Regression Code in R for Diagnosis Team Manuscripts

The purpose of this R code is to assess the risk (odds ratio) of autism and ADHD associated with exposure to physical and social exposome as part of the Equal-Life project. Major steps of this code include a) Data Preparation, b) Data Imputation, c) Descriptive statistics, and d) Regression.

1) You can use "master_data.csv" which was used to run RF. Copy the file into *your local disk/Logit_Code/Data* folder.
2) Open "Logistic_Regression.R".
3) On **line 7** write the pathway to your working directory. An example of a pathway is provided in **Figure 1**. The correct pathway to your working directory may be automatically detected.

```
setwd("C:/Users/libsul/R/EQL/Logit_code")
```

Figure 1 Example of pathway to a working directory.

4) Run **lines 12 to 18** to install packages and load libraries, load "maste_data.csv" and generate a "variable_type.xlsx" file.
5) In the "variable_type.xlsx", sheet "variable_type" you can change the variable type (categorical, numeric), inspect the number of unique values, and missing data in your data set, and select dependent, independent, and adjusting variables for the logistic regression (**Figure 2**).
    a. **Variables** = Names of all variables included in "maste_data.csv".
    b. **n_Uniques** = Number of unique values per variable.
    c. **Type** = type of each variable (guestimated).
        i. O = Ordered categorical.
        ii. U = Unordered categorical variable.
        iii. C = Continuous variable (a numerical variable).
    d. **Missing_Perc** = Percentage of missing data in "maste_data.csv".
    e. **Dependent_Variable** = Select which variable (write "1") is dependent in the model (outcome, e.g., autism), only <u>one</u> such variable can be selected.
    f. **Independent_Variables** = Select which variables (write "1") are independent in the model (exposures, e.g., NO_2, road length, population density). At least one variable must be selected.
    g. **Covariates** = Select which variables (write "1") are "covariates" (these are sex, mother's age at birth, child's year of birth, and mother's country of birth).
    h. Other named columns (e.g., "leave_empty1") are columns that could be used for an additional layer of adjustment. But these as their names suggest should be left empty.

A

Variable_Name	n_Uniques	Type	Missing_Perc	Dependent_Variable	Independent_Variables	Covariates	leave_empty1	leave_empty2	leave_empty3
X	10000	C	0						
child_id	10000	C	0						
Dob_ch_report	2152	C	0						
Sex_ch_r	2	U	0						
birthyear	6	U	0						
Agebirth_m_report	37	C	0						
father_age	55	C	1,17						
bmi	1570	C	3,68						
preg_smk_mr_r_t	2	U	0						
Edu_m_r_t	4	U	1,28						
Edu_f_r_t	4	U	24,83						
Hh_income_r_t	5	U	0,08						
occupcode_m	11	U	0						
occuphour_m	4	U	2,7						
occupcode_f	11	U	86,52						
Cob_m_r_t	2	U	0						
hh_adultsnr_r_t	3	U	0,08						
famsize_child	3	U	0,08						
loweight	3	U	0,25						
marital	2	U	0						
parity	4	U	0						
nbirths	1	U	0						
preterm	2	U	0						
ADHD	2	U	0						
ASD	2	U	0						

B

Outcome variable only (ADHD/ASD) ↘ | All other variables selected by the RF (see analysis protocol) ↘ | Only sex/COB/smoking/age ↘ | Please ignore these columns

Variable_Name	n_Uniques	Type	Missing_Perc	Dependent_Variable	Independent_Variables	Covariates	leave_empty1	leave_empty2	leave_empty3
X	10000	C	0						
child_id	10000	C	0						
Dob_ch_report	2152	C	0						
Sex_ch_r	2	U	0			1			
birthyear	6	O	0			1			
Agebirth_m_report	37	C	0			1			
father_age	55	C	1,17						
bmi	1570	C	3,68						
preg_smk_mr_r_t	2	U	0		1				
Edu_m_r_t	4	O	1,28		1				
Edu_f_r_t	4	O	24,83						
Hh_income_r_t	5	O	0,08						
occupcode_m	11	U	0						
occuphour_m	4	U	2,7						
occupcode_f	11	U	86,52						
Cob_m_r_t	2	U	0			1			
hh_adultsnr_r_t	3	C	0,08		1				
famsize_child	3	C	0,08						
loweight	3	U	0,25						
marital	2	U	0						
parity	4	U	0						
nbirths	1	U	0						
preterm	2	U	0						
ADHD	2	U	0	1					
ASD	2	U	0						

