Project for BS805 _ Fall 2022

Data Overview.

In this study, our interest is to examine the associations between plasma homocysteine levels and cognition function for people. The dataset we will use is a subset of the data used in the Framingham Study which found increased plasma homocysteine is an independent risk factor for the development of Alzheimer's disease (NEJM 2002).

The three separate datasets named *DEMOG_BS805_F22.SAS7BDAT*, *LABS_BS805_F22.SAS7BDAT*, and *NEURO_BS805_F22.SAS7BDAT*, respectively. Each of them has 900 subjects, and the number of variables for them are five, five, and three respectively. The details of them are shown in table 1.

Sample size: Subjects = 900 for all three data sets.

Table 1 The description of all three original dataset

DATASET NAME	VARIABLES	DESCRIPTION
DEMOG_BS805_F22	DEMOGID	Subject ID in Demographics data set
	AGE	age in years
	MALE	=1 if male, =0 if female
	EDUCG	=1 if education < 8 years, =2 if education >=8 years but no HS degree, =3 if HS degree but no college, = 4 if at least some college
	PKYRS	pack years of cigarette smoking
LABS_BS805_F22	LABSID	Subject ID in Labs data set
	HCY	plasma homocysteine level (μmol/L)
	FOLATE	plasma folate (nmol/mL)
	VITB12	plasma vitamin B12 (pmol/L)
	VITB6	plasma vitamin B6 (nmol/L)
NEURO_BS805_F22	NEUROID	Subject ID in Neuro data set
	MMSE	Mini-Mental State Examination (a measure of cognitive function with range 0-30)
	ADIN7YRS	=0 if no AD in 7 years of follow-up, =1 if AD in 7 years of follow-up

1. Combine the three data sets

After importing the three original files, three temporary SAS data sets named '*dem*' (with 900 subjects and 5 variables), '*lab*' (with 900 subjects and 5 variables), '*neu*' (with 900 subjects and 3 variables) was obtained. For the convenience of subsequent process, the names of the variables indicating ID in the three files 'DEMOGID', 'LABSID', and 'NEUROID' were all uniformly renamed as 'ID'.

Then, after sorting these three data sets using the proc sort procedure according to the 'ID' variable, the three data sets were finally combined into one single, temporary SAS data set, which named '*question1*' (with 900 subjects and 11 variables), using the merge statement.

2. Create new variables

A permanent SAS dataset named 'question2' (with 900 subjects and 17 variables) was created and stored in the library 'pjt805', which containing 6 new variables and the details of them are shown in the table 2.

Table 2 The description of the 6 new variables just created

DATASET	NEW VARIABLES	DESCRIPTION
QUESTION2	LHCY	the natural log of HCY
	HCYGE14	=1 for those whose HCY is at least 14; 0 for those whose homocysteine is less than 14
	AGEGRP	1 for age 65-74, 2 for age 75-79, 3 for age 80-84, 4 for age 85-89
	HSDEG	1 for High school degree or higher; 0 for Less than high school degree
	EXCLUDE	1 for those with missing ADIN7YRS, and 0 for those with non-missing ADIN7YRS
	MMSEF	flags subjects with cognitive deficits according to the MMSE

3. Perform statistical hypothesis tests to compare those excluded to those not excluded

For those excluded, there are 237 total subjects, whereas there are 237 total subjects for those not excluded. To compare the distribution between those excluded and those not excluded with respect to age (AGE), sex (MALE), education (use HSDEG), cigarette smoking (use PKYRS), cognitive status (MMSE), and homocysteine (use both LHCY and HCYGE14), different process would be performed.

(1) For continuous variable 'AGE' (using T-test): there are 237 values (no missing data) for those excluded, whereas there are 663 values (no missing data) for those not excluded. We have enough evidence to reject the null hypothesis $\boldsymbol{H_0}$ that the mean age of those excluded is identical to the mean age of those not excluded and conclude that the mean age for those excluded and not excluded is significantly different (p <0.0001, df=898, t-value= -4.63). Based on the Folded F test (p=0.0162<0.05, F=1.29), the Satterthwaite version of the test statistic was used. Therefore, the average age for those not excluded was about 1.67 years younger than the average age for those excluded (95% confidence interval -2.43, -0.92).

