I had made some changes, mostly based on Alaitz suggestions. The dependent variable bmi is not log transformed and not standardized, while the independent variables (the ones being analyzed) are categorized according to the recommendations, where only those variables are included for which specific recommended cut points are available.

Beta coefficients are calculated, using the categorized variables in an independent dataset and used to multiply the categories in our subset. Beta coefficients are also calculated in our subset to check that the direction is the same as in the independent dataset.
Note, that simple multiplication of categories with the coefficients will result in negative scores, which might lead to some weird interpretations, for example if we get negatively associated sugar, it will not only mean that people who eat more sugar have a lower risk for obesity, but also that it lowers their obesogenic score, implying that sugar reduces the risk from other variables…The effect of that is dependent on the size of the beta coefficient, if it is really small, then the resulting score will be very small and will almost have no effect on the final score, which is the same as excluding that variable in the first place.

Alaitz mentioned the paper Sarkisian at al. 2010 , where the risk score was also constructed by multiplication with the beta coefficient, but their beta coefficients were all positive as they had expected and they had used an arbitrary number to multiply the coefficients with, in order to get a resulting score of 1 or 2. At the moment we are using categories from 0 to 2. So zero will remain and 1 and 2 will get multiplied with the coefficient. We still need to discuss what to do with those variables where there are only two categories.

At this point we are talking about working with variables expressed in % of total energy intake (except for fiber and salt) and categorized based on guidelines, but nevertheless, I had put together details of several associations, just to get a clearer picture.
I checked the association with bmi, using:

- raw continuous variables(log transformed and standardized), fitted together or separately,

- continuous variables expressed in % of total energy intake, fitted together or separately,

- variables categorized based on guidelines, also fitted together or separately (most of variables have to be expressed in % of TEI to be able to apply recommendations, except for fiber and salt).

- I had also made another attempt to create a more robust risk score, by categorizing variables based on the location in the standard normal distribution. The details and results for that at the end of document.

Before looking at the different outcomes, I had checked for collinearity between the diet variables and there is plenty. I guess this is expected, since the final macro nutrients were extracted from the same food item self reports. This is a problem if fitting all variables together.

Regarding the categorizing each variable based on the guidelines, since the focus is on the obesogenic type of unhealthy environment, 0 score is given to subjects below the recommended intake of the nutrient, 1 to those that are in the limits of the recommended intake and 2 to those that are above the upper limit of the recommended nutrient intake.
There are variables where the guidelines will only recommend one limit. In those cases, there can only be two categories and we will have decide how to code for this, for now I assign only 1 and 2, without the 0, that is why some variables will have no subjects in the category 0.

I have marked some numbers/variables:

	insignificant
	wrong direction of association
	direction of association different in other data or model

The independent dataset of Swedish only (47107 subjects):

Pairwise Pearson correlation coefficient for all variables(raw, as in g per day) included in the model:

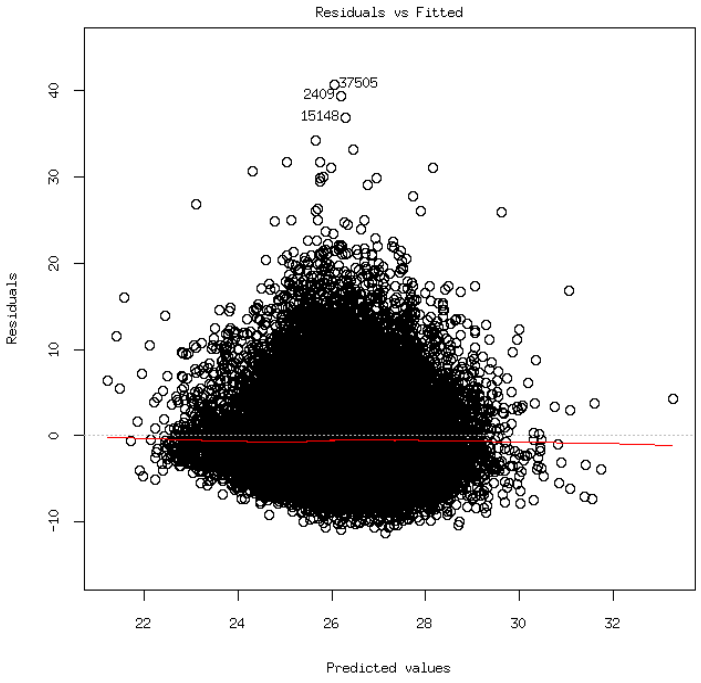
```
rcorr( cbind(bmi,POLYsum1_transformed,MONOsum1_transformed,mfetsum1_transformed,fettsum1_transformed,sacksum1_transformed,kolhsum1_transformed,FA_transformed,protsum1_transformed,fibesum1_transformed,NATRsum1_transformed),type="pearson")
```

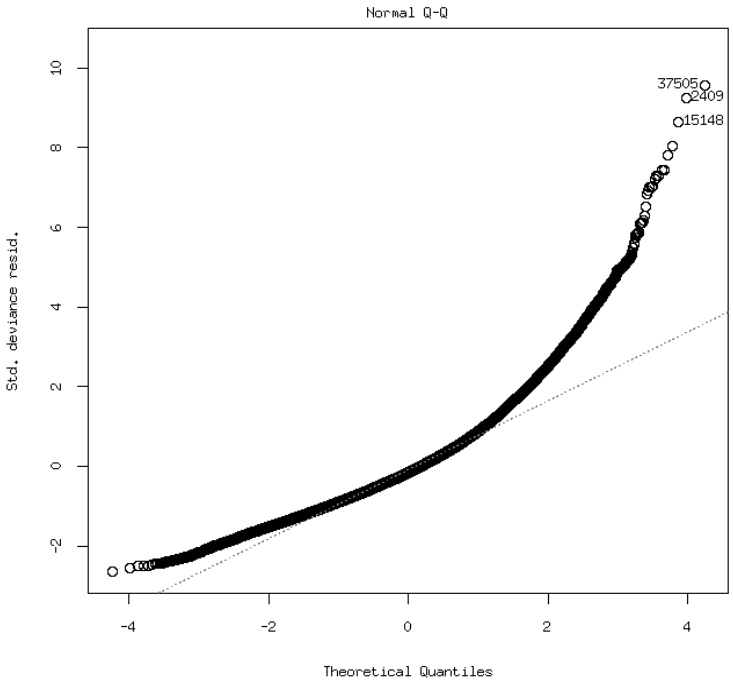
	bmi	POLYsum1	MONOsum1	mfetsum1	fettsum1	sacksum1	kolhsum1	FA	protsum1	fibesum1	NATRsum1
bmi	1										
POLYsum1	0.07	1									
MONOsum1	0.09	0.80	1								
mfetsum1	0.05	0.63	0.91	1							
fettsum1	0.07	0.80	0.97	0.96	1						
sacksum1	-0.02	0.23	0.29	0.30	0.32	1					
kolhsum1	0.02**	0.42	0.45	0.44	0.48	0.70	1				
FA	0.05	0.98	0.74	0.57	0.75	0.24	0.42	1			
protsum1	0.10	0.60	0.73	0.69	0.74	0.34	0.68	0.56	1		
fibesum1	0.00 i.s.	0.35	0.24	0.19	0.26	0.41	0.78	0.36	0.52	1	
NATRsum1	0.10	0.68	0.81	0.71	0.79	0.32	0.63	0.64	0.88	0.48	1

