Analysis of dietary patterns and breast cancer

1 Abstract

The goal of this project is to

1. Identify potential factors (both dietary and non-dietary) that associate with the risk of breast cancer.
2. Identify dietary patterns based on consumption of food groups.
3. Identify the association between dietary patterns and the risk of breast cancer, controlling for potential confounding factors.

2 Author

Tue Nguyen

3 Data processing

3.1 Missing data

I used the codebook to convert all the hard-coded values for missing data such as `0`, `9`, `99`, and `999` (depending on specific variables) **to the true, proper `NA` values** for unified treatment.

3.2 Occasional intake

Most variables on diet intake are represented by the frequency of consumption per week and range from `0` to `97`. The code `98` indicates "occasional consumption" around 1-3 times a month. Thus, I decided to **impute the code `98` with `0.5`** (i.e. twice per 4 weeks).

3.3 Smoking data

I used `fum1` to derive a new binary variable named `is_smoke`, which assumes values **`0` for never smoking subjects and `1` for current and ex-smokers**.

3.4 Body Mass Index

I used weight (`antr1`) and height (`antr2`) to compute the **BMI** using the formula:

$$BMI = \frac{weight}{height^2} \times 10000$$

3.5 Physical activities

Physical activities are divided into **work** (`fis1`, `fis3`, `fis5`, `fis7`) and **sport** (`fis2`, `fis4`, `fis6`, `fis8`). For the work category, I have:

- `fis1`: work activity at 12 years old
- `fis3`: work activity at 15-19 years old
- `fis5`: work activity at 30-39 years old
- `fis7`: work activity at 50-59 years old

These variables are in ordinal level, ranging from `1` (very heavy) to `5` (sedentary). This causes a problem of how to represent physical activity for each subject because the subjects are in different age ranges. For example, subjects younger than 50 years old don't have data on `fis7`. And similarly, those that are younger than 30 years old don't have data on both `fis5` and `fis7`.

I want to utilize the data on physical activity but also don't want a large number of missing values. Thus, for each subject, I used **the activity closest to their age level** as a proxy for physical activity. Specifically, I derived a new variable named `act_work` as follows.

- Younger than 12: use `fis1` (but `min(age) = 19`, so this is not used)
- Younger than 30: use `fis2`
- Younger than 50: use `fis3`
- 50 or older: use `fis5`

I did **the same for sports activities** with a new variable named `act_sport`.

> **Note:** In the original coding for physical activities, `1` is used for the highest level while `4` and `5` are used for the lowest levels. I **chose to reverse this ordering** so that it will be more convenient later when selecting the reference group for computing odd ratios and running logistic regressions.
>
> This means in the analysis, `1` is the group with **lowest exposure** to physical activity.

3.6 Family history of cancer

I created a binary variable named `rel_cancer_1st_deg` to indicate whether the subject has **any relatives of the first degree who had cancer**. A first-degree relative person is defined as a parent, brother, sister, or child.

3.7 Reproductivity

I decided to include some variables related to the subjects' Reproductivity such as:

- Age at menarche (`gin1`)
- Number of miscarriages (`gin8`)
- Number of abortions (`gin9`)
- Number of children (`v11`)

3.8 Food intake

I chose to retain 83 food items and group them into 24 groups depending on their nature. The food groups and their corresponding individual food items are presented in the table below.

Definition of food groups

food_group	n	items
beef	3	steak, boiled, stew
beer	1	regular
coffee	3	capu, other, decaff
dairy	2	ricotta, cheese
egg	2	boiled, fried
fat	1	general
fish	3	boiled, fried, tuna
fruit	9	apple_pear, banana, kiwi, cooked, citrus, peaches, melon, grape, berry

food_group	n	items
grain	5	bread, wheat_bread, crackers, maize, rice
juice	2	unsweetened, sweetened
liquor	4	grappa, whisky, after_dinner, other
milk	4	whole, skimmed_part, skimmed, yoghurt
organ	1	liver
pasta	5	butter, tomato, ragu, pesto, lasagne
pizza	1	general
pork	2	weiner, general
potato	2	boiled, fried
poultry	2	boiled, roasted
processed_meat	3	prosciutto, ham, salami
snack	10	biscuit, croissant, custard, cake, jampie, chocolate, soft_drink, candy, icecream, general
soup	2	light, veggie
sweet	4	sugar, sacca, other, honey_jam
veggie	11	bean, salad, carrot_raw, carrot_cooked, onion, artichoke, cruciferae, spinach, tomato, mixed_salad, zucchini
wine	1	regular

To keep the analysis simple, I decided to **work at food group levels instead of food item levels**. To do so, I computed the **total intake for each group by summing up the frequency of consumption of each individual food item**. This is not perfect, but it is the best I could do.

One note is that I **tried to utilize the serving sizes** (`porz` variables) as weights for frequency consumption, but most of them have **large portion of missing values** (code `0` and `9`). One possible solution is to impute the missing serving size by the mode of the distribution (because serving sizes are in the ordinal scale). But there is still a lot of uncertainty in doing so, and I chose to leave this option out. Thus, I only use the raw frequency count in the analysis.

In summary, there are **two problems in my aggregation method**.

1. Simple, unweighted summation of the frequencies of different food items as a representation of overall intake for each food group is not perfect. If we have additional information such as calories per portion of food or something similar to use as weights, then it would be nicer.
2. I couldn't able to incorporate the serving size in the aggregation (which I should) due to missing data.

3.9 Retained variables

I decide to retain the following variables for further analyses.

1. Ground truth for case-control
    - `has_cancer` : `0` = not having cancer, `1` = having cancer
2. Smoking information