(2) For continuous variable 'PKYRS' (using T-test): there are 215 values (22 missing data) for excluded, whereas there are 596 values (67 missing data) for those not excluded. We fail to reject the null hypothesis $\boldsymbol{H_0}$ that the mean PKYRS of those excluded is identical to the mean age of those not excluded and could conclude that the mean PKYRS for those excluded and not excluded is significantly identical (p=0.0687>0.05, df=809, t-value= -1.82). Based on the Folded F test (p=0.0801>0.05, F=1.21), the Pooled

version of the test statistic was used. Therefore, the average PKYRS for those not excluded was about 3.19 lesser than the average PKYRS for those excluded (95% confidence interval -6.63, 0.25).

(3) For continuous variable 'MMSE' (using T-test): there are 234 values (3 missing data) for excluded, whereas there are 661 values (2 missing data) for those not excluded. We have enough evidence to reject the null hypothesis H_0 that the mean MMSE of those excluded is identical to the mean age of those not excluded and conclude that the mean MMSE for those excluded and not excluded is significantly different ($p<0.0001$, df=302.8, t-value=3.86). Based on the Folded F test ($p<0.0001$, F =2.46), the Satterthwaite version of the test statistic was used. Therefore, the average MMSE for those not excluded was about 1.16 larger than the average MMSE for those excluded (95% confidence interval 0.57, 1.76).

(4) For continuous variable 'LHCY' (using T-test): there are 237 values (no missing data) for excluded, whereas there are 663 values (no missing data) for those not excluded. We have enough evidence to reject the null hypothesis H_0 that the mean LHCY of those excluded is identical to the mean age of those not excluded and conclude that the mean LHCY for those excluded and not excluded is significantly different ($p=0.0036<0.05$, df=898, t-value= -2.92). Based on the Folded F test ($p=0.3822>0.05$, F =1.10), the Pooled version of the test statistic was used. The average LHCY for those not excluded was about 0.09 lesser than the average LHCY for those excluded (95% confidence interval -0.14, -0.03).

(5) For categorical variable 'MALE' (using Chi-Square test): there are 237 values (no missing data) for those excluded, whereas there are 663 values (no missing data) for those not excluded. For those excluded, the proportion of female (MALE=0) is 51.90% (123 of 237), and the proportion of male (MALE =1) is 48.10% (114 of 237). For those not excluded, the proportion of female (MALE=0) is 65.91% (437 of 663), and the proportion of male (MALE =1) is 34.09% (226 of 663). We have enough evidence to reject the null hypothesis H_0 that gender is not associated with variable EXCLUDE and conclude that there is a significant association between sex and whether subjects are excluded ($X^2(1) = 14.5864$, $p=0.0001<0.05$).

(6) For categorical variable 'HSDEG' (using Chi-Square test): there are 234 values (3 missing data) for those excluded, whereas there are 651 values (12 missing data) for those not excluded. For those excluded, the proportion of those who have less than high school degree (HSDEG=0) is 35.04% (82 of 234), and the proportion of those who have high school degree or higher (HSDEG =1) is 64.96% (152 of 234). For those not excluded, the proportion of those who have less than high school degree (HSDEG=0) is 30.41% (198 of 651), and the proportion of those who have high school degree or higher (HSDEG =1) is 69.59% (453 of 651). We fail to reject the null hypothesis H_0 that education is not associated with variable EXCLUDE and conclude that there is no significant association between education and whether subjects are excluded ($X^2(1) =1.7046$, $p=0.1917>0.05$).

(7) For categorical variable 'HCYGE14' (using Chi-Square test): there are 237 values (no missing data) for those excluded, and there are 663 values (no missing data) for those not excluded. For those excluded, the proportion of those whose homocysteine is less than 14 (HCYGE14=0) is 67.09% (159 of 237), and the proportion of those whose HCY is at least 14 (HCYGE14=1) is 32.91% (78 of 237). For those not

excluded, the proportion of those whose homocysteine is less than 14 (HCYGE14=0) is 74.96% (497 of 663), and the proportion of those whose HCY is at least 14 (HCYGE14=1) is 25.04% (166 of 663). We have enough evidence to reject the null hypothesis $\boldsymbol{H_0}$ that categorical homocysteine is not associated with variable EXCLUDE and conclude that there is a significant association between categorical homocysteine and whether subjects are excluded ($X^2(1) = 5.4773$, p= 0.0193 <0.05).