Figure 2 (A) "variable_type.xlsx" file, sheet "variable_type" as generated by the R code, (B) "variable_type.xlsx" file, sheet "variable_type" with changed variable types and selected ADHD as the dependent variable, preg_smk_mr_r_t, Edu_m_r_t, hh_adultsnr_r_t as independent variables, and Sex_ch_r, birthyear, Agebirth_m_report, and Cob_m_r_t as covariates.

6) In the "variable_type.xlsx", sheet "desc_stat" you can assign variables for which will be generated descriptive statistic (pre-selected according to Table 1 of the ADHD manuscript analysis protocol). Copy-paste variable names included in the sheet "variable_type" in the appropriate cells. Cells for variables that are not available in your cohort can be left empty (**Figure 3**).

Note: The variable coding is expected to follow a harmonisation protocol (e.g., sex: 0=boys, 1=girls, smoking: 0=no smoking, 1=smoking).

Parameter	Variable_Name
ADHD or Autism variable	
Sex of the child	
Age of child at diagnosis	
Mothers age at birth	
Fathers age at birth	
Mothers BMI	
Smoking during pregnancy	
Income variable	
Mothers education	
Mothers county of birth	
Fathers education	
Fathers country of birth	

A

Parameter	Variable_Name
ADHD or Autism variable	ADHD
Sex of the child	Sex_ch_r
Age of child at diagnosis	age_diag_ADHD
Mothers age at birth	Agebirth_m_report
Fathers age at birth	father_age
Mothers BMI	bmi
Smoking during pregnancy	preg_smk_mr_r_t
Income variable	Hh_income_r_t
Mothers education	Edu_m_r_t
Mothers county of birth	Cob_m_r_t
Fathers education	Edu_f_r_t
Fathers country of birth	

B

Figure 3 (A) "variable_type.xlsx" file, sheet "desc_stat" as generated by the R code, (B) "variable_type.xlsx" file, sheet "desc_stat" with copy-pasted variable names. Cells can be left empty for variables not available in your study.

7) In the "variable_type.xlsx", sheet "sex_differences" you can assign variables which will be used for sex differences analyses (logistic and linear regression). Copy-paste variable names included in the sheet "variable_type" in the appropriate cells. Cells for variables that are not available in your cohort can be left empty (**Figure 4**).

Variable	Variable_Name
Autism/ADHD	
Child Sex	
Mothers country of birth	
Mothers education	
Pop. density (GHS)	
Child birthyear	
Child age at diagnosis	
Child age at the last measurement	

A

Variable	Variable_Name
Autism/ADHD	ADHD
Child Sex	Sex_ch_r
Mothers country of birth	Cob_m_r_t
Mothers education	Edu_m_r_t
Pop. density (GHS)	
Child birthyear	birthyear
Child age at diagnosis	age_diag_ADHD
Child age at the last measurement	

B

Figure 4 (A) "variable_type.xlsx" file, sheet "sex_differences" as generated by the R code, (B) "variable_type.xlsx" file, sheet "sex_differences" with copy-pasted variable names. Cells can be left empty for variables not available in your study.

8) If you are happy with the selection, save "variable_type.xlsx" and proceed further.
9) On **line 44** of the code, you can change the maximum fraction of missing data (imp_max) for which the imputation will be done. The default is set up at 0.25 (= imputation will be done **only** for variables with a maximum of **25%** of missing data).
10) On **line 46** of the code, you can change the seed for the imputation (seed).
11) On **line 48** of the code, you can choose to generate a correlation between all imputed data ("ON") or disable this feature ("OFF").
12) Run **line 50** of the code to initiate imputation (it will take time based on the size of your data).
13) On **line 61** of the code, write the name (analysis_run) of your analysis (e.g., ADHD, Autism, analysis1).

