Lindsay Kastner

BS723 Summer 2024 – Project 1

Introduction

A sufficient amount of vitamin D is necessary for musculoskeletal health. Evidence links vitamin D insufficiency to a variety of medical disorders such as: type 1 and type 2 diabetes, cardiovascular disease (CVD) and certain types of cancer. The concentration of vitamin D in the blood can be measured through the biomarker 25-hydroxyvitamin D. Important determinants of that biomarker are diet and sun exposure, and it is believed that some genes, such as those related to skin pigmentation, may influence the amount of circulating vitamin D. Vitamin D insufficiency is defined as having a 25-hydroxyvitamin D concentration below 30 ng/ml.

Data Description

A cross-sectional study was conducted to examine the relationship between two genetic factors and Vitamin D levels. Individuals were recruited, in approximately equal numbers, from a community health center located in the following three cities: Boston, MA; Washington, DC; and Chicago, IL. Genetic data was obtained on two separate single nucleotide polymorphisms (SNP). SNPs are genetic variants at a particular location in an individual's DNA. The investigators coded each SNP is as a dichotomous: having at least one "A" allele (genotype AA, AB) versus having 0 "A" alleles (genotype BB).

Information on 1500 individuals were collected and stored in the file **vitamind2024.sas7bdat**. The dataset includes the following variables:

Variable	Description	Details
ID	Observation Number	
vitaminD	Blood concentration of 25-hydroxyvitamin D (ng/ml).	Concentrations below 10 ng/ml were marked as missing (.).
age	Age (years)	numeric
height	Height (meters)	numeric
weight	Weight (kilograms)	numeric
diet	Vitamin D Supplementation	0 = No, 1 = Yes

sex	Gender	M = Male, F = Female
month	Month of exam	1 = January, 2 = February, … , 12 = December
snp1	First Genetic Marker	1=genotype AA or AB; 0=genotype BB
snp2	Second Genetic Marker	1=genotype AA or AB; 0=genotype BB

Include your SAS code and SAS log. DO NOT INCLUDE YOUR SAS OUTPUT (except for the necessary graphs). [10 points]

Section 1 – Data Processing

1. Create a library called *analysis* to read in the **vitamin_sum2024.sas7bdat** dataset. **[2 points]**

```
libname analysis '/home/u63889606/BS 723/Project 1';
run;
```

2. Create a temporary dataset called *vit_data* from *analysis.vitamind* dataset with the following changes:

NOTE: All the following data manipulation should be done within the vit_data data step.

a. Calculate body mass index (*BMI*) as: (weight/height^2) **[3 points]**

b. Create a new indicator variable called *obese*. Obesity should be defined as having a BMI of greater than or equal to 30 kg/m^2. If the subject is obese, the value for this variable should be 1, otherwise, non-obese individuals should be given a value of 0. **[5 points]**

c. Create a new indicator variable called *malesex*. This numeric variable should be coded as 1 if the subject is male and 0 if the subject is female. **[5 points]**

d. Create a new categorical variable called *season*, which represents the time-frame of the blood draw. Those with blood drawn in the winter months are expected to have lower concentrations of circulating Vitamin D, because of lower exposure to sunlight. **[5 points]**

Assign the four seasons using the following months of blood draw:

Winter: December, January, February

Spring: March, April, May

Summer: June, July, August

Fall: September, October, November

e. Apply appropriate formats to the *obese*, *diet*, *malesex, snp1,* and *snp2* variables. **[5 points]**

f. Delete any observations with missing values for *vitaminD* variable. **[5 points]**

Code:

```
*Question 2;

proc format ;
     value $malesex 'M'=1 'F'=0;
     value obesef 1="obese" 0="non-obese";
     value dietf 0="No" 1="Yes";
     value snp1f 1="genotype AA or AB" 0="genotype BB";
     value snp2f 1="genotype AA or AB" 0="genotype BB";
run;
data vit_data;
     set analysis.vitamind_sum2024;
     *Question 2a;
     bmi = (weight)/(height**2);

     *Question 2b;
     if bmi ge 30 then obese = 1;
     else if bmi lt 30 then obese = 0;

     *Question 2d;
     if month in (1,2,12) then season="Winter";
     else if month in (3,4,5) then season="Spring";
     else if month in (6,7,8) then season="Summer";
     else if month in (9,10,11) then season="Fall";

     *Question 2e;
```

```
      format sex $malesex. obese obesef. diet dietf. snp1 snp1f.
       snp2 snp2f.;

      *Question 2f;
      if vitaminD="." then delete;
run;
```