all p-values < 2e-16, except where marked

Raw continuous variables, fitted together by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_transformed + MONOsum1_transformed + mfetsum1_transformed + fettsum1_transformed + sacksum1_transformed + kolhsum1_transformed + FA_transformed + protsum1_transformed + fibesum1_transformed + NATRsum1_transformed, family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.5990539	0.000359	5.21e-05
MONOsum1	0.4934630	0.000418	1.24e-05
mfetsum1	-1.6356688	0.003114	< 2e-16
fettsum1	1.1059859	0.000555	4.79e-07
acids	-0.0175674	0.001208	1.10e-13
kolhsum1	-0.3281385	0.000748	5.14e-09
sacksum1	-0.9322543	7e-06	0.585453
protsum1	0.5610824	0.002982	< 2e-16
NATRsum1	0.3152880	0.000718	1.04e-08
fibesum1	-0.2264846	0.00073	7.88e-09





Note that the deviation from normality is now bigger, since the dependent variable was not log transformed, the p-values could be subjected to invalidity.

Raw continuous, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.2040732	0.001759	< 2e-16
MONOsum1	0.3031544	0.003339	< 2e-16
mfetsum1	0.0614008	0.000147	0.0097074
fettsum1	0.1881207	0.001329	< 2e-16
acids	0.1363734	0.000809	< 2e-16
kolhsum1	-0.040163	7.3e-05	0.0677203
sacksum1	-0.1268329	0.000767	< 2e-16
protsum1	0.3610829	0.005985	< 2e-16
NATRsum1	0.4121994	0.007223	< 2e-16
fibesum1	-0.0541303	0.000149	0.009225

Continuous expressed in % of TEI, fitted together by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_ofTEI_transformed + MONOsum1_ofTEI_transformed + mfetsum1_ofTEI_transformed + fettsum1_ofTEI_transformed + sacksum1_ofTEI_transformed + kolhsum1_ofTEI_transformed + FA_ofTEI_transformed + protsum1_ofTEI_transformed + fibesum1_ofTEI_transformed + NATRsum1_ofTEI_transformed, family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.4800456	0.000457	4.94e-06
MONOsum1	0.2430660	0.00031	0.000171
mfetsum1	-0.9320629	0.002762	< 2e-16
fettsum1	0.7560991	0.000814	1.08e-09
acids	-0.7062288	0.001276	2.28e-14
kolhsum1	0.0445714	1.5e-05	0.406487
sacksum1	0.0016700	0	0.950096
protsum1	0.3475617	0.003203	< 2e-16
NATRsum1	0.1723705	0.000773	2.84e-09
fibesum1	-0.1721206	0.000654	4.67e-08

Continuous expressed in % of TEI, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.1588141	0.001181	< 2e-16
MONOsum1	0.3235834	0.004089	< 2e-16
mfetsum1	-0.0412288	7.7e-05	0.0609774
fettsum1	0.1415299	0.000853	< 2e-16
acids	0.0759219	0.000279	0.0003564
kolhsum1	-0.281902	0.003236	< 2e-16
sacksum1	-0.2124712	0.002211	< 2e-16
protsum1	0.4837918	0.011704	< 2e-16
NATRsum1	0.5100663	0.012844	< 2e-16
fibesum1	-0.191649	0.001513	< 2e-16

Categorized according to the recommended cut points, fitted together by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_ofTEI_categorized_g + MONOsum1_ofTEI_categorized_g + mfetsum1_ofTEI_categorized_g + fettsum1_ofTEI_categorized_g + sacksum1_ofTEI_categorized_g + kolhsum1_ofTEI_categorized_g + FA_ofTEI_categorized_g + protsum1_ofTEI_categorized_g + fibesum1_categorized_g + NATRsum1_categorized_g, family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.1099463	0.000154	0.00810
MONOsum1	0.3012363	0.000467	3.87e-06
mfetsum1	-0.6099049	0.000954	4.08e-11
fettsum1	0.1496284	0.000146	0.00974
acids	0.0630988	8e-06	0.53503
kolhsum1	-0.0949347	6.1e-05	0.09418
sacksum1	0.0826964	2.3e-05	0.30691
protsum1	1.1729414	0.002883	< 2e-16
NATRsum1	0.4858922	0.001805	3.73e-05
fibesum1	0.2635006	0.000373	< 2e-16

Categorized according to the recommended cut points, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.2215384	0.000743	< 2e-16
MONOsum1	0.3588432	0.000943	< 2e-16
mfetsum1	-0.1572909	8.6e-05	0.0471491
fettsum1	0.2623853	0.000851	< 2e-16
acids	-0.1876798	8.1e-05	0.0541129
kolhsum1	-0.3118413	0.00124	< 2e-16
sacksum1	-0.159644	9e-05	0.0424843
protsum1	1.2827382	0.00356	< 2e-16
NATRsum1	0.487762	0.002123	< 2e-16
fibesum1	0.1398369	0.000132	0.0141008

The 2 visits subset of Swedish only (33114 subjects in each visit),
visit==1 (33114 subjects):

Pairwise Pearson correlation coefficient for all variables(raw, as in g per day) included in the model:

```
rcorr( cbind(bmi,POLYsum1_transformed,MONOsum1_transformed,mfetsum1_transformed,fettsum1_transformed,sacksum1_transformed,kolhsum1_transformed,FA_transformed,protsum1_transformed,fibesum1_transformed,NATRsum1_transformed),type="pearson")
```