14) On **line 63** of the code, you can choose to generate correlations between variables used in the analyses ("ON") or disable this feature ("OFF"). By default, this is set to "OFF".

Note: If you disable correlation between imputed variables (line 48) and correlations between variables used in the analysis (line 63) no correlations will be generated.

15) You can run **line 65** which will calculate descriptive statistics. This will be done only if you previously assigned correct variable names (see step 6).
16) On **line 70** of the code, you can choose to generate linearity plots ("ON") or disable this feature ("OFF").
17) Now you can run **line 72** of the code. This will run logistic regression models based on the variables selected in "variable_type.xlsx". Running **line 75** will print the model results with all variables in the console.
18) All results will be saved in *your local disk/Logit_Code/Results*.
19) You can freely change the variable selection in the "variable_type.xlsx" file, sheet "variable_type" and re-run line 75 of the code until you are satisfied with the model results and fit.

Note: Each time you re-run line 72 (logistic regression) the results will be generated and if there already are results, these will be overwritten. If you want to keep multiple versions of the same analysis, please change the "analysis_run" variable (line 61) and run also descriptive statistics (line 65) each time you want to run the regression again.

20) You can run **line 83** which will run logistic and linear regression for the sex differences paper. This will be done only if you previously assigned correct variable names (see step 7). This code works with a specific selection of variables and therefore **can be run only once**.

Result files

Results generated according to steps 1 to 17 are saved in: *your local disk/Logit_Code/Results*/name of your analysis run.

There are two folders:

1) Logistic = results for the diagnosis team logistic regression paper.
2) Sex_Differences = results for the sex differences paper.

The following files can be found in the "Logistic" folder:

A) descriptives.xlsx (include descriptive statistics results and correlations of the non-imputed data set).
B) fit.xlsx (include the goodness of fit and multicollinearity test).
C) regression.xlsx (include regression model results).
D) variable_type.xlsx (copy of the variable type Excel file).
E) folder Linearity (include linearity visualisation).
F) folder ROC_AUC (include ROC curve and area under the curve).

The following files can be found in the "Sex_Differences" folder:

A) descriptives.xlsx (include descriptive statistics results and correlations of the non-imputed data set).
B) linear_multicol.xlsx (include multicollinearity test for linear regression).
C) linear_regr.xlsx (include linear regression model results, **Figures 5** and **6**).
D) logit_fit.xlsx (include the goodness of fit and multicollinearity test for the logistic regression).
E) logit_regr.xlsx (include logistic regression model results).
F) folder Linearity (include linearity visualization).
G) folder LM_assump (include visualization of linear model's assumptions).
H) folder ROC_AUC (include ROC curve and area under the curve).

In addition, the correlation matrix on the imputed data set is also included in *your local disk/Logit_Code/Results/*correlation_imputed_data.xlsx.

Variable	Formula	Estimate	Std.Error	p_value	sig	OR	CI_low	CI_high	Variance-Covariance Matrix	X.Intercept.	RT_LDEN_BY	res_NO2_mr_t	Sex_ch_r1	bmi	preg_smk_mr_r_t1
(Intercept)	ADHD ~ RT_LDEN_BY	-3,187963376	0,705206241	6,90656E-06	****	0,041256	0,010356	0,1643496	->	0,49732	-0,00848				
RT_LDEN_BY	ADHD ~ RT_LDEN_BY	0,011460091	0,012205425	0,347992199	ns	1,011526	0,987615	1,0360161	->	-0,00848	0,000149				
(Intercept)	ADHD ~ res_NO2_mr_t	-2,756344064	0,356189908	2,46272E-14	****	0,063524	0,031604	0,1276821	->	0,12687		-0,00852			
res_NO2_mr_t	ADHD ~ res_NO2_mr_t	0,016525688	0,025438637	0,516081387	ns	1,016663	0,967215	1,0686385	->	-0,00852		0,000647			
(Intercept)	ADHD ~ Sex_ch_r	-2,06664819	0,138184931	8,98679E-46	****	0,126609	0,09657	0,1659937	->	0,0191			-0,0191		
Sex_ch_r1	ADHD ~ Sex_ch_r	-1,427692523	0,304454443	3,12242E-06	****	0,239862	0,13207	0,4356301	->	-0,0191			0,092693		
(Intercept)	ADHD ~ bmi	-3,260888739	0,680212181	1,89459E-06	****	0,038354	0,010111	0,1454864	->	0,46269				-0,01829	
bmi	ADHD ~ bmi	0,029651804	0,027326636	0,278156052	ns	1,030096	0,976375	1,0867722	->	-0,01829				0,000747	
(Intercept)	ADHD ~ preg_smk_mr_r_t	-2,540476501	0,124144299	5,73586E-78	****	0,078829	0,061803	0,1005446	->	0,01541					-0,01541
preg_smk_mr_r_t1	ADHD ~ preg_smk_mr_r_t	-0,024472857	0,611870193	0,968103685	ns	0,975824	0,294129	3,2374718	->	-0,01541					0,374385