Section 2 – Univariate Statistics

3. Fill in the following Table 1 with the appropriate descriptive statistics (depending on if the variable is continuous or categorical) about the complete analytic sample using the dataset you created in Section 1. **[15 points]**

Table 1. Patient Characteristics

Characteristic	Statistics
Vitamin D (ng/mL)	37.73 (10.46)
Age (years)	42.45 (7.30)
Sex (male)	641 (42.85)
Obese (yes)	757 (50.60)
Vitamin D supplementation (yes)	442 (29.55)
Season of blood draw	
Winter	365 (24.40)
Spring	409 (27.34)
Summer	348 (23.26)
Fall	374 (25.00)
SNP1 (AA/AB genotype)	992 (66.31)
SNP2 (AA/AB genotype)	996 (66.58)

Code:

```
proc univariate data=vit_data;
      var vitaminD age;
run;
proc freq data=vit_data;
      tables sex obese diet season snp1 snp2;
```

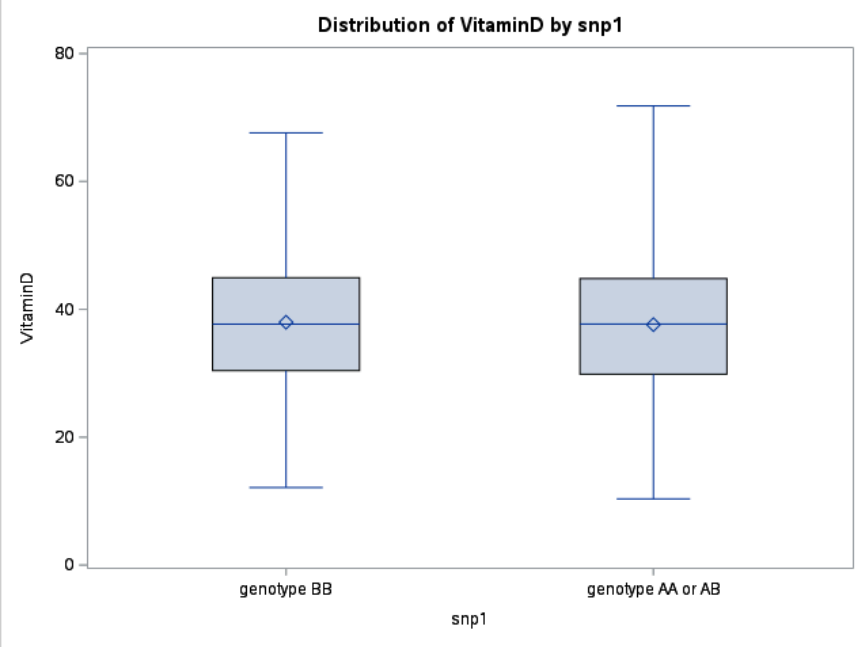
```
run;
```

Section 3 – Bivariate Associations

4. Now we wish to examine if there is an association between vitamin D and the two SNPs.

a. Create boxplots of blood levels of vitamin D stratified by the SNP genotypes (you should have two boxplots for each of the SNPs). Comment on what you observe from the box plots comparing the vitamin D levels for each of the SNP genotypes. **[15 points]**