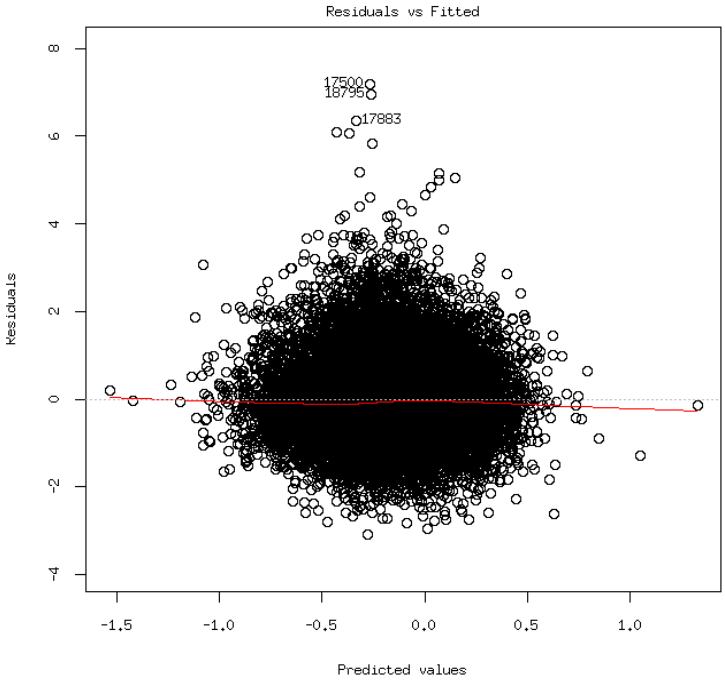
	bmi	POLYsum1	MONOsum1	mfetsum1	fettsum1	sacksum1	kolhsum1	FA	protsum1	fibesum1	NATRsum1
bmi	1										
POLYsum1	0.07	1									
MONOsum1	0.08	0.80	1								
mfetsum1	0.01**	0.67	0.89	1							
fettsum1	0.04	0.82	0.94	0.97	1						
sacksum1	-0.03	0.32	0.39	0.42	0.43	1					
kolhsum1	0.03	0.56	0.59	0.57	0.61	0.68	1				
FA	0.07	0.98	0.75	0.62	0.77	0.33	0.56	1			
protsum1	0.09	0.62	0.72	0.71	0.74	0.40	0.79	0.60	1		
fibesum1	0.03	0.42	0.33	0.28	0.34	0.35	0.78	0.44	0.58	1	
NATRsum1	0.09	0.71	0.81	0.74	0.80	0.39	0.74	0.69	0.89	0.54	1

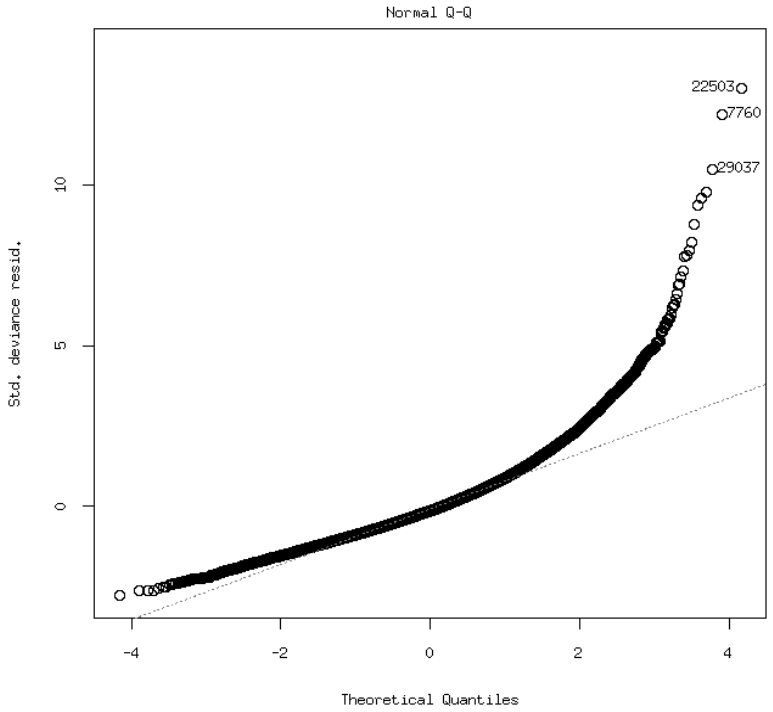
all p-values < 2e-16, except where marked

Raw continuous variables, fitted together by:

```
glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_transformed + MONOsum1_transformed + mfetsum1_transformed + fettsum1_transformed + sacksum1_transformed + kolhsum1_transformed + FA_transformed + protsum1_transformed + fibesum1_transformed + NATRsum1_transformed, family = gaussian(link = "identity"))
```

	Regression coefficient	Variance explained	p-value
POLYsum1	4.329e-01	0.000378	0.000522
MONOsum1	9.475e-02	4.2e-05	0.246351
mfetsum1	-1.798e+00	0.004927	< 2e-16
fettsum1	1.497e+00	0.001505	4.42e-12
acids	-7.683e-01	0.001663	3.40e-13
kolhsum1	-3.002e-01	0.000701	2.33e-06
sacksum1	-1.583e-02	7e-06	0.635929
protsum1	6.007e-01	0.004119	< 2e-16
NATRsum1	1.934e-01	0.000342	0.000978
fibesum1	-3.677e-02	2.7e-05	0.354084





Raw continuous, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.1060793	0.000685	3e-06
MONOsum1	0.0781947	0.000323	0.0013459
mfetsum1	-0.1051945	0.000634	7e-06
fettsum1	0.0025845	0	0.9135259
acids	0.0809107	0.00041	0.0003029
kolhsum1	0.0303871	6.5e-05	0.1504342
sacksum1	-0.1086266	0.000842	2e-07
protsum1	0.2449353	0.004022	< 2e-16
NATRsum1	0.276346	0.004523	< 2e-16
fibesum1	0.0572951	0.000257	0.0042038

Continuous expressed in % of TEI, fitted together by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_ofTEI_transformed + MONOsum1_ofTEI_transformed + mfetsum1_ofTEI_transformed + fettsum1_ofTEI_transformed + sacksum1_ofTEI_transformed + kolhsum1_ofTEI_transformed + FA_ofTEI_transformed + protsum1_ofTEI_transformed + fibesum1_ofTEI_transformed + NATRsum1_ofTEI_transformed, family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	3.242e-01	0.000492	7.56e-05
MONOsum1	4.483e-02	3e-05	0.329414
mfetsum1	-9.775e-01	0.004545	< 2e-16
fettsum1	8.942e-01	0.001775	5.54e-14
acids	-5.353e-01	0.001746	8.91e-14
kolhsum1	8.690e-02	6.9e-05	0.139751
sacksum1	-1.156e-02	6e-06	0.666132
protsum1	3.347e-01	0.004085	< 2e-16
NATRsum1	9.843e-02	0.000383	0.000481
fibesum1	-3.314e-02	3.6e-05	0.283955

