The Journal of Open Source Software

{discord}: An R Package for Discordant-Kinship Regressions

Jonathan D. Trattner[1] **and S. Mason Garrison**[2]

1 Department of Neuroscience, Wake Forest School of Medicine **2** Department of Psychology, Wake Forest University

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

As a field, behavior genetics studies the genetic and environmental sources of individual differences in psychological traits and characteristics. More technically, the field focuses on decomposing the sources of phenotypic variation into genetic (Additive (A)+ Dominance (D)) and environmental (Shared Environment (C) + Non-Shared Environment (E)) variance components, by leveraging twin and family studies. These models can do more than merely describe sources of variance; they can be used to infer causation (Burt, Plaisance, & Hambrick, 2019). Here, we present software to facilitate genetically-informed quasi-experimental designs primarily for kinship modeling. Specifically, it facilitates discordant-kinship regressions by comparing kin, such as siblings. These designs account for genetic-and-environmental variance when examining causal links in the realm of 'nature vs. nurture.'

Statement of Need

Kin-comparison designs distinguish "within-family variance" from "between-family variance" (Chamberlain & Griliches, 1975). Within-family variance indicates how individuals of a specific family differ from one another; the between-family variance reflects sources that make family members more similar to one another (**garrison2016?**). By partitioning these sources of variance, scholars may greatly reduce confounds when testing hypotheses (Lahey & D'Onofrio, 2010). Our R package, {discord}, has customizable, efficient code for generating genetically-informed simulations and provides user-friendly functions to help researchers use kin-based quasi-experimental designs.

{discord} augments the NlsyLinks R package, which provides kinship links for the National Longitudinal Surveys of Youth – a series of cross-generational, nationally representative surveys of over 30,000 participants (Beasley et al., 2016; Rodgers et al., 2016). It has been used in multiple studies (cite, Mason, cite!).

Mathematics

To facilitate kinship comparisons, {discord} implements a modified reciprocal standard dyad model (Kenny, Kashy, & Cook, 2006) known as the discordant-kinship model (see (**garrison2016?**) for an extension). Consider the simplified case where a behavioral outcome, Y, is predicted by variable, X. The discordant-kinship model relates the difference in the outcome, $Y_{i\Delta}$, for the ith kinship pair, where $\bar{Y}_i$ is the mean level of the outcome, $\bar{X}_i$ is the mean level of the predictor, and $X_{i\Delta}$ is the between-kin difference in the predictor.

$$Y_{i\Delta} = \beta_0 + \beta_1 \bar{Y}_i + \beta_2 \bar{X}_i + \beta_3 X_{i\Delta} + \epsilon_i$$

This model partitions variance in line with the above discussion to support causal inference. Specifically, the within-family variance is described by Y_Δ and X_Δ; between-family variance is captured by $\bar{Y}$ and $\bar{X}$ (Garrison & Rodgers, 2021).

A non-significant association between Y_Δ and X_Δ suggests that the variables are not causally related and are not apparent within a family where gene and shared-environmental factors are controlled. In contrast, a significant association may provide support for a causal relationship between variables depending on the relatedness of each kin pair. That is, the discordant-kinship model is applicable for any set of kin: monozygotic twins who share 100% of their DNA; full-siblings who share 50%; half-siblings who share 25%; cousins who share 12.5%; etc. Thus, a significant relationship found with monozygotic twins would provide stronger support for a causal claim than the same relationship between cousins.

Following (Garrison & Rodgers, 2021), we recommend interpreting significant associations as *not disproving a causal relationship*. Although this design controls for much (sibling) if not all (monozygotic twins) background heterogeneity, it is possible that a significant relationship between a phenotype and plausible covariates is possible due to non-shared environmental influences.

The next section illustrates two (HOW MANY?) examples of discordant-kinship regressions with the {discord} package.