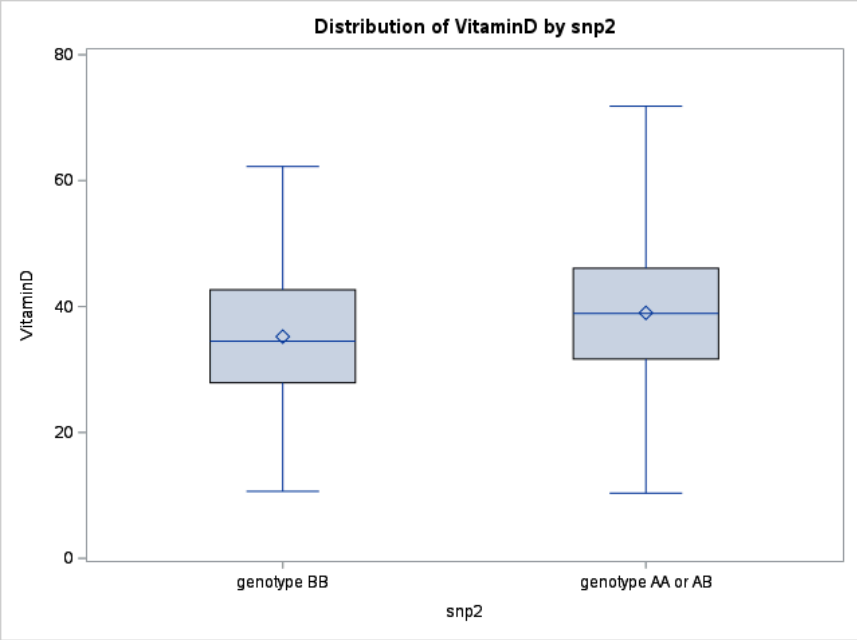
```
proc sort data=vit_data;
      by snp1;
run;
proc boxplot data=vit_data;
      plot vitaminD*snp1 / cboxes = black;
run;
```



Comment: The variances in vitamin D appear to be very similar across both groups (BB, and AA and AB) of genotypes for the first genetic marker (snp1). We can assume equal variances from this boxplot.

```
proc sort data=vit_data;
```

```
      by snp2;
run;
proc boxplot data=vit_data;
      plot vitaminD*snp2 / cboxes = black;
run;
```



Comment: The variances in vitamin D appear to be very similar across both groups (BB, and AA and AB) of genotypes for the first genetic marker (snp2). We can assume equal variances from this boxplot.

b. Fill in Tables 2 and 3 with the SNP-stratified descriptive statistics for the outcome, *vitaminD,* and each potential confounder of the association between the genetic factors and concentration of vitamin D: *age, malesex, obesity, supplementation,* and *season*. **[10 points]**

Code:

```
*Table 2;
proc sort data=vit_data;
      by snp1;
run;
proc means data=vit_data;
      var vitaminD age;
      by snp1;
```

```
run;
proc freq data = vit_data;
     by snp1;
     tables sex obese diet season;
run;
*Table 3;
proc sort data=vit_data;
     by snp2;
run;
proc means data=vit_data;
     var vitaminD age;
     by snp2;
run;
proc freq data = vit_data;
     by snp2;
     tables sex obese diet season;
run;
```

c. Test for association between each of the two SNP variables with the factors in tables 2 and 3 (separately). Write-up a report of the results in the last column of each table (2 and 3). Make sure to include the null and alternative hypotheses, test statistic, df, pvalue, an interpretation of the measures of effect, and conclusion. **[20 points]**

Code:

```
*Table2 t-tests;
proc sort data=vit_data;
     by snp1;
run;

proc boxplot data=vit_data;
     plot vitaminD*snp1 / cboxes=black;
run;
proc boxplot data=vit_data;
     plot age*snp1 / cboxes=black;
run;

proc ttest data=vit_data;
     class snp1;
     var vitaminD age;
run;

*Table3 t-tests;
proc sort data=vit_data;
     by snp2;
```

```
run;

proc boxplot data=vit_data;
     plot vitaminD*snp2 / cboxes=black;
run;
proc boxplot data=vit_data;
     plot age*snp2 / cboxes=black;
run;

proc ttest data=vit_data;
     class snp2;
     var vitaminD age;
run;

*Table 2 chi-squares;
proc sort data=vit_data;
     by descending obese descending sex descending
     diet descending season descending snp1;
run;
proc freq data=vit_data order=data;
     tables sex*snp1 obese*snp1 diet*snp1 season*snp1
     / chisq measures;
run;

*Table 3 chi-squares;
proc sort data=vit_data;
     by descending obese descending sex descending
     diet descending season descending snp2;
run;
proc freq data=vit_data order=data;
     tables sex*snp2 obese*snp2 diet*snp2 season*snp2
     / chisq measures;
run;
```