4. Create a new data set excluding those with missing ADIN7YRS values

At first, there are 237 subjects have missing ADIN7YRS variable, and 663 subjects have non-missing ADIN7YRS variable. After excluding the subjects whose ADIN7YRS variable is missing and drop the column 'EXCLUDE', a new temporary data set named '*qustion4*' was created, which finally have 663 subjects and 16 variables (ID, AGE, MALE, EDUCG, PKYRS, FOLATE, VITB6, VITB12, HCY, MMSE, ADIN7YRS, LHCY, HCYGE14, AGEGRP, HSDEG, and MMSEF).

5. Vertical bar charts and Descriptive statistics

The vertical bar charts of the five continuous variables are shown in Figure 1.

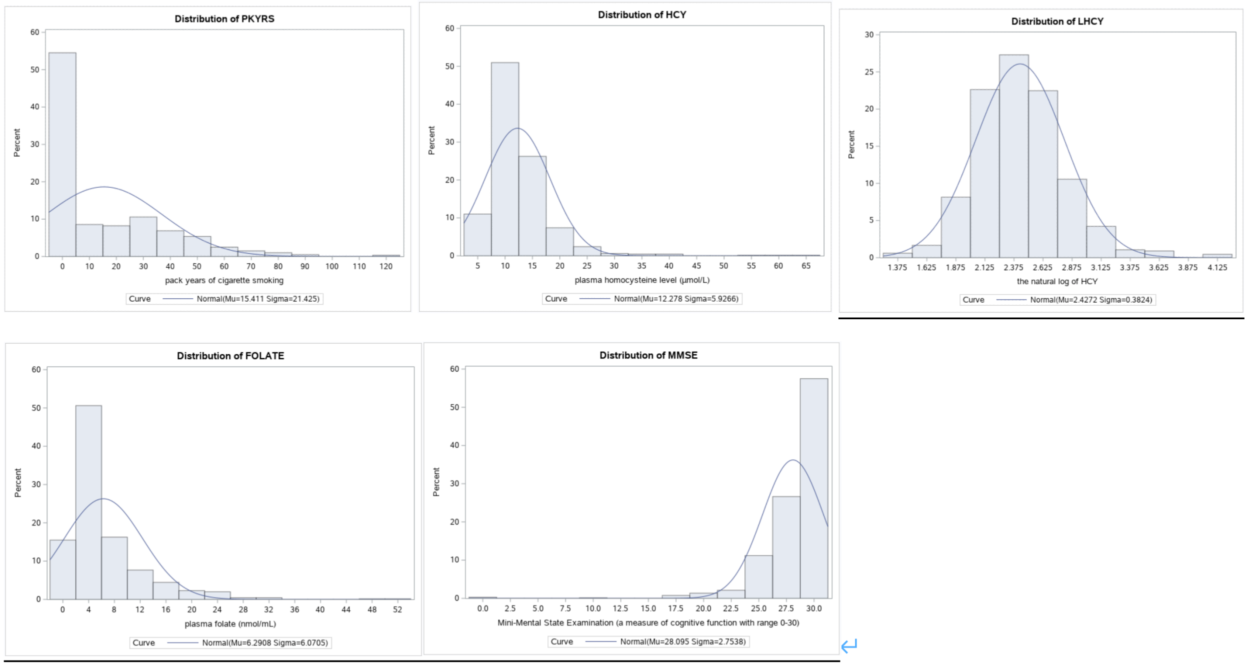

Figure 1 vertical bar charts for 5 variables

And since all the five variables (PKYRS, HCY, LHCY, FOLATE, and MMSE) are all continuous variables, so we could generate descriptive statistics for each of them using PROC UNIVARIATE procedure, which are shown as Table 3.