Continuous expressed in % of TEI, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.0856582	0.000535	3.7e-05
MONOsum1	0.0465608	0.000133	0.0397995
mfetsum1	-0.2083109	0.0031	< 2e-16
fettsum1	-0.072357	0.000367	0.0006278
acids	0.0539485	0.000216	0.0086657
kolhsum1	-0.0345755	8.5e-05	0.0998065
sacksum1	-0.1588909	0.001899	< 2e-16
protsum1	0.3724594	0.010604	< 2e-16
NATRsum1	0.3768771	0.010549	< 2e-16
fibesum1	0.0367767	8.7e-05	0.0957426

Categorized according to the recommended cut points, fitted together by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_ofTEI_categorized_g + MONOsum1_ofTEI_categorized_g + mfetsum1_ofTEI_categorized_g + fettsum1_ofTEI_categorized_g + sacksum1_ofTEI_categorized_g + kolhsum1_ofTEI_categorized_g + FA_ofTEI_categorized_g + protsum1_ofTEI_categorized_g + fibesum1_categorized_g + NATRsum1_categorized_g, family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	6.525e-02	7.7e-05	0.117809
MONOsum1	4.347e-02	1.7e-05	0.465114
mfetsum1	-5.020e-01	0.000869	1.44e-07
fettsum1	-1.820e-01	0.000241	0.005616
acids	-1.728e-01	7.6e-05	0.119654
kolhsum1	-2.162e-01	0.000418	0.000263
sacksum1	-1.266e-01	0.000125	0.046424
protsum1	8.620e-01	0.001104	3.05e-09
NATRsum1	3.097e-01	0.001179	9.05e-10
fibesum1	5.002e-02	2.3e-05	0.392393

Categorized according to the recommended cut points, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.0907435	0.000175	0.0182951
MONOsum1	-0.0126073	2e-06	0.8064681
mfetsum1	-0.4241771	0.000848	2e-07
fettsum1	-0.0809161	9.9e-05	0.0765824
acids	-0.2191146	0.000131	0.0408488
kolhsum1	-0.110991	0.000213	0.0092061
sacksum1	-0.2487395	0.000513	5.33e-05
protsum1	0.9824838	0.00148	< 2e-16
NATRsum1	0.3305855	0.001572	< 2e-16
fibesum1	-0.1027124	0.000123	0.0476973

The 2 visits subset (66228 subjects),
visit==2 (33114 subjects):

Pairwise Pearson correlation coefficient for all variables(raw, as in g per day) included in the model, bmi added for comparison:

```
rcorr( cbind(bmi,POLYsum1_transformed,MONOsum1_transformed,mfetsum1_transformed,fettsum1_transformed,sacksum1_transformed,kolhsum1_transformed,FA_transformed,protsum1_transformed,fibesum1_transformed,NATRsum1_transformed),type="pearson")
```

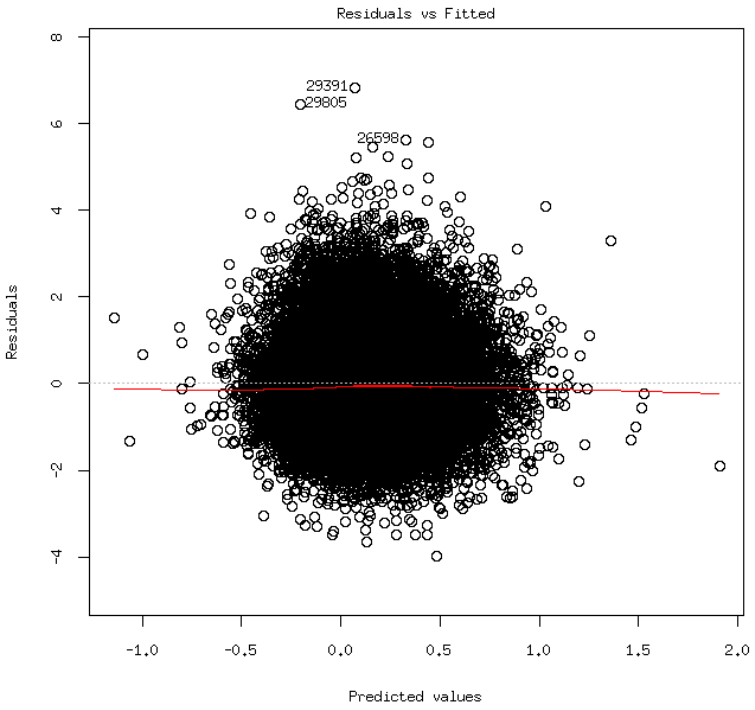
	bmi	POLYsum1	MONOsum1	mfetsum1	fettsum1	sacksum1	kolhsum1	FA	protsum1	fibesum1	NATRsum1
bmi	1										
POLYsum1	0.08	1									
MONOsum1	0.11	0.80	1								
mfetsum1	0.08	0.62	0.92	1							
fettsum1	0.10	0.79	0.98	0.96	1						
sacksum1	-0.02**	0.28	0.34	0.33	0.36	1					
kolhsum1	0.03	0.46	0.49	0.46	0.50	0.70	1				
FA	0.06	0.98	0.74	0.57	0.74	0.29	0.46	1			
protsum1	0.13	0.59	0.73	0.69	0.74	0.36	0.70	0.56	1		
fibesum1	0.00 i.s.	0.39	0.26	0.20	0.28	0.40	0.78	0.40	0.54	1	
NATRsum1	0.13	0.67	0.81	0.69	0.77	0.34	0.65	0.63	0.88	0.50	1

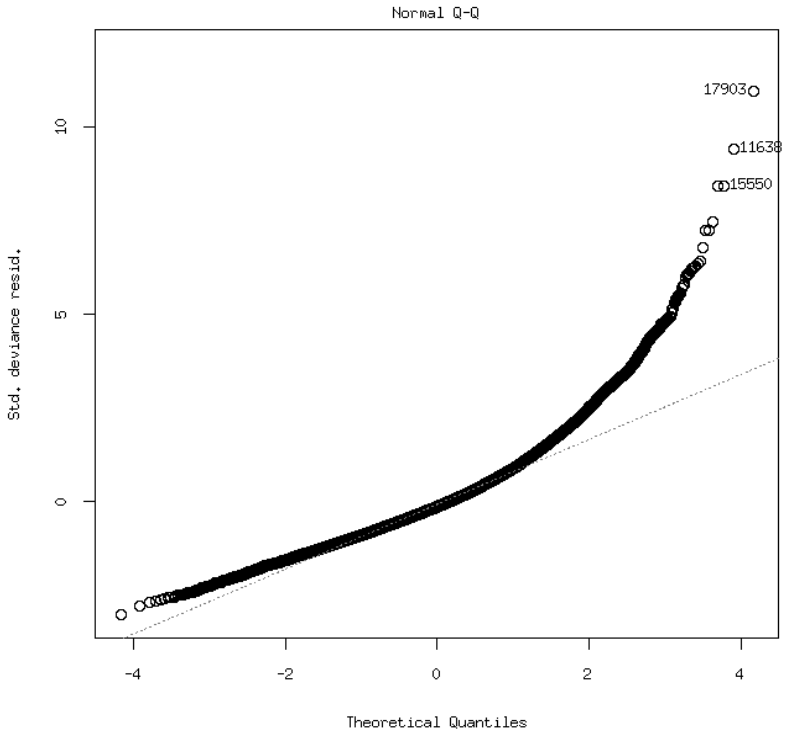
all p-values < 2e-16, except where marked