Vaccine willingness and socioeconomic status

Introduction

The following analysis is a simple case based on work presented elsewhere (Trattner, Kennon, & Garrison, 2020). The original project was inspired by reports detailing health disparities among ethnic minorities during the COVID-19 pandemic (Hooper, Nápoles, & Pérez-Stable, 2020). Data came from the 1979 National Longitudinal Survey of Youth (NLSY79), a nationally representative household probability sample jointly sponsored by the U.S. Bureau of Labor Statistics and Department of Defense. Participants were surveyed annually from 1979 until 1994 at which point surveys occurred biennially. The data are publicly available at https://www.nlsinfo.org/ and include responses from a biennial flu vaccine survey administered between 2006 and 2016. Our work originally examined whether SES at age 40 is a significant predictor for vaccination rates using the discordant-kinship model.

The data for this analysis was downloaded with the NLS Investigator and can be found here. The SES at age 40 data can be found here. For clarity and to emphasize the functionality of {discord}, the data has been pre-processed using this script. This discordant-kinship analysis is possible thanks to recent work that estimated relatedness for approximately 95% of the NLSY79 kin pairs (Rodgers et al., 2016). These kinship links are included in the {NlsyLinks} R package (Beasley et al., 2016) and are easily utilized with the {discord} package.

Data Cleaning

For this example, we will load the following packages.

```r
# For easy data manipulation
library(dplyr)
# For kinship linkages
library(NlsyLinks)
# For discordant-kinship regression
library(discord)
# To clean data frame names
library(janitor)
# tidyup output
library(broom)
# pipe
library(magrittr)
```

After some pre-processing, we have a data frame containing subject identifiers, demographic information such as race and sex, and behavioral measurements like flu vaccination rates and SES at age 40. A random slice of this data looks like:

CASEID	RACE	SEX	FLU_total	S00_H40
381	1	1	2	58.16532
572	0	0	0	39.13405
562	0	1	3	80.17893
577	0	0	2	91.86056
382	1	0	0	38.82502
331	1	1	1	18.91381

Using the kinship relationships from the {NlsyLinks} package, we can create a data frame that lends itself to discordant analysis. For each kin pair, the function `CreatePairLinksSingleEntered()` takes a data set like the one above, [**a specification of the NLSY database and the kin's relatedness**], and the variables of interest. It returns a data frame where every row is a kin-pair and each column is a variable of interest with a suffix indicating to which individual the value corresponds.

For this example, we will examine the relationship between flu vaccinations received between 2006-2016 and SES at age 40 between full siblings. As such, we specify the following variables from the pre-processed data frame previewed above.

```r
# Get kinship links for individuals with the following variables:
link_vars <- c("FLU_total", "FLU_2008", "FLU_2010",
               "FLU_2012", "FLU_2014", "FLU_2016",
               "S00_H40", "RACE", "SEX")
```

We now link the subjects by the specified variables using `CreatePairLinksSingleEntered()`, from the {NlsyLinks} package.

```r
# Specify NLSY database and kin relatedness
link_pairs <- Links79PairExpanded %>%
  filter(RelationshipPath == "Gen1Housemates" & RFull == 0.5)

df_link <- CreatePairLinksSingleEntered(outcomeDataset = flu_ses_data,
                                        linksPairDataset = link_pairs,
                                        outcomeNames = link_vars)
```

We have saved this data frame as `df_link`. A random subset of this data is:[1]

ExtendedID	SubjectTag_S1	SubjectTag_S2	FLU_total_S1	FLU_total_S2	S00_H40_S1	S00_H40_S2
281	28100	28200	4	1	81.25000	78.51113
949	94900	95000	0	0	55.66002	46.23236
137	13700	13800	3	0	81.06099	76.36864
1184	118400	118500	3	4	85.50907	92.97715
272	27200	27300	5	3	83.25019	62.60165
1042	104200	104300	4	1	24.42666	19.51865

