017).
Early detection of properties at risk of blight using spatiotemporal data.
http://doi.org/10.5281/zenodo.556510

Breidt, F. Jay, Opsomer, Jean (2017)
Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science* **32**(2). p. 190–205.

Buskirk, Trent D., Kolenikov, Stanislav (2015)
Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification. *Survey Insights: Methods from the Field*. http://surveyinsights.org/?p=5108.

Daumé, Hal (2017)
A Course in Machine Learning. http://ciml.info/

References

Chandler, Dana, Levitt, Steven D., List, John A. (2011).
Predicting and Preventing Shootings among At-Risk Youth. *American Economic Review* **101**(3). 288–292.

Donoho, David (2017)
50 Years of Data Science. *Journal of Computational and Graphical Statistics* **26**(4). p. 745–766.

Hainmueller, Jens, Hazlett, Chad (2014)
Kernel Regularized Least Squares: Reducing Misspecification Bias with a flexible and interpretable machine learning approach. *Political Analysis* **22**. p. 143–168

Kern, Christoph (2017)
Data-driven prediction of panel nonresponse. *ESRA Conference Presentation*.

Klausch, Thomas (2017)
Predicting panel attrition using panel-metadata: A machine learning approach. *ESRA Conference Presentation*.

References

Kleinberg, Jon, Lakkaraju, Himabindu, Leskovec, Jure, Ludwig, Jens, Mullainathan, Sendhil (2017)
Human Decisions and Machine Predictions. *NBER Working Paper No. 23180*.

Kleinberg, Jon, Ludwig, Jens, Mullainathan, Sendhil, Obermeyer, Ziad (2015)
Prediction Policy Problems. *American Economic Review* **105**(5). p. 491–495

Lakkaraju, Himabindu, Aguiar, Everaldo, Shan, Carl, Miller, David, Bhanpuri, Nasir, Ghani, Rayid, Addison, Kecia
A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Lee, Brian K., Lessler, Justin, Stuart, Elizabeth A. (2010)
Improving propensity score weighting using machine learning. *Statistics in Medicine* **29**(3). 337–346

References

Liberman, Mark (2015)
Reproducible Research and the Common Task Method. Simmons Foundation Lecture, April 1, 2015: `https://www.simonsfoundation.org/event/reproducible-research-and-the-common-task-method/`

McCaffrey, Daniel F., Ridgeway, Greg, Morral, Andrew R. (2004)
Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* **9**(4). 403–425.

Mullainathan, Sendhil, Spiess, Jann (2017)
Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* **31**(2). p. 87–106.

Pallett, David (2003)
A look at NIST'S benchmark ASR tests: past, present, and future. IEEE Workshop on Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003.

References

Ridgeway, Greg, McCaffrey, Daniel F., Morral, Andrew R., Griffin, Beth A., Burgette, Lane (2017).

twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. R package version 1.5. `https://CRAN.R-project.org/package=twang`

Rosenbaum, Paul R., Rubin, Donald B. (1983)

The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1). 41–55.

Schierholz, Malte, Gensicke, Miriam, Tschersich, Nikolai, Kreuter, Frauke (2018)

Occupation coding during the interview. *J. R. Statist. Soc. A* **181**. p. 379–407.