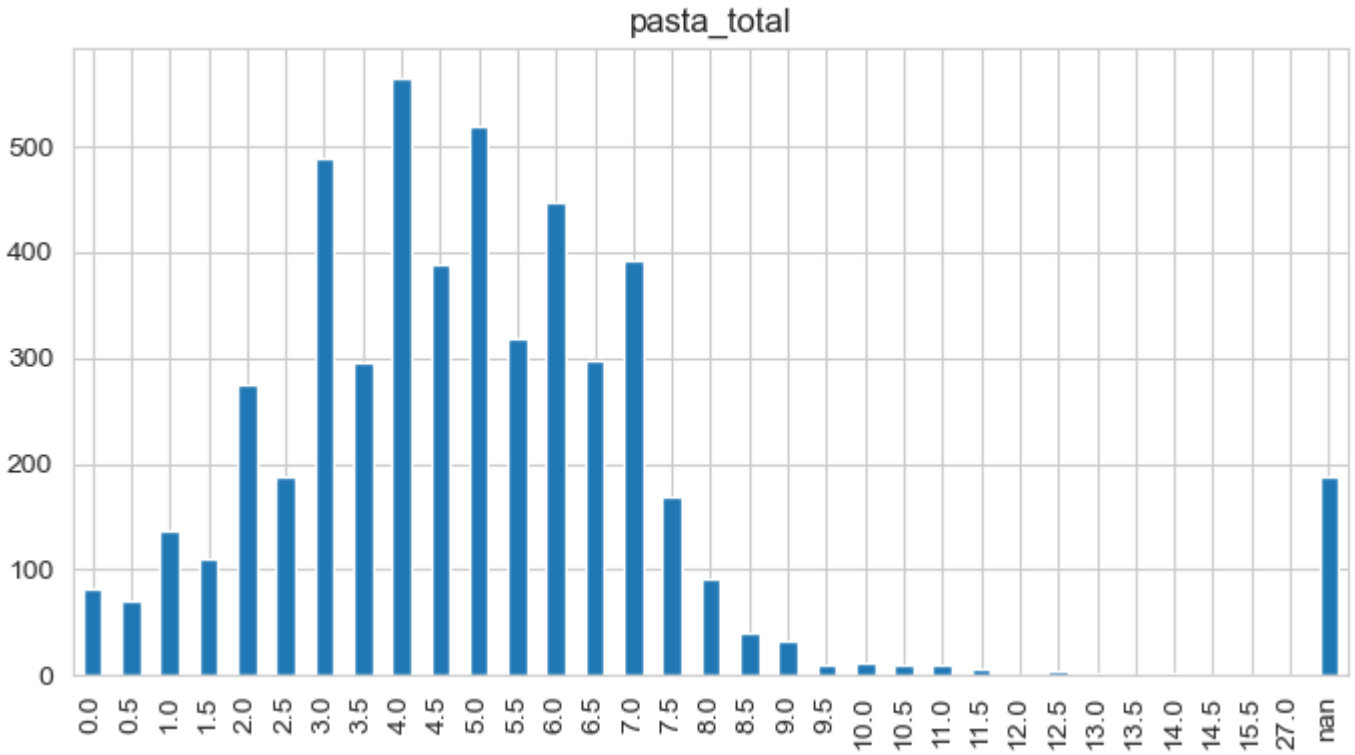
    - `is_smoke` : `0` = non-smoker, `1` = current/ex-smokers
    - `smoke_yrs` : number of years smoking
3. Height and weight
    - `height` : height
    - `weight` : weight
    - `bmi` : BMI computed from height and weight
4. Education
    - `edu` : years of schooling
5. Marital status
    - `marital_status` : with 3 levels: `never` , `married_coliving` , and `divorced_widowed`
6. Physical activities
    - `act_work` : level of work activity: `1` = mostly seated to `5` = very heavy
    - `act_sport` : level of sport activity: `1` = 2h/week to `4` = more than 7h/week
7. Family history of cancer
    - `rel_cancer_1st_deg` : whether the subject has any first-degree relative having cancer
8. Reproducibility characteristics
    - `age_menarche` : age at menarche
    - `n_children` : number of children
    - `n_abortions` : number of abortions
    - `n_miscarriages` : number of miscarriages
9. Aggregate frequencies of consumption for each food group
    - These are 24 variables on the aggregated frequencies of consumption per week for 24 food groups.
    - Examples: `fat_total` , `beef_total` , `pork_total` , `veggie_total` , etc.

3.10 Check data quality

After the cleaning step, I run a `for` loop to generate the plot of distribution for each retained variable to make sure there is no anomaly. For example, the below plots is the distribution of total pasta consumption. We can see that

- Code `98` is already replaced with `0.5`
- There is a bar at the end for true missing value instead of code `99`

I repeat this check for all the retained variables.



3.11 Finalized data

The original dataset contains 5157 rows. After the decision on what variables to keep, I explore the proportion of missing data in the cleaned dataset to decide whether to just discard the missing data or to perform some imputation.

Here are the top 10 variables with the highest number of missing values.