This data is almost ready for analysis, but we want to ensure that the data are representative of actual trends. The `FLU_total` column is simply a sum of the biennial survey responses. So for a given sibling-pair, one or both individuals may not have responded to the survey indicating their vaccination status. If that's the case, we want to exclude those siblings to reduce [**non-response bias**]. We can do this by examining the biennial responses and removing any rows with missingness.

```
# Take the linked data, group by the sibling pairs and
# count the number of responses for flu each year. If there is an NA,
# then data is missing for one of the years, and we omit it.
consistent_kin <- df_link %>%
  group_by(SubjectTag_S1, SubjectTag_S2) %>%
  count(FLU_2008_S1, FLU_2010_S1,
        FLU_2012_S1, FLU_2014_S1,
        FLU_2016_S1, FLU_2008_S2,
        FLU_2010_S2, FLU_2012_S2,
        FLU_2014_S2, FLU_2016_S2) %>%
  na.omit()

# Create the flu_modeling_data object with only consistent responders.
# Clean the column names with the {janitor} package.
flu_modeling_data <- semi_join(df_link,
                               consistent_kin,
                               by = c("SubjectTag_S1",
                                      "SubjectTag_S2")) %>%
  clean_names()
```

To avoid violating assumptions of independence, in our analysis we specify that the sibling-pairs should be from unique households (i.e. we randomly select one sibling pair per household).

```
flu_modeling_data <- flu_modeling_data %>%
  group_by(extended_id) %>%
  slice_sample() %>%
  ungroup()
```

The data we will use for modeling now contains additional information for each kin pair, including sex and race of each individual, flu vaccination status for the biennial survey

[1] Notice that, with the exception of the first column indicating the specific pair, each column name has the suffix "_S1" and "_S2." As mentioned above, these suffices identify which sibling the column values correspond.

between 2006-2016, and a total flu vaccination count for that period. The total vaccination count ranges from 0 - 5, where 0 indicates that the individual did not get a vaccine in any year between 2006-2016 and 5 indicates that an individual got at least 5 vaccines between 2006-2016. Although our data set has individual years, we were only interested in the aggregate as we felt that was a measure of general tendency. A subset of the data to use in this regression looks like:

extended_id	subject_tag_s1	subject_tag_s2	flu_total_s1	flu_total_s2	race_s1	race_s2	sex_s1	sex_s2	ses_age_40_s1	ses_age_40_s2
17	1700	1800	0	0	0	0	1	1	49.26537	74.92440
29	2900	3000	2	0	0	0	0	0	56.80481	32.05423
37	3700	3800	1	5	0	0	0	0	58.55547	50.45408
40	4000	4100	2	0	0	0	1	1	78.19220	73.41860
58	5800	5900	5	0	0	0	0	1	80.56835	49.68414
61	6100	6200	3	4	0	0	0	0	74.43720	50.56920
67	6700	6800	4	4	0	0	1	0	89.67767	82.68649
74	7500	7600	0	0	0	0	0	1	88.15524	61.54234
83	8300	8400	0	3	1	1	1	1	46.41507	64.12765
85	8500	8700	0	4	1	1	1	1	40.12748	64.14045

Modeling and Interpretation

To perform the regression using the {discord} package, we supply the data frame and specify the outcome and predictors. It also requires a kinship pair id, `extended_id` in our case, as well as pair identifiers – the column name suffixes that identify to which kin a column's values correspond ("_s1" and "_s2" in our case).[2] Optional, though recommended, are columns containing sex and race information to control for as additional covariates. In our case, these columns are prefixed "race" and "sex." Per the pre-processing script, these columns contain dummy variables where the reference group for race is "non-Black, non-Hispanic" and the reference group for sex is female.

By entering this information into the `discord_regression()` function, we can run the model as such:

```r
# Setting a seed for reproducibility
set.seed(18)
flu_model_output <- discord_regression(
                          data = flu_modeling_data,
                          outcome = "flu_total",
                          predictors = "s00_h40",
                          id = "extended_id",
                          sex = "sex",
                          race = "race",
                          pair_identifiers = c("_s1", "_s2")
                          )
```

The default output of `discord_regression()` is an `lm` object. The metrics for our regression can be summarized as follows:

Looking at this output, the intercept can be thought of as the average difference in outcomes between siblings, controlling for all other variables. That is, it looks like the average difference for two sisters of a non-minority ethnic background (the reference groups for sex and race) is approximately 1.4. The term `flu_total_mean` is essentially an extra component of the intercept that captures some non-linear trends and allows the difference score to change as a function of the average predictors. Here, this is the mean socioeconomic status for the siblings, `s00_h40_mean`. We also accounted for sex and race, neither of which have a statistically significant effect on the differences in flu vaccine shots between siblings (different families) or within a sibling pair (same family).