Figure 5 Example of univariate logistic regression models.

Models are “glued” under each other and include:

Variable: The variable in question in the given model

Formula: formula (set of variables) used in each respective model

Estimate: Beta coefficient

Std. Error, **p_value**, **Variance-Covariance matrix**: self-explanatory…

OR: odds ratio based on Estimate and increment of change (1)

Lower CI and **Upper CI**: 95% confidence interval accompanying OR

Sig. Level: Significance level (p-value) expressed as * where:

ns	not significant
*	p-value ≤0.05 & >0.01
**	p-value ≤0.01 & >0.001
***	p-value ≤0.001 & >0.0001
****	p-value ≤0.0001

Variable	Formula	Estimate	Std.Error	p_value	sig	OR	CI_low	CI_high	Variance-Covariance Matrix	X.Intercept.	RT_LDEN_BY	res_NO2_mr_t	Sex_ch_r1	bmi	preg_smk_mr_r_t1
(Intercept)	ADHD~RT_LDEN_BY+res_NO2_mr_t+Sex_ch_r+bmi+preg_smk_mr_r_t	-3,553493	1,0525247	0,000764	***	0,028624	0,003638	0,22525	->	1,107808	-0,01116	0,004207	-0,01852	-0,02085	0,019093
RT_LDEN_BY	ADHD~RT_LDEN_BY+res_NO2_mr_t+Sex_ch_r+bmi+preg_smk_mr_r_t	0,0125694	0,01722	0,465606	ns	1,012649	0,979041	1,04741	->	-0,01116	0,000297	-0,00044	-5,5E-05	2,69E-06	-0,00019
res_NO2_mr_t	ADHD~RT_LDEN_BY+res_NO2_mr_t+Sex_ch_r+bmi+preg_smk_mr_r_t	-0,000509	0,036714	0,988939	ns	0,999491	0,930095	1,074065	->	0,004207	-0,00044	0,001348	0,000195	0,000117	0,001298
Sex_ch_r1	ADHD~RT_LDEN_BY+res_NO2_mr_t+Sex_ch_r+bmi+preg_smk_mr_r_t	-1,426847	0,3048049	3,25E-06	****	0,240065	0,132091	0,436298	->	-0,01852	-5,5E-05	0,000195	0,092906	-1,1E-05	0,004571
bmi	ADHD~RT_LDEN_BY+res_NO2_mr_t+Sex_ch_r+bmi+preg_smk_mr_r_t	0,0325773	0,0280251	0,245346	ns	1,033114	0,977896	1,091449	->	-0,02085	2,69E-06	0,000117	-1,1E-05	0,000785	-0,00174
preg_smk_mr_r_t1	ADHD~RT_LDEN_BY+res_NO2_mr_t+Sex_ch_r+bmi+preg_smk_mr_r_t	-0,151104	0,6265232	0,809467	ns	0,859759	0,251808	2,935512	->	0,019093	-0,00019	0,001298	0,004571	-0,00174	0,392531

Figure 6 Example of multivariate logistic regression (include all variables).