For PKYRS, there are 596 values for this variable (with 67 missing data), and the Mean is 15.41, Standard Deviation is 21.43, Median is 0.61, Mode is 0.00, and the Interquartile Range is 28.20. For HCY, there are 663 values for this variable (no missing data), and the Mean is 12.28, Standard Deviation is 5.93, Median is 11.00, Mode is 8.50, and the Interquartile Range is 5.10. For LHCY, there are 663 values for this variable (no missing data), and the Mean is 2.43, Standard Deviation is 0.38, Median is 2.40, Mode is 2.14, and the Interquartile Range is 0.45.For FOLATE, there are 652 values for this variable (with 11 missing data), and the Mean is 6.29, Standard Deviation is 6.07, Median is 4.21, Mode is 2.00, and the Interquartile Range is 5.39.For MMSE, there are 661 values for this variable (with 2 missing data), and the Mean is 28.10, Standard Deviation is 2.75, Median is 29.00, Mode is 30.00, and the Interquartile Range is 3.00.

Table 3 descriptive statistics for 5 variables in Q5

VARIABLE	N	MISSING-	MEAN	STD DEV	MEDIAN	MODE	IQR
PKYRS	596	67	15.41	21.43	0.61	0.00	28.20
HCY	663	0	12.28	5.93	11.00	8.50	5.10
LHCY	663	0	2.43	0.38	2.40	2.14	0.45
FOLATE	652	11	6.29	6.07	4.21	2.00	5.39
MMSE	661	2	28.10	2.75	29.00	30.00	3.00

6. Both homocysteine and cognitive function may change with age

6.1 Test if LHCY (dependent) linearly associated with continuous age

A simple linear regression analysis was performed with outcome variable ‘LHCY’ and the ‘AGE’ variable. For this analysis, there are no missing values for LHCY, so the total number of observations used are 663.

Table 4 Simple linear regression of LHCY with continuous age

	DF	**F**	**P**	**N_obs**	**R^2**		
Model	1	17.93	<0.0001	663	0.0264		
Error	661						
	Beta estimate	**95% Lower CI**	**95% upper CI**	**Standard Error**	**Standard Estimate**	**t**	**p**
AGE	0.01349	0.00724	0.01975	0.00319	0.16253	4.23	<0.0001

H_0: There is no linear association between LHCY and continuous age.

H_1: There is a linear association between LHCY and continuous age.

Results: In the linear regression analysis of the association of 'LHCY' and 'AGE', we found that there was a significant linear association at the 0.05 level (P < 0.0001, F-value=17.93, df=(1, 661), R^2=0.0264). The β-estimated (slope) = 0.0135 (P<0.0001, t-value=4.23, df=661, SE=0.1625, Std Error=0.0032; 95% confidence interval 0.0072, 0.01975). Hence, we have strong evidence to reject the null hypothesis *H_0* and conclude that there is a linear association between LHCY and age. The R^2=0.0264 for this model indicates that the variable 'AGE' accounts for 2.64% of the variability of the variable 'LHCY' in the dataset. The β-estimated means that, for a one-unit increase in age, on average, LHCY increases by 0.0135.

To sum up, we have strong evidence to conclude that LHCY is linearly associated with continuous age.

6.2 Test if MMSE (dependent) linearly associated with continuous age

A simple linear regression analysis was performed with outcome variable 'MMSE' and the 'AGE' variable. For this analysis, there are 2 missing values for MMSE, so the total number of observations used are 661.

Table 5 Simple linear regression of MMSE with continuous age

	DF	F	P	N_obs	R^2		
Model	1	38.02	<0.0001	661	0.0546		
Error	659						
	Beta estimate	**95% Lower CI**	**95% upper CI**	**Standard Error**	**Standard Estimate**	**t**	**p**
AGE	-0.14061	-0.18538	-0.09583	0.02280	-0.23357	-6.17	<0.0001

H_0: There is no linear association between MMSE and continuous age.

H_1: There is a linear association between MMSE and continuous age.

Results: In the linear regression analysis of the association of 'MMSE' and 'AGE', we found that there was a significant linear association at the 0.05 level (P < 0.0001, F-value=38.02, df=(1, 659), R^2=0.0546). The β-estimated (slope) = -0.1406 (P<0.0001, t-value= -6.17, df=659, SE= -0.2336, Std Error=0.0228; 95% confidence interval -0.1854, 0.0958). Hence, we have strong evidence to reject the null hypothesis *H_0* and conclude that there is a linear association between MMSE and age. The R^2=0.0546 for this model indicates that the variable 'AGE' accounts for 5.46% of the variability of the variable 'LHCY' in the dataset. The β-estimated means that, for a one-unit increase in age, on average, MMSE decreases by 0.1406.