Table 2. Patient Characteristics Stratified by SNP1

Characteristic	Genotype AA or AB (n=992)	Genotype BB (n=504)	Test statistic, df, pvalue
Vitamin D (ng/mL)	37.60 (10.57)	37.98 (10.25)	0.65, 1494 df, p=0.5132
Age (years)	42.32 (7.34)	42.69 (7.23)	0.93, 1494 df, p=0.3509
Sex (male)	425 (42.84)	216 (42.86)	0.000, 1 df, p=0.9958
Obese (yes)	494 (49.80)	263 (52.18)	0.7600, 1 df, p=0.3833
Vitamin D supplementation (yes)	288 (29.03)	154 (30.56)	0.3725, 1 df, p=0.5416

Season			
Winter	241 (24.29)	124 (24.60)	
Spring	266 (26.81)	143 (28.37)	0.6752, 3 df, p=0.8790
Summer	236 (23.79)	112 (22.22)	
Fall	249 (25.10)	125 (24.80)	

Write Up:

The 1,496 individuals sampled were divided by their genotypes for the first genetic marker (SNP1), with 992 having genotype AA or AB and 504 having genotype BB. All descriptive statistics are shown in Table 2. For the two continuous variables Vitamin D and Age, descriptive statistics were calculated using means and standard deviations. For the categorical variables Sex, Obese, Vitamin D supplementation, and Season, descriptive statistics were calculated using counts and percentages.

The null hypothesis for both continuous variables is that the mean vitamin D blood concentration and mean age would be equal between both groups of genotypes. We ran a two-sample test of means to test the alternative hypothesis that the mean vitamin D concentration and mean age are different between genotypes. First, we examined the variability between both groups for the two variables by generating boxplots, both of which were determined to have similar variances. We therefore used a t-test assuming equal variances to test the alternative hypothesis at the 0.05 level. The mean vitamin D concentration in individuals with genotype AA or AB (37.60 ng/mL ± 10.57) was not significantly different than the vitamin D concentration in individuals with genotype BB (37.98 ng/mL ± 10.25), t-value=0.65, df=1,494, p=0.5132. Similarly, the mean age of individuals with genotype AA or AB (42.32 years ±7.34) was not significantly different than the mean age of individuals with genotype BB (42.69 years ± 7.23), t-value=0.93, df=1,494, p=0.3509. We therefore fail to reject the null hypothesis for both Vitamin D concentration and Age, since both p-values (0.5132 and 0.3509, respectively) are above the 0.05 level of significance.

For the categorical variables Sex, Obese, Vitamin D supplementation, and Season, the null hypothesis for each of these variables is that the odds of being male, being obese, taking vitamin D supplementation, or completing the exam during a specific seasonal category in the genotype AA and AB group is the same in the genotype BB group (OR=1). We therefore used a chi-square test for each variable to test the alternative hypothesis, which states that the odds of being male, being obese, taking vitamin D supplements, or completing the exam during a specific seasonal category in the genotype AA and AB group is not the same in the genotype BB group (OR≠1). For Sex, the estimated OR=0.9994 (95% CI = 0.8047, 1.2412) with the odds of being male in the group with the AA or AB genotype being 0.9994 times the odds of being male in the group with the genotype BB. For Obesity, the estimated OR=0.9090 (95% CI = 0.7335, 1.1265) with the odds of being obese in the group with the AA or AB genotype being 0.9090

times the odds of being obese in the group with the BB genotype. For Vitamin D supplementation, the estimated OR=0.9298 (95% CI = 0.7358, 1.1748) with the odds of taking a vitamin D supplement in the AA or AB genotype group being 0.9298 times the odds of taking a vitamin D supplement in the BB genotype group. We fail to reject the null hypothesis for each variable. There is no significant evidence (at α=0.05) that the OR is not equal to 1 for Sex (t=0.000, 1 df, p=0.995), Obesity (t=0.7600, 1 df, p=0.3833), Vitamin D supplementation (t=0.3725, 1 df, p=0.5416), or season of exam (t=0.6752, 3 df , p=0.8790).