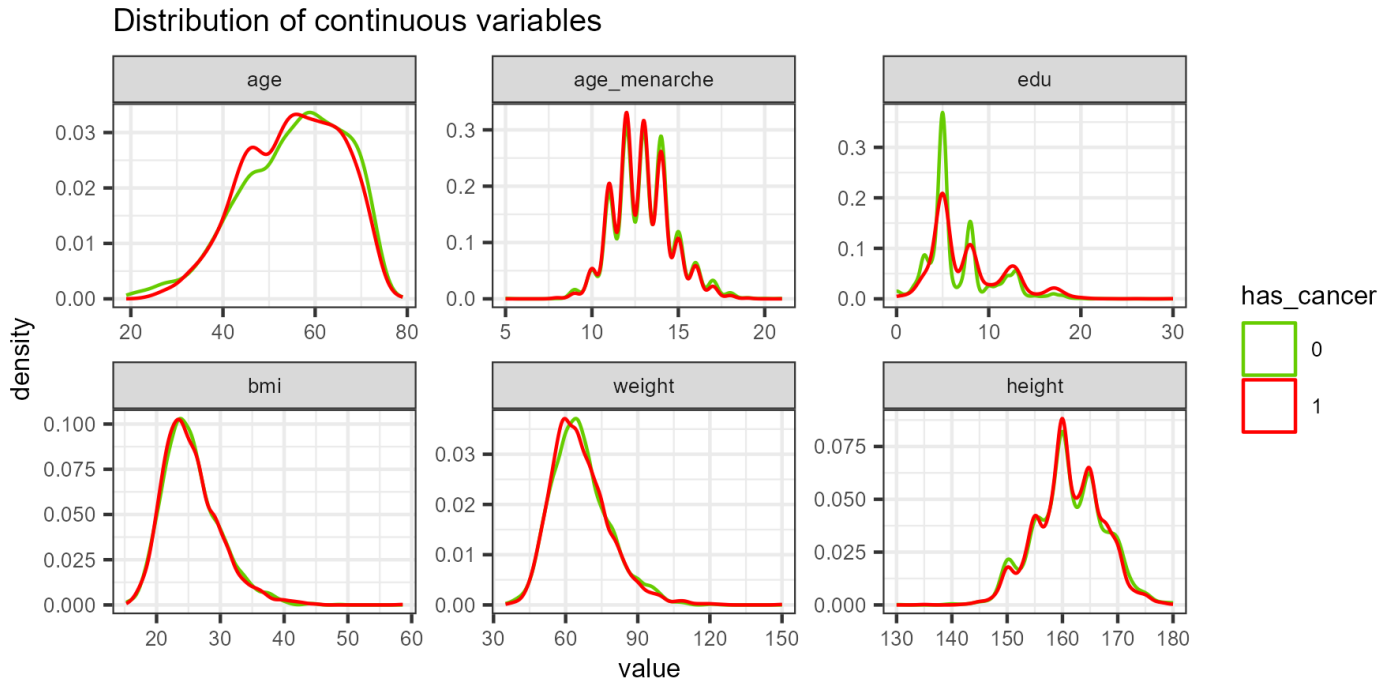
Top 10 variables having NAs

variable	num_NA	pct_NA
rel_cancer_1st_deg	362	7.02
act_work	187	3.63
pasta_total	187	3.63
act_sport	186	3.61
sweet_total	59	1.14
edu	37	0.72
bmi	15	0.29
fruit_total	10	0.19
weight	9	0.17
height	9	0.17

Since the amount of missing data is negligible, I decided to simply discard rows that have `NA` in any cell. And the finalized data without any missing values have 4307 rows. I were still able to retained 83.52% of the original dataset.

4 Exploratory analysis

4.1 Univariate analysis



> **Comment:** no clear signal.

For discrete and categorical variables, I compute the (unadjusted) odd ratio (95% confidence interval included). However, to have reliable estimates, only groups with more than 100 observations are kept.

Odd ratio for is_smoke (95%-CI)

value	odd_ratio	lower	upper	case	control
0	**1**	•	•	1398	1520
1	**1.13**	0.995	1.285	708	681

> **Comment:** no clear signal.

Odd ratio for rel_cancer_1st_deg (95%-CI)

value	odd_ratio	lower	upper	case	control
0	**1**	•	•	1153	1348
1	**1.306**	1.157	1.475	953	853

Comment: Having first-degree relatives who have cancer tend to **increase** the risk of breast cancer.

Odd ratio for n_children (95%-CI)

value	odd_ratio	lower	upper	case	control
0	**1**	•	•	334	327
1	**1.106**	0.905	1.352	480	425
2	**0.994**	0.829	1.192	802	790
3	**0.789**	0.639	0.974	327	406
4	**0.678**	0.502	0.916	97	140

Comment:

- Having more children tends to **reduce** the risk of breast cancer (perhaps reproductivity capability is the common cause)
- Other levels were removed due to too few observations

Odd ratio for n_abortions (95%-CI)

value	odd_ratio	lower	upper	case	control
0	**1**	•	•	1841	1969
1	**1.405**	1.097	1.800	155	118
2	**1.121**	0.787	1.597	65	62

Comment:

- Having had abortions tends to **increase** the risk of breast cancer
- Other levels were removed due to too few observations

Odd ratio for n_miscarriages (95%-CI)

value	odd_ratio	lower	upper	case	control
0	**1**	•	•	1650	1672
1	**0.835**	0.706	0.987	299	363
2	**1.08**	0.822	1.418	114	107

Comment:

- Experiencing miscarriage tends to **reduce** the risk of breast cancer (very counter-intuitive)
- Other levels were removed due to too few observations

Odd ratio for act_work (95%-CI)

value	odd_ratio	lower	upper	case	control
1	**1**	•	•	231	194
2	**0.791**	0.638	0.980	772	820
3	**0.795**	0.643	0.983	890	940
4	**0.754**	0.574	0.990	193	215

Comment:

- `act_work = 5` (very heavy work) was removed due to too few observations
- Heavier level of work tends to **reduce** the risk of breast cancer

Odd ratio for act_sport (95%-CI)

value	odd_ratio	lower	upper	case	control
1	**1**	•	•	1562	1628
2	**1.042**	0.892	1.218	396	396
3	**0.877**	0.667	1.153	101	120
4	**0.859**	0.580	1.272	47	57

Comment:

- No clear signal

Odd ratio for marital_status (95%-CI)

value	odd_ratio	lower	upper	case	control
never	**1**	•	•	179	203
married_coliving	**1.155**	0.933	1.429	1562	1534
divorced_widowed	**0.892**	0.699	1.138	365	464

Comment:

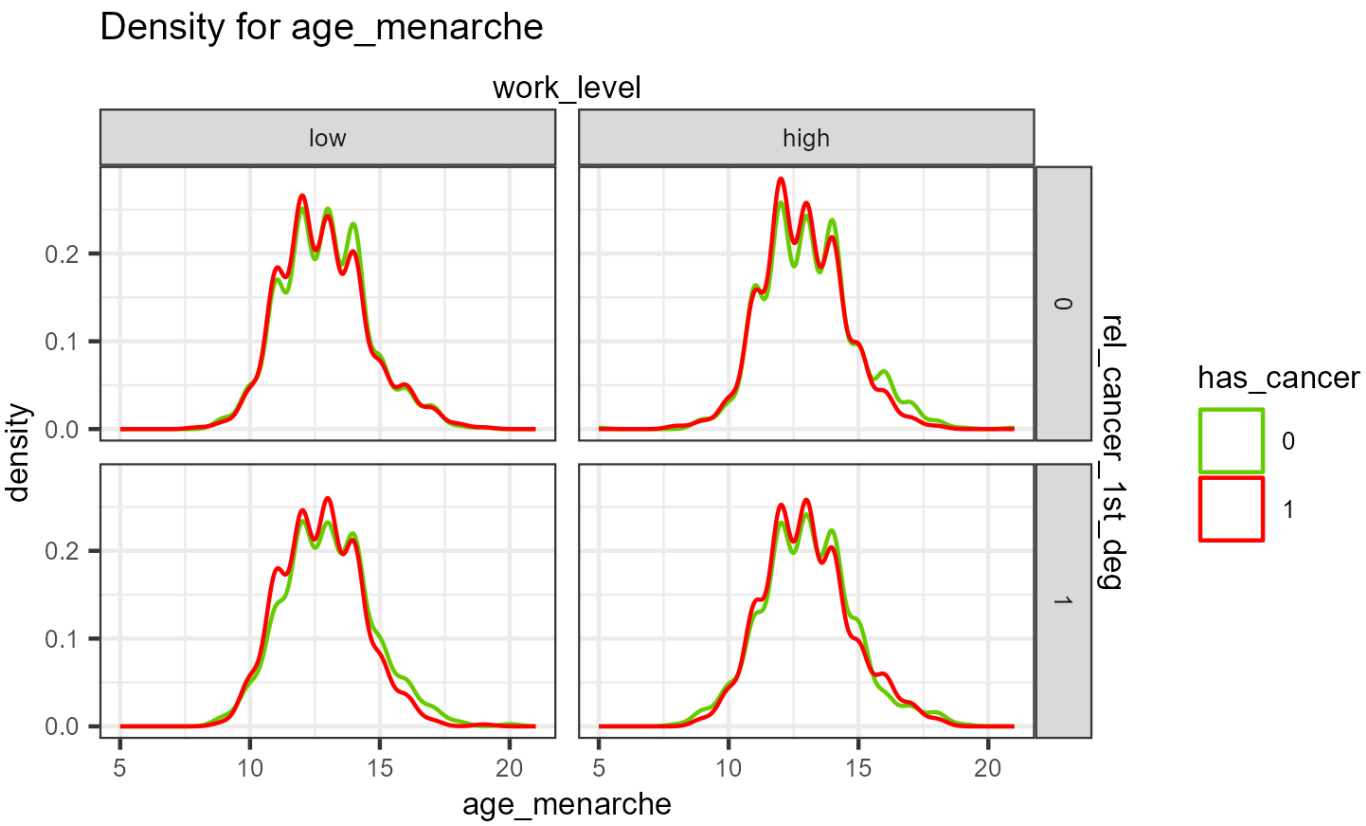
- `marital_status = 'never'` was chosen as the reference group
- No clear signal