[2] Note these ids were previously "_S1" and "_S2," however, we used the `clean_names()` function which coerced the column names to lowercase.

Term	Estimate	Standard Error	T Statistic	P Value
(Intercept)	1.374	0.195	7.059	p<0.001
flu_total_mean	0.187	0.034	5.522	p<0.001
s00_h40_diff	0.005	0.002	2.421	p=0.016
s00_h40_mean	0.003	0.003	1.102	p=0.271
sex_1	-0.147	0.098	-1.491	p=0.136
race_1	-0.029	0.103	-0.281	p=0.779
sex_2	0.095	0.098	0.965	p=0.335

The most important metric from the output, though, is the difference score, `s00_h40_diff`. Here, it is statistically significant. An interpretation of this might be, "the difference in socioeconomic status between siblings at age 40 is positively associated with the difference in the number of flu vaccinations received between 2006-2016." This means that a sibling with 10% higher SES is expected to have 0.0491088 more flu shots.

The goal of performing a discordant-kinship regression is to see whether there is a significant difference in some behavioral measure while controlling for as much gene-and-environmental variance as possible. In this section, we walked through an analysis showing a statistically significant difference in the number of flu shots a sibling received and their socioeconomic status. From this, we *could not* claim the relationship is causal. However, we cannot eliminate causality because there are statistically significant within- and between-family differences in our predictors and outcomes.

Conclusion

In its current implementation, the {discord} package encourages best practices for performing discordant-kinship regressions. For example, the main function has the default expectation that sex and race indicators will be supplied. These measures are both important covariates when testing for causality between familial background and psychological characteristics.

This, and other design choices, are crucial to facilitating transparent and reproducible results. Software ever-evolves, however, and to further support reproducible research we plan to provide improved documentation and allow for easier inspection of the underlying model implementation and results.

Acknowledgements

We acknowledge contributions from Cermet Ream, Joe Rodgers, and support from Lucy D'Agostino McGowan on this project.

References

Beasley, W., Rodgers, J., Bard, D., Hunter, M., Garrison, S. M., & Meredith, K. (2016). *NlsyLinks: Utilities and kinship information for research with the NLSY*. Retrieved from https://CRAN.R-project.org/package=NlsyLinks

Burt, S. A., Plaisance, K. S., & Hambrick, D. Z. (2019). Understanding "What Could Be": A Call for 'Experimental Behavioral Genetics'. *Behavior Genetics*, *49*(2), 235–243. doi:10.1007/s10519-018-9918-y

Chamberlain, G., & Griliches, Z. (1975). Unobservables with a Variance-Components Structure: Ability, Schooling, and the Economic Success of Brothers. *International Economic Review*, *16*(2), 422–449.

Garrison, S. M., & Rodgers, J. L. (2021, June 28). *Using genetically-informed designs to test causal claims without experiments: Discordant-sibling designs and applications for differential psychology.*

Hooper, M. W., Nápoles, A. M., & Pérez-Stable, E. J. (2020). COVID-19 and Racial/Ethnic Disparities. *JAMA*. doi:10.1001/jama.2020.8598

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. Dyadic data analysis. New York, NY, US: Guilford Press.

Lahey, B. B., & D'Onofrio, B. M. (2010). All in the Family: Comparing Siblings to Test Causal Hypotheses Regarding Environmental Influences on Behavior. *Current Directions in Psychological Science*, *19*(5), 319–323. doi:10.1177/0963721410383977

Rodgers, J. L., Beasley, W. H., Bard, D. E., Meredith, K. M., Hunter, M. D., Johnson, A. B., Buster, M., et al. (2016). The NLSY kinship links: Using the NLSY79 and NLSY-children data to conduct genetically-informed and family-oriented research. *Behavior genetics*, *46*(4), 538–551. doi:10.1007/s10519-016-9785-3

Trattner, J., Kennon, L., & Garrison, S. M. (2020). Vaccine willingness and socioeconomic status: a biometrically controlled design. *Behavior Genetics*, Behavior Genetics Association 50th Annual Meeting Abstracts, *50*(6), 483–483. doi:10.1007/s10519-020-10018-8