To sum up, we have strong evidence to conclude that MMSE is linearly associated with continuous age.

7. Test if mean LHCY are the same in the four age groups

To determine whether the mean LHCY differed across the four age groups, we performed a one-way ANOVA with the outcome variable 'LHCY' and one independent factor 'AGEGRP'. For this analysis, there are no missing values for LHCY, so the total number of observations used are 663.

Table 6 Linear regression of LHCY with categorical age

	DF	F	P	N_obs	R^2
Model	3	6.64	0.0002	663	0.0293
Error	659				
AGEGRP	3	6.64	0.0002		
	Estimate	**Std Err**	**t**	**p**	
AGEGRP 1	-0.1475	0.0680	-3.64	0.0003	
AGEGRP 2	-0.2103	0.0700	-3.00	0.0028	
AGEGRP 3	-0.1110	0.0760	-1.46	0.1448	
AGEGRP 4	0.0000	.	.	.	

H_0: The mean LHCY are the same in all age groups.

H_1: At least one age group differs from the others with respect to mean LHCY.

Results: Reject the null hypothesis. We found significant evidence of an overall difference in LHCY between age groups ($P=0.0002<0.05$; F-value=6.64, df=(3, 659), $R^2=0.0293$) at the 0.05 significance level. Hence, we have strong evidence to reject ***H_0*** and conclude that there is at least one group differs from the others with respect to mean LHCY. The $R^2=0.0293$ for this model indicates that the variable '*AGEGRP*' accounts for 2.93% of the variability of the variable in LHCY.

Test between age group 1 and 4: Reject the null hypothesis that the mean LHCY are the same for group 1 and group 4. The age group 1 (age 65-74) took a statistically significantly lower LHCY (on average 0.2475 less) than age group 4 (age 85-89), with $P=0.0003<0.05$, t-value=-3.64, df=659, Std Error=0.0680.

Test between age group 2 and 4: Reject the null hypothesis that the mean LHCY are the same for group 2 and group 4. The age group 2 (age 75-79) took a statistically significantly lower LHCY (on average 0.2103 less) than age group 4 (age 85-89), with $P=0.0028<0.05$, t-value=-3.00, df=659, Std Error =0.0700.

Test between age group 3 and 4: Fail to reject the null hypothesis that the mean LHCY are the same for group 3 and group 4. The age group 3 (age 80-84) took a slightly lower LHCY (on average 0.1110 less) than age group 4 (age 85-89), with $P=0.1448>0.05$, t-value=-1.46, df=659, Std Error =0.0760.

Tukey's post hoc test: To identify specific between-group differences, the Tukey's post hoc procedure was employed. We found that age group 1 (age 65-74) took a statistically lower LHCY (on average 0.25 less) than age group 4 (age 85-89) at 0.05 level. The age group 2 (age 75-79) took a statistically lower LHCY (on average 0.14 less) than age group 4 (age 85-89) at 0.05 level. And the age group 3 (age 80-84) took a slightly lower LHCY (on average 0.04 less) than age group 4 (age 85-89), which is not significant at the 0.05 confidence level.

And the mean LHCY for each age group and standard deviation was as follows:

Age Group 1 (age 65-74): 2.38 ± 0.37 (n=337),

Age Group 2 (age 75-79): 2.42 ± 0.34 (n=202),

Age Group 3 (age 80-84): 2.52 ± 0.45 (n=90),

Age Group 4 (age 85-89): 2.63 ± 0.45 (n=34).

In conclusion, we have strong evidence to conclude that there exist at least one age group differs from the others with respect to mean LHCY. And after performing the Tukey's post hoc test, we could find that the mean LHCY for age group 3 (age 80-84) and age group 4 (age 85-89) are statistically identical; whereas the mean LHCY for age group 1 (age 65-74) and age group 4 (age 85-89) are statistically different, and the mean LHCY for age group 2 (age 75-79) and age group 4 (age 85-89) are also statistically different.