4.2 Stratification

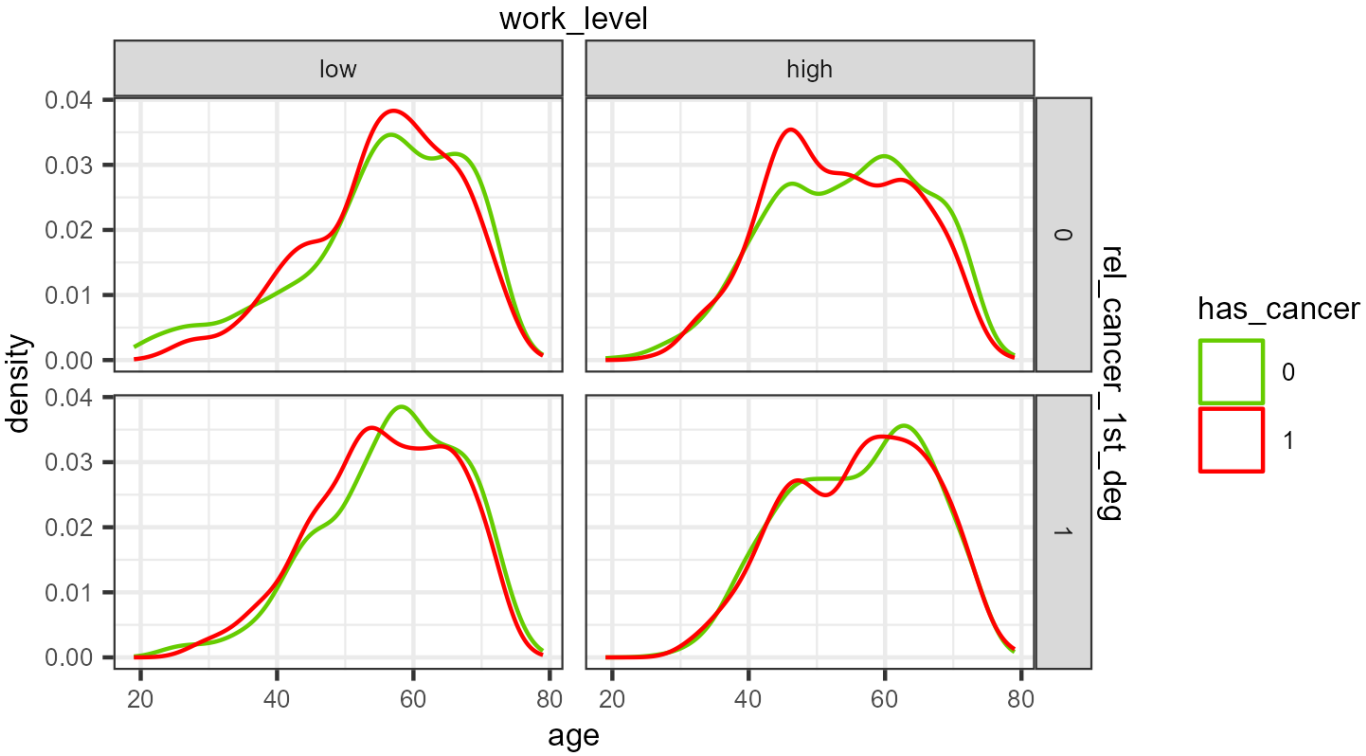
From the previous analysis, we already identified `rel_cancer_1st_deg` and `act_work` as potential factors that associate with the risk of breast cancer. Thus, conditioning on these two factors, we will explore the distribution of other (continuous) variables such as `age_menarche`, `age`, `edu`, and `bmi` to see if there is any signal.

Since `act_work` has 5 levels and `rel_cancer_1st_deg` has 2 levels, they make up 10 groups in total. This will make some groups to have too few observations to be meaningful. Thus, I created a variable named `work_level` with only two values: `"low"` (for `act_work = 1, 2`) and `"high"` (for `act_work = 3, 4, 5`)