Raw continuous variables, fitted together by:

```
glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_transformed + MONOsum1_transformed + mfetsum1_transformed + fettsum1_transformed + sacksum1_transformed + kolhsum1_transformed + FA_transformed + protsum1_transformed + fibesum1_transformed + NATRsum1_transformed, family = gaussian(link = "identity"))
```

	Regression coefficient	Variance explained	p-value
POLYsum1	5.144e-01	0.000287	0.00230
MONOsum1	4.466e-01	0.000262	0.00360
mfetsum1	-1.927e+00	0.004702	< 2e-16
fettsum1	1.610e+00	0.001109	2.08e-09
acids	-1.010e+00	0.001481	4.35e-12
kolhsum1	-8.809e-02	6.1e-05	0.15853
sacksum1	-1.813e-01	0.000826	2.33e-07
protsum1	5.398e-01	0.00312	< 2e-16
NATRsum1	2.470e-01	0.000511	4.74e-05
fibesum1	-2.839e-01	0.001278	1.25e-10





Raw continuous, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.2075034	0.002141	< 2e-16
MONOsum1	0.3415456	0.004986	< 2e-16
mfetsum1	0.1156133	0.000583	1.39e-05
fettsum1	0.2407354	0.00245	< 2e-16
acids	0.1315379	0.00088	1e-07
kolhsum1	-0.0007506	0	0.9752463
sacksum1	-0.1812365	0.001842	< 2e-16
protsum1	0.4033973	0.008609	< 2e-16
NATRsum1	0.4496075	0.010221	< 2e-16
fibesum1	-0.0300287	5.4e-05	0.1857612

Continuous expressed in % of TEI, fitted together by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_ofTEI_transformed + MONOsum1_ofTEI_transformed + mfetsum1_ofTEI_transformed + fettsum1_ofTEI_transformed + sacksum1_ofTEI_transformed + kolhsum1_ofTEI_transformed + FA_ofTEI_transformed + protsum1_ofTEI_transformed + fibesum1_ofTEI_transformed + NATRsum1_ofTEI_transformed, family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.3992368	0.000351	0.000748
MONOsum1	0.1763958	0.00013	0.040343
mfetsum1	-1.1323844	0.004352	< 2e-16
fettsum1	1.2026031	0.00188	6.09e-15
acids	-0.7423517	0.001502	3.07e-12
kolhsum1	0.3109844	0.000867	1.17e-07
sacksum1	-0.1255568	6e-04	1.05e-05
protsum1	0.3799312	0.004332	< 2e-16
NATRsum1	0.1475537	0.000635	5.81e-06
fibesum1	-0.2148042	0.001155	9.61e-10

Continuous expressed in % of TEI, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.1271074	0.000929	< 2e-16
MONOsum1	0.3168217	0.004969	< 2e-16
mfetsum1	-0.0076495	3e-06	0.7546639
fettsum1	0.162972	0.001328	< 2e-16
acids	0.0402379	9.5e-05	0.0795438
kolhsum1	-0.2624236	0.003475	< 2e-16
sacksum1	-0.2939958	0.005146	< 2e-16
protsum1	0.4935001	0.014374	< 2e-16
NATRsum1	0.508124	0.015301	< 2e-16
fibesum1	-0.1922141	0.001817	< 2e-16

Categorized according to the recommended cut points, fitted together by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_ofTEI_categorized_g + MONOsum1_ofTEI_categorized_g + mfetsum1_ofTEI_categorized_g + fettsum1_ofTEI_categorized_g + sacksum1_ofTEI_categorized_g + kolhsum1_ofTEI_categorized_g + FA_ofTEI_categorized_g + protsum1_ofTEI_categorized_g + fibesum1_categorized_g + NATRsum1_categorized_g, family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.0570714	4.7e-05	0.216702
MONOsum1	0.2352016	0.000354	0.000715
mfetsum1	-0.4311723	0.000695	2.09e-06
fettsum1	0.2128727	0.000319	0.001322
acids	0.2400015	0.000119	0.050148
kolhsum1	-0.0298615	7e-06	0.625836
sacksum1	-0.2140844	0.000144	0.031122
protsum1	1.0617234	0.002321	< 2e-16
NATRsum1	0.6812422	0.003633	< 2e-16
fibesum1	0.2377727	0.000397	0.000337

Categorized according to the recommended cut points, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.1821857	0.000586	1.33e-05
MONOsum1	0.3238795	0.001015	< 2e-16
mfetsum1	-0.0302204	5e-06	0.6909235
fettsum1	0.2967705	0.001178	< 2e-16
acids	-0.0352855	3e-06	0.7655599
kolhsum1	-0.2952981	0.001295	< 2e-16
sacksum1	-0.4216125	0.000581	1.45e-05
protsum1	1.2621739	0.003386	< 2e-16
NATRsum1	0.6820745	0.004294	< 2e-16
fibesum1	0.0585875	3.1e-05	0.318958

Number of subjects in each category for each variable:

Visit 1:

	Number of subject with 0	Number of subject with 1	Number of subject with 2
POLYsum1	16615	14781	565
MONOsum1	7206	24707	48
mfetsum1	0	2089	29872
fettsum1	1629	25027	5305
acids	0	30809	1152
kolhsum1	8071	22895	995
sacksum1	0	28072	3889
protsum1	456	31336	169
NATRsum1	0	21555	10406
fibesum1	0	5772	26189

	bmi<18.5	18.5=<bmi<25	25=<bmi<30	30=<bmi
bmi	239	18067	11841	2834

Visit 2:

	Number of subject with 0	Number of subject with 1	Number of subject with 2
POLYsum1	13017	18247	1132
MONOsum1	6915	25363	118
mfetsum1	0	3275	29121
fettsum1	2106	23745	6545
acids	0	31186	1210
kolhsum1	11314	20042	1040
sacksum1	0	30569	1827
protsum1	332	31265	799
kolesum1	10691	10690	11015
NATRsum1	0	25636	6760
fibesum1	0	5769	26627

	bmi<18.5	18.5=<bmi<25	25=<bmi<30	30=<bmi
bmi	185	13890	14043	4971

Three different diet scores:
-diet score 1: combining all individual scores as they are, not a good idea, since some are very correlated, some are categorized opposite the direction of association with bmi...
-diet score 2: combining all individual scores, but multiplying them with the standardized beta coefficient of the continuous variable from an independent data set
-diet score 3: combining only scores from the variables which are complying with the guidelines used to create the scores and are not included in the other variables

visit=1:

	Regression coefficient, unstandardized	Variance explained	p-value
diet score 1	-1.055e-02	1e-05	0.57117
diet score 2	2.952e-01	0.001622	6.65e-13
diet score 3	5.322e-01	0.002004	1.35e-15

visit=2:

	Regression coefficient, unstandardized	Variance explained	p-value
diet score 1	1.720e-01	0.002377	< 2e-16
diet score 2	0.6205540	0.006883	<2e-16
diet score 3	1.0357356	0.007899	< 2e-16

Diet score details:

diet score 1 with categories 7 to 16:

visit=1:

score	Number of subjects	bmi (number of subjects who are)			
		underweight	normal	overweight	obese
7	69	0	34	25	8
8	520	0	263	199	57
9	152	2	623	392	122
10	3917	26	2240	1308	326
11	1060	74	6048	3655	819
12	11131	85	5947	4099	963
13	4044	33	2008	1579	400
14	464	0	24	167	51
15	22	0	14	5	3
16	2	0	1	0	1

visit=2:

score	Number of subjects	bmi (number of subjects who are)			
		underweight	normal	overweight	obese
7	143	0	66	61	16
8	814	2	400	312	100
9	1568	12	717	626	210
10	4256	18	1866	1802	570
11	10254	68	4441	4279	1451
12	11696	61	4901	4965	1764
13	3232	14	1091	1485	641
14	393	4	104	187	97
15	34	0	12	13	9
16	6	0	2	2	2

diet score 2 is continuous due to multiplication and ranges from -0.86547 to 4.84430, might not be a good scale and having negative scoring. To get a feeling of the distribution I made a standardized score and took part based on SD, so I divided the score in ten parts(as first score) and looked at the subject counts:

visit=1:

part of score	Number of subjects	bmi (number of subjects who are)			
		underweight	normal	overweight	obese
1	48	2	32	10	4
2	90	0	56	27	7
3	687	7	380	247	50
4	3967	34	2325	1273	314
5	13231	103	7590	4386	1091
6	8630	49	4506	3311	740
7	5082	34	2447	2081	501
8	156	0	64	64	27
9	56	4	17	25	12
10	14	0	5	5	4

visit=2:

part of score	Number of subjects	bmi (number of subjects who are)			
		underweight	normal	overweight	obese
1	25	0	12	13	0
2	32	0	15	14	3
3	651	1	283	288	79
4	4051	27	1860	1672	492
5	12990	82	5849	5269	1773
6	10784	53	4365	4685	1676
7	3019	14	971	1426	608
8	446	2	140	186	115
9	378	0	98	172	108
10	20	0	7	7	6

diet score 3 is continuous due to multiplication and ranges from 0.6275989 to 4.3454448. To get a feeling of the distribution I divided the score in 10 parts(as first score) and looked at the subject counts::

visit=1:

part of score	Number of subjects	bmi (number of subjects who are)			
		underweight	normal	overweight	obese
1	235	3	148	70	14
2	149	3	80	57	7
3	72	0	34	31	5
4	1313	9	712	458	130
5	17187	127	9768	5799	1421
6	6079	45	3256	269	491
7	6757	42	3355	2678	652
8	5	0	1	4	0
9	88	0	42	35	11
10	76	1	26	28	19

visit=2:

part of score	Number of subjects	bmi (number of subjects who are)			
		underweight	normal	overweight	obese
1	44	0	19	25	0
2	266	2	116	113	35
3	12	0	4	7	1
4	1872	6	890	750	226
5	18904	122	8446	7779	2537
6	8743	41	3357	3874	1469
7	1756	7	525	836	387
8	24	0	12	9	3
9	484	1	140	222	119
10	291	0	91	117	83

For PA only the following variables had sufficient data: g1_a, g1_b, g1_c, g1_d, g3_a, g3_b, g6. The variables g1_N are related since they are seasonal variables and we were discussing to combine them, so I had checked several different combinations:

1. Separate variables.
2. Combining all g1_N
3. Combining only the significant g1_N

Visit 1:

1.Separate variables:

	Regression coefficient, unstandardized	Regression coefficient, standardized	p-value
g1_a	-0.026707	-0.035912	0.01397
g1_b	-0.001490	-0.002117	0.85631
g1_c	-0.006834	-0.009016	0.52182
g1_d	-0.028987	-0.031211	0.00242
g3_a	-0.005845	-0.006559	0.31450
g3_b	-0.004055	-0.005348	0.46532
g6	-0.054040	-0.063570	< 2e-16

2.Combining all g1_N:

	Regression coefficient, unstandardized	Regression coefficient, standardized	p-value
g1_abcd	-0.014986	-0.070790	< 2e-16
g3_a	-0.005632	-0.006320	0.32869
g3_b	-0.003804	-0.005016	0.48647
g6	-0.053058	-0.062415	< 2e-16

3.Combining only the significant g1_N:

	Regression coefficient, unstandardized	Regression coefficient, standardized	p-value
g1_ad	-0.031370	-0.071011	< 2e-16
g3_a	-0.005596	-0.006279	0.33173
g3_b	-0.005429	-0.007159	0.31257
g6	-0.053258	-0.062651	< 2e-16

Visit 2:

1.Separate variables:

	Regression coefficient, unstandardized	Regression coefficient, standardized	p-value
g1_a	-0.0253117	-0.0332984	0.02985
g1_b	0.0006195	0.0008662	0.94453
g1_c	-0.0009063	-0.0011778	0.93659
g1_d	-0.0666347	-0.0716823	3.49e-11
g3_a	-0.0402458	-0.0420984	3.26e-10
g3_b	-0.0169811	-0.0228948	0.00373
g6	-0.0680579	-0.0912764	< 2e-16

2.Combining all g1_N:

	Regression coefficient, unstandardized	Regression coefficient, standardized	p-value
g1_abcd	-0.020930	-0.097848	< 2e-16
g3_a	-0.042628	-0.044591	2.12e-11
g3_b	-0.013237	-0.017847	0.021643
g6	-0.068313	-0.091619	< 2e-16

3.Combining only the significant g1_N:

	Regression coefficient, unstandardized	Regression coefficient, standardized	p-value
g1_ad	-0.044993	-0.100966	< 2e-16
g3_a	-0.042335	-0.044284	2.83e-11
g3_b	-0.014702	-0.019822	0.00904
g6	-0.068315	-0.091621	< 2e-16

Variance explained calculated separately and adjusting only for the basic variables, since PA variables have missing values in mixed places.