Table 3. Patient Characteristics Stratified by SNP2

Characteristic	**Genotype AA or AB (n=996)**	**Genotype BB (n=500)**	**Test statistic, df, pvalue**
Vitamin D (ng/mL)	38.99 (10.31)	35.22 (10.32)	-6.67, 1494 df, p<0.0001
Age (years)	42.50 (7.30)	42.34 (7.32)	-0.38, 1494 df, p=0.7059
Sex (male)	414 (41.57)	227 (45.40)	1.9979, 1 df, p=0.1575
Obese (yes)	490 (49.20)	267 (53.40)	2.3528, 1 df, p=0.1251
Vitamin D supplementation (yes)	294 (29.52)	148 (29.60)	0.0011, 1 df, p=0.9739
Season Winter Spring Summer Fall	 241 (24.20) 280 (28.11) 235 (23.59) 240 (24.10)	 124 (24.80) 129 (25.80) 113 (22.60) 134 (26.80)	 1.8155, 3 df, p=0.6116

Write Up:

The 1,496 individuals sampled were divided by their genotypes for the first genetic marker (SNP2), with 996 having genotype AA or AB and 500 having genotype BB. All descriptive statistics are shown in Table 2. For the two continuous variables Vitamin D and Age, descriptive statistics were calculated using means and standard deviations. For the categorical variables Sex, Obese, Vitamin D supplementation, and Season, descriptive statistics were calculated using counts and percentages.

The null hypothesis for both continuous variables is that the mean vitamin D blood concentration and mean age would be equal between both groups of genotypes. We ran a two-sample test of means to test the alternative hypothesis that the mean vitamin D concentration and mean age are different between genotypes. First, we examined the variability between both groups for the two variables by generating boxplots, both of which were determined to have similar variances. We therefore used a t-test assuming equal variances to test the alternative hypothesis at the 0.05 level. **The mean vitamin D concentration in individuals with genotype AA or AB (38.99 ng/mL ± 10.31) was significantly different than the vitamin D**

concentration in individuals with genotype BB (35.22 ng/mL ± 10.32), t-value=-6.67, df=1,494, p<0.0001. We therefore reject the null hypothesis of no difference between means as the p-value is less than 0.05. However, the mean age of individuals with genotype AA or AB (42.50 years ±7.30) was not significantly different than the mean age of individuals with genotype BB (42.34 years ± 7.32), t-value=-0.38, df=1,494, p=0.7059. We therefore fail to reject the null hypothesis for Age, since the p-value of 0.7059 is greater than 0.05.

For the categorical variables Sex, Obese, Vitamin D supplementation, and Season, the null hypothesis for each of these variables is that the odds of being male, being obese, taking vitamin D supplementation, or completing the exam during a specific seasonal category in the genotype AA and AB group is the same in the genotype BB group (OR=1). We therefore used a chi-square test for each variable to test the alternative hypothesis, which states that the odds of being male, being obese, taking vitamin D supplements, or completing the exam during a specific seasonal category in the genotype AA and AB group is not the same in the genotype BB group (OR≠1). For Sex, the estimated OR=0.8555 (95% CI = 0.6890, 1.0623) with the odds of being male in the group with the AA or AB genotype being 0.8555 times the odds of being male in the group with the genotype BB. For Obesity, the estimated OR=0.8451 (95% CI = 0.6815, 1.0480) with the odds of being obese in the group with the AA or AB genotype being 0.8451 times the odds of being obese in the group with the BB genotype. For Vitamin D supplementation, the estimated OR=0.9961 (95% CI = 0.7872, 1.2604) with the odds of taking a vitamin D supplement in the AA or AB genotype group being 0.7872 times the odds of taking a vitamin D supplement in the BB genotype group. We fail to reject the null hypothesis for each variable. There is no significant evidence (at α=0.05) that the OR is not equal to 1 for Sex (t=1.9979, 1 df, p=0.1575), Obesity (t=2.3528, 1 df, p=0.1251), Vitamin D supplementation (t=0.0011, 1 df, p=0.9739), or Season of exam (t=1.8155, 3 df , p=0.6116).