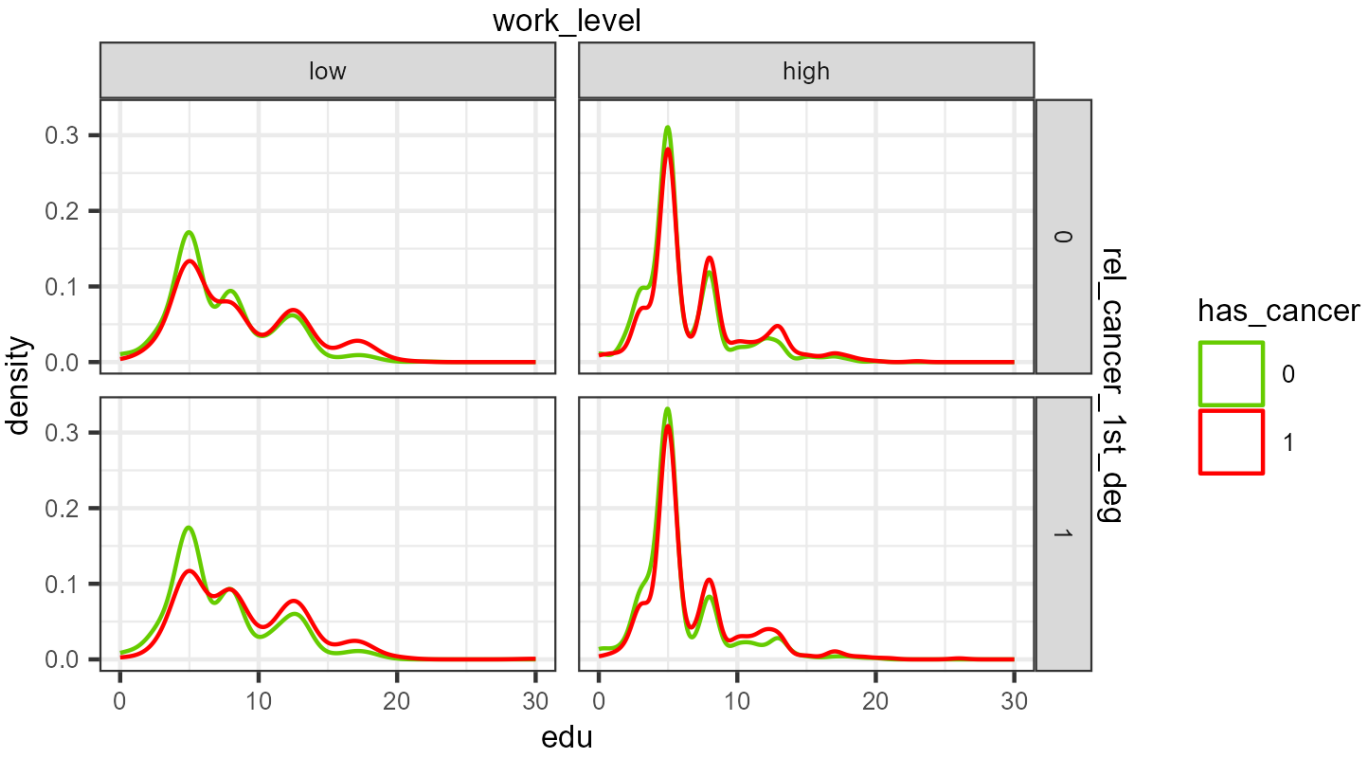
Unfortunately, no obvious signal was found.