Visit 1:

	Variance explained	p-value
g1_a	0.006326	< 2e-16
g1_b	0.005051	< 2e-16
g1_c	0.005886	< 2e-16
g1_d	0.005701	< 2e-16
g3_a	0.000649	5.212801e-06
g3_b	0.002416	< 2e-16
g6	0.005676	< 2e-16
g1_abcd	0.006849	< 2e-16
g1_ad	0.006977	< 2e-16

Visit 2:

	Variance explained	p-value
g1_a	0.012533	< 2e-16
g1_b	0.009751	< 2e-16
g1_c	0.01193	< 2e-16
g1_d	0.013767	< 2e-16
g3_a	0.006293	< 2e-16
g3_b	0.010456	< 2e-16
g6	0.012445	< 2e-16
g1_abcd	0.014299	< 2e-16
g1_ad	0.01496	< 2e-16

Probably best to keep just g1_a, g1_d and g6 for the final PA score, out of all combinations, the beta coefficient and r^2 were the biggest:

Visit 1:

	Regression coefficient, unstandardized	Regression coefficient, standardized	Variance explained	p-value
PA score	-0.038565	-0.102699	0.010857	< 2e-16

Visit 2:

	Regression coefficient, unstandardized	Regression coefficient, standardized	Variance explained	p-value
PA score	-0.0559818	-0.1538940	0.023603	< 2e-16

PA score details:

PA score with categories 0 to 10:

visit=1:

category	Number of subjects	bmi (number of subjects who are)			
		underweight	normal	overweight	obese
0	7514	60	3562	3032	834
1	4669	26	2250	1934	440
2	3882	16	2118	1422	310
3	2803	17	1614	970	190
4	3226	30	1845	1058	285
5	2596	17	1527	843	197
6	2593	25	1593	804	160
7	1919	12	1174	605	120
8	1243	9	846	326	58
9	625	7	442	156	18
10	213	2	152	51	6

visit=2:

category	Number of subjects	bmi (number of subjects who are)			
		underweight	normal	overweight	obese
0	7230	35	2476	3292	1424
1	3774	19	1276	1769	707
2	2969	10	1185	1325	448
3	3823	17	1622	1694	488
4	3370	23	1463	1403	477
5	2313	14	1093	896	309
6	2487	12	1250	939	285
7	1947	14	998	762	171
8	1299	6	773	416	103
9	929	9	546	338	36
10	326	2	208	100	16

Combining diet score and PA(PA reverse scoring):

Separate:

diet score 1
Visit 1:

	Regression coefficient, standardized	p-value
diet score 1	0.002307	0.71732
PA score	0.103639	< 2e-16

Visit 2:

	Regression coefficient, standardized	p-value
diet score 1	0.057348	< 2e-16
PA score	0.147512	< 2e-16

diet score 2
Visit 1:

	Regression coefficient, standardized	p-value
diet score 2	0.072328	< 2e-16
PA score	0.104734	< 2e-16

Visit 2:

	Regression coefficient, standardized	p-value
diet score 2	0.101192	< 2e-16
PA score	0.151720	< 2e-16

diet score 3
Visit 1:

	Regression coefficient, standardized	p-value
diet score 3	0.048991	< 2e-16
PA score	0.100540	< 2e-16

Visit 2:

	Regression coefficient, standardized	p-value
diet score 3	0.092354	< 2e-16
PA score	0.145231	< 2e-16

Summed (best results when summing the standardized diet and PA scores):

diet score 1
Visit 1:

	Regression coefficient, standardized	Variance explained	p-value
diet score 1 + PA score	0.067369	0.004034	< 2e-16

Visit 2:

	Regression coefficient, standardized	Variance explained	p-value
diet score 1 + PA score	0.1213241	0.012926	< 2e-16

diet score 2
Visit 1:

	Regression coefficient, standardized	Variance explained	p-value
diet score 2 + PA score	0.076094	0.005849	< 2e-16

Visit 2:

	Regression coefficient, standardized	Variance explained	p-value
diet score 2 + PA score	0.102899	0.010333	< 2e-16

diet score 3

Visit 1:

	Regression coefficient, standardized	Variance explained	p-value
diet score 3 + PA score	0.088926	0.007851	< 2e-16

Visit 2:

	Regression coefficient, standardized	Variance explained	p-value
diet score 3 + PA score	0.1319158	0.016556	< 2e-16

More robust score:

I had also made another attempt to create a more robust risk score, by categorizing variables based on the location in the standard normal distribution. I tried different cutpoints and categories, using 1*SD or 2*SD. For example, if using 1*SD as cutpoint and 4 categories, then for variables which are to be positively associated with bmi, I had set
0 : [,-1) ,
1 : [-1, 0),
2: [0,1),
3: [1,)
For variables which are to be negatively associated with bmi, I had set the categories the other way around. I had tried this since I observed better results with some variables, where the categorizing based on recommendations changed the direction of association with bmi or the association became insignificant. But with some variables, the results were worst. I had used the distribution of the raw continuous variables and expressed as % of TEI. The best results are using…

The independent dataset of Swedish only (47107 subjects):

Categorizing the % of TEI based on location in standardized normal distribution, fitted together by:

```
glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_ofTEI_categorized_n + MONOsum1_ofTEI_categorized_n
+ mfetsum1_ofTEI_categorized_n + fettsum1_ofTEI_categorized_n + sacksum1_ofTEI_categorized_n +
kolhsum1_ofTEI_categorized_n + FA_ofTEI_categorized_n + protsum1_ofTEI_categorized_n + fibesum1_categorized_n +
NATRsum1_categorized_n, family = gaussian(link = "identity"))
```

	Regression coefficient	Variance explained	p-value
POLYsum1	0.2613463	0.00028	0.000349
MONOsum1	0.4604947	0.001224	7.70e-14
mfetsum1	-0.3915533	0.001003	1.30e-11
fettsum1	0.0330560	4e-06	0.659415
acids	0.2187564	0.000212	0.001873
kolhsum1	-0.0448882	1e-05	0.503266
sacksum1	-0.0573420	4.1e-05	0.171973
protsum1	0.5994689	0.004369	< 2e-16
NATRsum1	0.5125242	0.002858	< 2e-16
fibesum1	0.2200713	0.000548	5.67e-07