Whole Code for Project 1:

```
*Section 1;
*Question 1;
libname analysis '/home/u63889606/BS 723/Project 1';
run;

*Question 2;

proc format ;
      value $malesex 'M'=1 'F'=0;
      value obesef 1="obese" 0="non-obese";
      value dietf 0="No" 1="Yes";
      value snp1f 1="genotype AA or AB" 0="genotype BB";
      value snp2f 1="genotype AA or AB" 0="genotype BB";
run;
data vit_data;
```

```
      set analysis.vitamind_sum2024;
      *Question 2a;
      bmi = (weight)/(height**2);

      *Question 2b;
      if bmi ge 30 then obese = 1;
      else if bmi lt 30 then obese = 0;

      *Question 2d;
      if month in (1,2,12) then season="Winter";
      else if month in (3,4,5) then season="Spring";
      else if month in (6,7,8) then season="Summer";
      else if month in (9,10,11) then season="Fall";

      *Question 2e;
      format sex $malesex. obese obesef. diet dietf. snp1 snp1f. snp2
snp2f.;

      *Question 2f;
      if vitaminD="." then delete;
run;

*Section 2, Question 3;
proc univariate data=vit_data;
      var vitaminD age;
run;
proc freq data=vit_data;
      tables sex obese diet season snp1 snp2;
run;

*Section 3 Question 4;
*4a;
proc sort data=vit_data;
      by snp1;
run;
proc boxplot data=vit_data;
      plot vitaminD*snp1 / cboxes = black;
run;

proc sort data=vit_data;
      by snp2;
run;
proc boxplot data=vit_data;
      plot vitaminD*snp2 / cboxes = black;
run;

*Section 3 Q 4b;
*Table 2;
proc sort data=vit_data;
      by snp1;
run;
proc means data=vit_data;
```

```
      var vitaminD age;
      by snp1;
run;
proc sort data=vit_data;
      by snp1;
run;
proc freq data = vit_data;
      by snp1;
      tables sex obese diet season;
run;
*Table 3;
proc sort data=vit_data;
      by snp2;
run;
proc means data=vit_data;
      var vitaminD age;
      by snp2;
run;
proc sort data=vit_data;
      by snp2;
run;
proc freq data = vit_data;
      by snp2;
      tables sex obese diet season;
run;

*Table2 t-tests;
proc sort data=vit_data;
      by snp1;
run;

proc boxplot data=vit_data;
      plot vitaminD*snp1 / cboxes=black;
run;
proc boxplot data=vit_data;
      plot age*snp1 / cboxes=black;
run;

proc ttest data=vit_data;
      class snp1;
      var vitaminD age;
run;

*Table3 t-tests;
proc sort data=vit_data;
      by snp2;
run;

proc boxplot data=vit_data;
      plot vitaminD*snp2 / cboxes=black;
run;
proc boxplot data=vit_data;
```

```
      plot age*snp2 / cboxes=black;
run;

proc ttest data=vit_data;
      class snp2;
      var vitaminD age;
run;

*Table 2 chi-squares;
proc sort data=vit_data;
      by descending obese descending sex descending diet descending
season descending snp1;
run;
proc freq data=vit_data order=data;
      tables sex*snp1 obese*snp1 diet*snp1 season*snp1 / chisq
measures;
run;

*Table 3 chi-squares;
proc sort data=vit_data;
      by descending obese descending sex descending diet descending
season descending snp2;
run;
proc freq data=vit_data order=data;
      tables sex*snp2 obese*snp2 diet*snp2 season*snp2 / chisq
measures;
run;
```