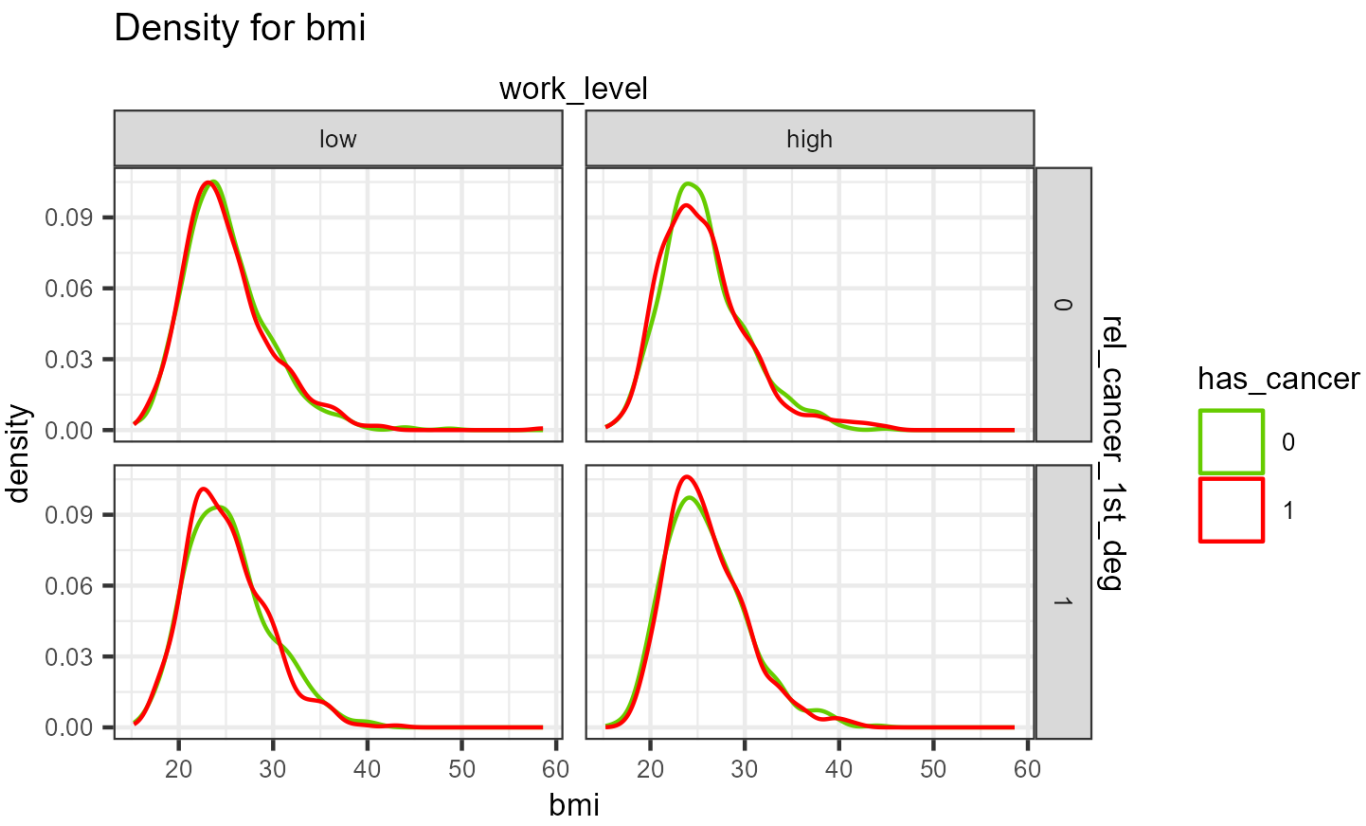


Density for age



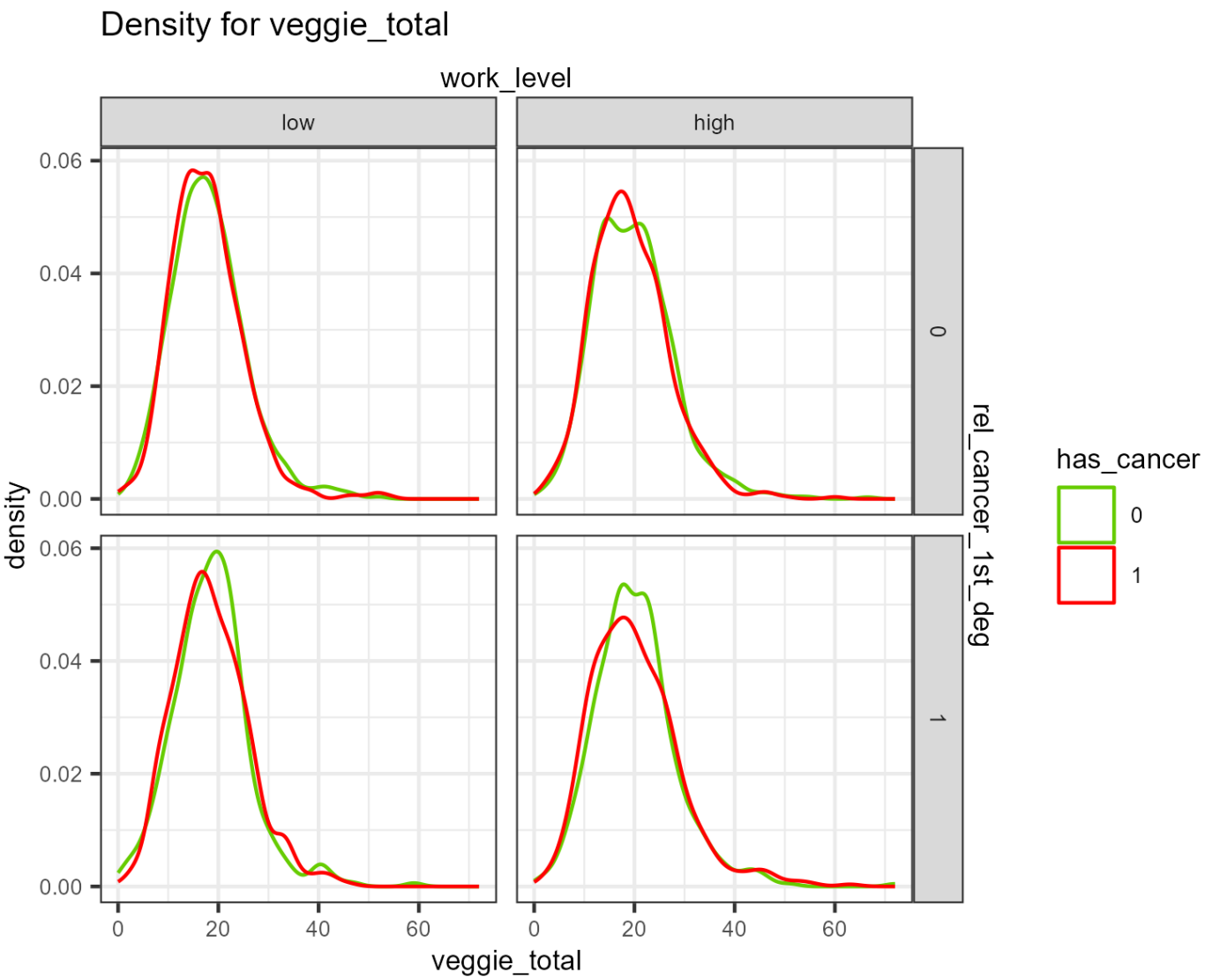
Density for edu





4.3 Dietary variables

I replicate a similar analysis for intake of each food group and review all the plots. Unfortunately, there is no clear signal. All the plots look similar to the one below.



5 Identify dietary patterns

I perform a PCA on 24 food groups (mean-centered and unit standard deviation).

5.1 PCs and variance explained

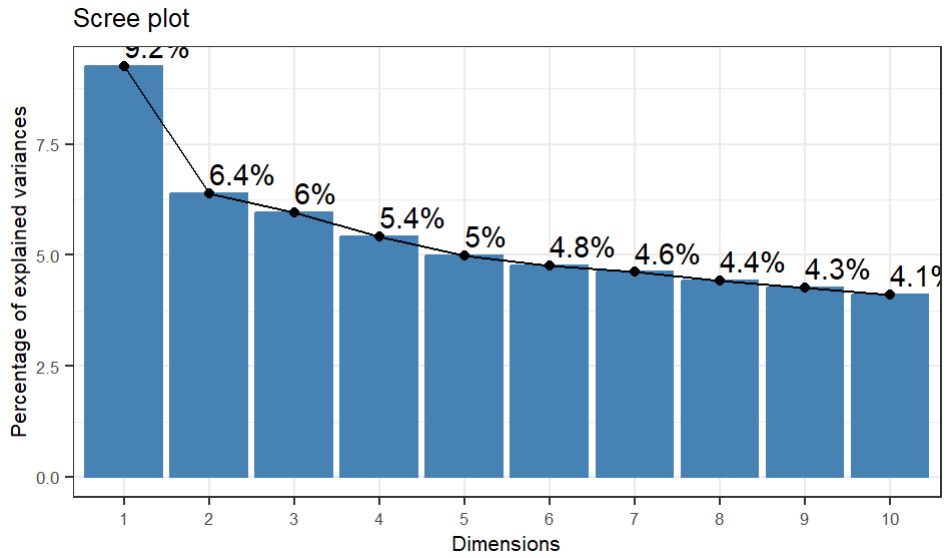
Principal components

pc	eigenvalue	pct_var	cum_pct_var
Dim.1	2.220	9.249	9.249
Dim.2	1.535	6.394	15.643
Dim.3	1.430	5.959	21.602
Dim.4	1.300	5.416	27.019
Dim.5	1.196	4.984	32.002
Dim.6	1.143	4.762	36.764
Dim.7	1.108	4.616	41.380
Dim.8	1.064	4.433	45.813