Categorizing the % of TEI based on location in standardized normal distribution, fitted separately by:

```
glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link =
"identity"))
```

	Regression coefficient	Variance explained	p-value
POLYsum1	0.2475245	0.000865	< 2e-16

MONOsum1	0.4823759	0.003044	< 2e-16
mfetsum1	-0.0504075	3.5e-05	0.2089986
fettsum1	0.2280955	0.000693	< 2e-16
acids	-0.1233607	0.000219	0.0015748
kolhsum1	-0.4235747	0.002052	< 2e-16
sacksum1	-0.3331994	0.001611	< 2e-16
protsum1	0.7470236	0.008214	< 2e-16
NATRsum1	0.5777928	0.004808	< 2e-16
fibesum1	0.0974485	0.000155	0.0077181

Categorizing the raw variable, based on location in standardized normal distribution, fitted together by:

```
glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_categorized_n + MONOsum1_categorized_n +
fettsum1_categorized_n + mfetsum1_categorized_n + sacksum1_categorized_n + kolhsum1_categorized_n + FA_categorized_n +
protsum1_categorized_n + fibesum1_categorized_n + NATRsum1_categorized_n, family = gaussian(link = "identity"))
```

	Regression coefficient	Variance explained	p-value
POLYsum1	0.1011252	3.1e-05	0.2354
MONOsum1	0.4798268	0.000821	9.22e-10
mfetsum1	-0.4747318	0.001102	1.31e-12
fettsum1	-0.1235618	4.1e-05	0.1733
acids	0.0400984	5e-06	0.6222
kolhsum1	-0.2242919	0.00035	6.40e-05
sacksum1	-0.1988698	0.000449	6.05e-06
protsum1	0.5289741	0.002022	< 2e-16
NATRsum1	0.4824163	0.00155	< 2e-16
fibesum1	0.3320282	0.001072	2.62e-12

Categorizing the raw variable, based on location in standardized normal distribution, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value
POLYsum1	0.2845415	0.001134	< 2e-16

MONOsum1	0.438291	0.002517	< 2e-16
mfetsum1	0.075288	7.7e-05	0.0607732
fettsum1	0.2632603	0.000921	< 2e-16
acids	-0.232696	0.00076	< 2e-16
kolhsum1	-0.0465577	3.3e-05	0.2229115
sacksum1	-0.20525	0.000626	1e-07
protsum1	0.5213218	0.004197	< 2e-16
NATRsum1	0.5777928	0.004808	< 2e-16
fibesum1	0.0974485	0.000155	0.0077181

The 2 visits subset of Swedish only (33114 subjects in each visit),
visit==1 (33114 subjects):

Categorizing the % of TEI based on location in standardized normal distribution, fitted together by:

```
glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_ofTEI_categorized_n + MONOsum1_ofTEI_categorized_n
+ mfetsum1_ofTEI_categorized_n + fettsum1_ofTEI_categorized_n + sacksum1_ofTEI_categorized_n +
kolhsum1_ofTEI_categorized_n + FA_ofTEI_categorized_n + protsum1_ofTEI_categorized_n + fibesum1_categorized_n +
NATRsum1_categorized_n, family = gaussian(link = "identity"))
```

	Regression coefficient	Variance explained	p-value
POLYsum1	2.820e-01	0.000565	2.22e-05
MONOsum1	2.497e-01	0.000621	8.76e-06
mfetsum1	-4.409e-01	0.001914	5.86e-15
fettsum1	-4.077e-02	1e-05	0.57149
acids	2.103e-01	0.000327	0.00125
kolhsum1	-2.831e-02	7e-06	0.64550
sacksum1	-1.048e-01	0.000211	0.00950
protsum1	4.610e-01	0.003943	< 2e-16
NATRsum1	3.099e-01	0.001494	0.21306
fibesum1	5.391e-02	4.9e-05	5.33e-12

Categorizing the % of TEI based on location in standardized normal distribution, fitted separately by:

```
glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link =
"identity"))
```

	Regression coefficient	Variance explained	p-value

POLYsum1	0.1395885	0.00041	0.000301
MONOsum1	0.0990943	0.000191	0.0137052
mfetsum1	-0.3037979	0.001958	< 2e-16
fettsum1	-0.1114747	0.000262	0.0038718
acids	-0.0546606	6.1e-05	0.163207
kolhsum1	-0.0478943	4.9e-05	0.2124676
sacksum1	-0.2619674	0.001571	0
protsum1	0.5487643	0.006925	0.5487643
NATRsum1	0.3867049	0.003099	< 2e-16
fibesum1	-0.0942184	0.000223	0.0077647

Categorizing the raw variable, based on location in standardized normal distribution, fitted together by:
```
glm(bmi~age + agesq + gender_factor + year + ffq_factor + POLYsum1_categorized_n + MONOsum1_categorized_n +
fettsum1_categorized_n + mfetsum1_categorized_n + sacksum1_categorized_n + kolhsum1_categorized_n + FA_categorized_n +
protsum1_categorized_n + fibesum1_categorized_n + NATRsum1_categorized_n, family = gaussian(link = "identity"))
```

	Regression coefficient	Variance explained	p-value
POLYsum1	2.701e-01	0.00037	0.000598
MONOsum1	8.608e-02	4.4e-05	0.238781
mfetsum1	-5.021e-01	0.001719	1.37e-13
fettsum1	1.765e-02	1e-06	0.843679
acids	2.110e-01	0.000242	0.005500
kolhsum1	-1.532e-01	0.000237	0.006048
sacksum1	-1.769e-01	0.000536	3.61e-05
protsum1	3.678e-01	0.001385	3.14e-11
NATRsum1	4.060e-01	0.001523	3.34e-12
fibesum1	2.498e-02	9e-06	0.587909

Categorizing the raw variable, based on location in standardized normal distribution, fitted separately by: `glm(bmi~age + agesq + gender_factor + year + ffq_factor + VIP_data_independant[,c(variable)], family = gaussian(link = "identity"))`

	Regression coefficient	Variance explained	p-value

POLYsum1	0.1697378	0.000592	1.41e-05
MONOsum1	0.1011289	0.000198	0.0119913
mfetsum1	-0.1371225	0.000378	0.0005243
fettsum1	0.0224524	1e-05	0.5736579
acids	-0.1081224	0.00024	0.0057262
kolhsum1	0.0377398	3.4e-05	0.2973921
sacksum1	-0.1674143	0.000637	6.7e-06
protsum1	0.3381225	0.002588	< 2e-16
NATRsum1	0.3867049	0.003099	< 2e-16
fibesum1	-0.0942184	0.000223	0.0077647