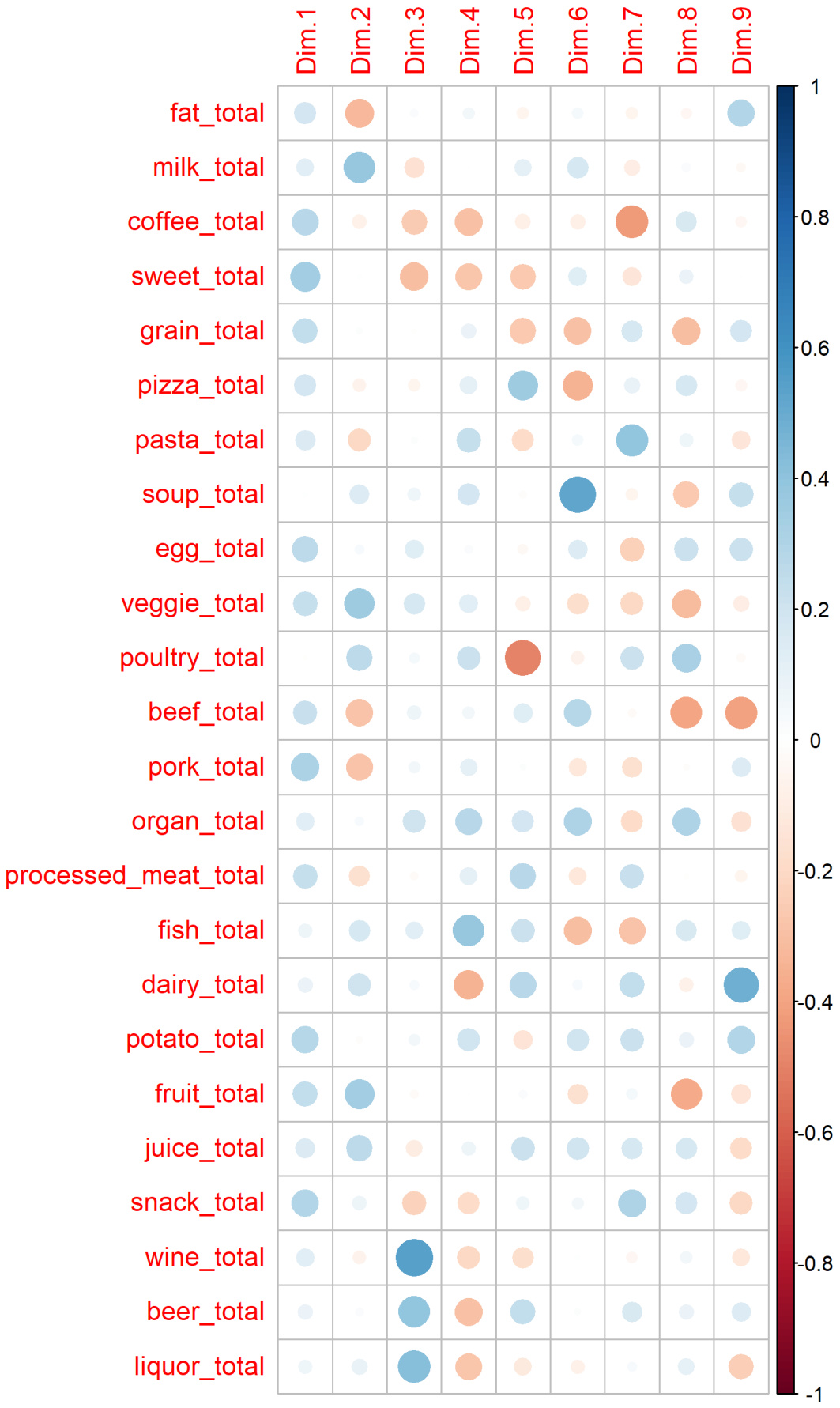
pc	eigenvalue	pct_var	cum_pct_var
Dim.9	1.024	4.265	50.078
Dim.10	0.984	4.101	54.180
Dim.11	0.939	3.911	58.091
Dim.12	0.903	3.764	61.855
Dim.13	0.886	3.693	65.548
Dim.14	0.864	3.602	69.150
Dim.15	0.849	3.539	72.689
Dim.16	0.842	3.509	76.197
Dim.17	0.829	3.455	79.652
Dim.18	0.803	3.347	82.999
Dim.19	0.776	3.234	86.233
Dim.20	0.744	3.102	89.335
Dim.21	0.706	2.941	92.276
Dim.22	0.675	2.811	95.087
Dim.23	0.627	2.611	97.699
Dim.24	0.552	2.301	100.000

Comment: The first 9 PCs have eigenvalues > 1 and together they explain around 50% of the variance of the data.

5.2 Screeplot



5.3 Factor loadings

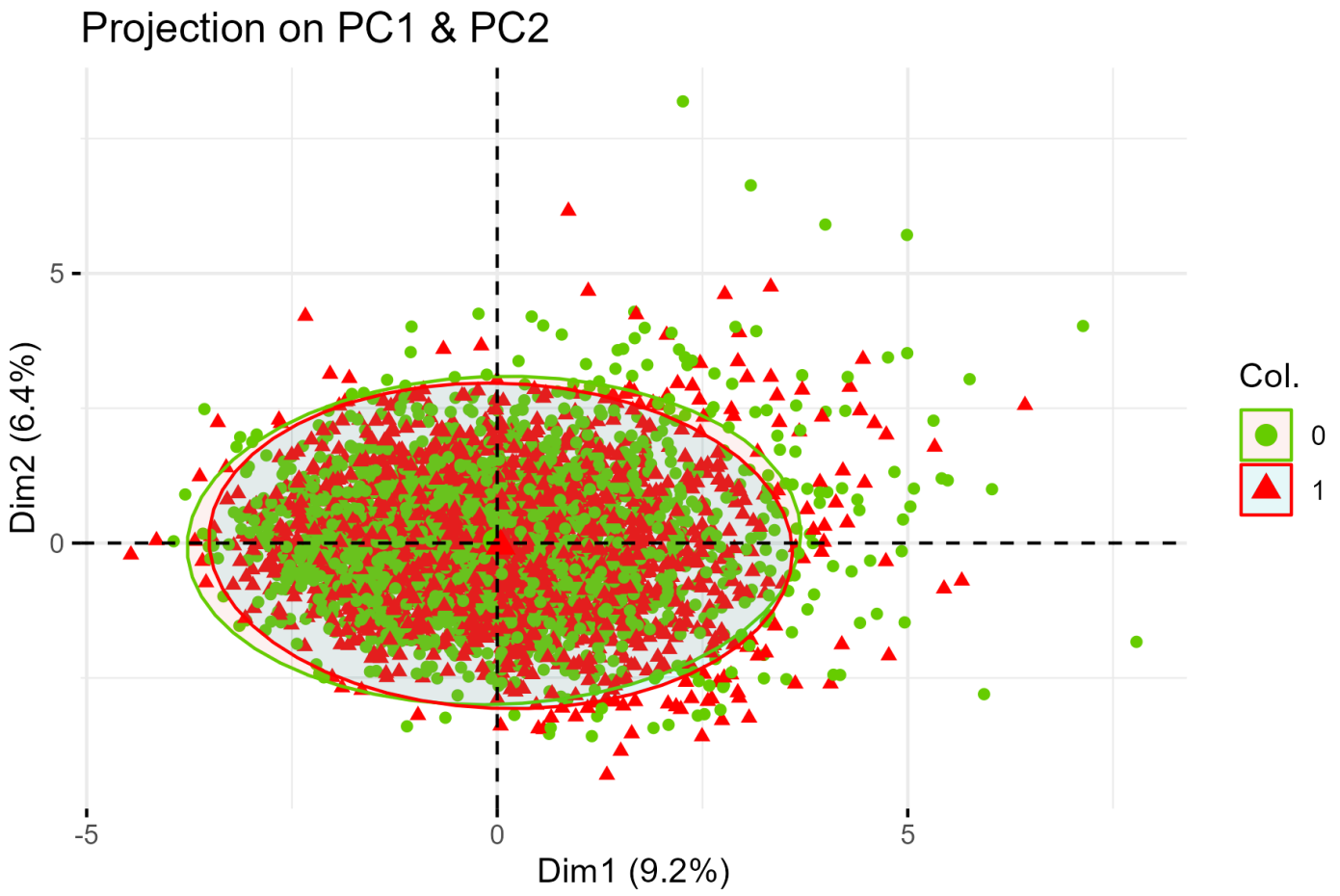


Comments:

- PC1: **traditional eating pattern** with balanced consumption of most traditional food groups including coffee, sweet, grain, eggs, vegetable, beef, pork, potato, fruit, and snacks.
- PC2: **healthy pattern** with milk, poultry (white meat), vegetable, fruit, and juice. Besides, this pattern avoids beef (red meat), fat, pork, and processed meat.
- PC3: **drinkers pattern** with main consumption of wine, beer, and liquor.
- PC4: **healthy 2** with main consumption of fish, organ, poultry, pasta, soup, vegetable, and potato. Besides, this pattern avoids alcohol, snacks, coffee, sweets, and processed dairy products.
- No clear interpretation found for PC >= 5

5.4 Projections on first two PCs

I tried to project the data on the first two PCs to see if they can well separate out cases and controls. However, they can't, which is expected because the first 2 PCs only explain only less than 16% of the total variation.



6 Dietary patterns and risk of BC

	odd_ratio	lower	upper	p_val	z_val	estimate	std_error
(Intercept)	**1.5077**	0.7657	2.9715	0.2351	1.1872	0.4106	0.3458
is_smoke1	**1.027**	0.8981	1.1743	0.6967	0.3898	0.0267	0.0684
rel_cancer_1st_deg1	**1.2977**	1.1476	1.4677	0.0000	4.1529	0.2606	0.0628
age_menarche	**0.9579**	0.9233	0.9937	0.0216	-2.2966	-0.0430	0.0187
n_children	**0.8871**	0.8451	0.9307	0.0000	-4.8652	-0.1198	0.0246
n_abortions	**1.0671**	0.9809	1.1615	0.1308	1.5109	0.0650	0.0430
n_miscarriages	**0.9574**	0.8847	1.0356	0.2779	-1.0851	-0.0435	0.0401
act_work2	**0.8517**	0.6836	1.0602	0.1513	-1.4349	-0.1606	0.1119
act_work3	**0.8674**	0.6967	1.0791	0.2022	-1.2754	-0.1423	0.1116
act_work4	**0.8326**	0.6279	1.1034	0.2027	-1.2739	-0.1832	0.1438
act_work5	**0.559**	0.3021	1.0126	0.0582	-1.8944	-0.5816	0.3070
bmi	**1.0082**	0.9939	1.0226	0.2629	1.1195	0.0081	0.0073
pc12	**1.1821**	0.9746	1.4340	0.0895	1.6979	0.1673	0.0985
pc13	**1.2519**	1.0307	1.5211	0.0236	2.2636	0.2247	0.0993
pc14	**1.34**	1.1003	1.6327	0.0036	2.9074	0.2927	0.1007
pc15	**1.4119**	1.1528	1.7304	0.0009	3.3298	0.3450	0.1036
pc22	**0.9002**	0.7427	1.0909	0.2835	-1.0725	-0.1052	0.0981
pc23	**0.9144**	0.7534	1.1097	0.3649	-0.9060	-0.0895	0.0987
pc24	**0.7984**	0.6578	0.9686	0.0225	-2.2816	-0.2252	0.0987
pc25	**0.8608**	0.7094	1.0444	0.1287	-1.5191	-0.1499	0.0986
pc32	**1.5461**	1.2649	1.8911	0.0000	4.2481	0.4357	0.1026
pc33	**1.4199**	1.1543	1.7476	0.0009	3.3134	0.3506	0.1058
pc34	**1.3722**	1.1158	1.6883	0.0027	2.9950	0.3164	0.1056
pc35	**1.3478**	1.1003	1.6518	0.0040	2.8805	0.2985	0.1036
pc42	**0.8195**	0.6747	0.9951	0.0446	-2.0082	-0.1991	0.0991
pc43	**0.7875**	0.6462	0.9593	0.0178	-2.3706	-0.2389	0.1008
pc44	**0.7674**	0.6271	0.9385	0.0100	-2.5755	-0.2648	0.1028
pc45	**0.6699**	0.5446	0.8234	0.0001	-3.7996	-0.4006	0.1054

Comments:

- **Traditional and drinker patterns** (PC1 and PC3) are associated with **increased** risk of breast cancer
- **Healthy patterns** (PC2 and PC4) are associated with **reduced risk of breast cancer**

7 Future work

Due to limited time and knowledge of the domain, my analysis has several limitations. However, it can be improved in the future in the following dimensions.

1. Improve the representation of diet intake beyond the simple frequency of food consumption with external data such as weighting intake by serving size and nutritional contents (There are some free and commercial databases providing detailed nutritional contents for each type of food)
2. Find a better representation of physical activities
3. Incorporate information on the subjects' history of medical treatment
4. Replicate the analysis for the food item level and nutrition level