The result supports a linear relationship between age and log homocysteine.

8. Piecewise linear model of log homocysteine and age

After creating four new continuous variables called age1, age2, age3, and age4, the piecewise linear model using age1, age2, age3, and age4 to predict LHCY was performed. For this analysis, there are no missing values for LHCY, so the total number of observations used are 663.

Table 7 Piecewise linear model of log homocysteine and age

	DF	**F**	**P**	**N_obs**	**R^2**		
Model	4	4.91	0.0007	663	0.0290		
Error	658						
	Beta estimate	**95% Lower CI**	**95% upper CI**	**Standard Error**	**Standard Estimate**	**t**	**p**
age1	0.00471	-0.01350	0.02291	0.00927	0.02483	0.51	0.6119
age2	0.01531	-0.00782	0.03844	0.01178	0.08037	1.30	0.1943

age3	0.02345	-0.01395	0.06086	0.01905	0.07792	1.23	0.2188
age4	0.02027	-0.06664	0.10719	0.04426	0.02304	0.46	0.6471

H_0: There is no association between LHCY and age1, age2, age3, and age4.

H_1: There is association between LHCY and age1, age2, age3, and age4.

<u>*Results:*</u> Reject H_0 and have strong evidence to conclude that there is association between LHCY and age1, age2, age3, and age4 ($P=0.0007<0.05$, F-value=4.91, df=(4,658), $R^2=0.0290$). The $R^2 = 0.0290$ for this model indicates nearly 2.90% of the variability in LHCY was accounted for by age1, age2, age3, and age4.

<u>*Test for 'age1'*</u>: When testing the null hypothesis H_0 that there is no association between LHCY and those who age from 65 to 74, we fail to reject the null hypothesis (β-estimate=0.0047, 95% confidence interval -0.0135, 0.0229; P=0.6119 > 0.05, t-value=0.51, df=658, SE=0.0248, Std Error=0.0093). The alternative hypothesis H_1 is that there is association between log homocysteine and those who age from 65 to 74.

<u>*Test for 'age2'*</u>: When testing the null hypothesis H_0 that there is no association between LHCY and those who age from 75 to 79, we fail to reject the null hypothesis (β-estimate=0.0153, 95% confidence interval -0.0078, 0.0384; P=0.1943>0.05, t-value=1.30, df=658, SE=0.0804, Std Error=0.0118). The alternative hypothesis H_1 is that there is association between log homocysteine and those who age from 75 to 79.

<u>*Test for 'age3'*</u>: When testing the null hypothesis H_0 that there is no association between LHCY and those who age from 80 to 84, we fail to reject the null hypothesis (β-estimate= 0.0235, 95% confidence interval -0.0140, 0.0609; P=0.2188>0.05, t-value=1.23, df=658, SE=0.0779, Std Error=0.0191). The alternative hypothesis H_1 is that there is association between log homocysteine and those who age from 80 to 84.

<u>*Test for 'age4'*</u>: When testing the null hypothesis H_0 that there is no association between LHCY and those who age from 85 to 89, we fail to reject the null hypothesis (β-estimate= 0.0203, 95% confidence interval -0.0666, 0.1072; P=0.6471>0.05, t-value=0.46, df=658, SE=0.0230, Std Error=0.0443). The alternative hypothesis H_1 is that there is association between LHCY and those who age from 85 to 89.

To sum up, though the overall model is statistically significant at 0.05 confidence level, the estimated individual slopes for age1, age2, age3, and age4 were not statistically significant. And the piecewise model generates a higher R^2 (0.0290 > 0.0264) than the previous simple linear regression model in Q6.1. I prefer to choose the categorical variable 'AGEGRP' (in Q7, the mean LHCY differences analysis by the dummy regression using PROC GLM with a class statement) for this dataset later, because it has higher R^2 (0.0293 with P=0.0002, which is larger than 0.0264 with P<0.0001 for simple linear regression and larger than 0.0290 with P=0.0007 for piecewise regression) compared with other two regression models. A higher R^2 indicates that the variability in log homocysteine could be better accounted for by the categorical variable 'AGEGRP'.

9. A multiple linear regression with interaction with a dummy variable for gender

To determine whether there is an interaction between log homocysteine and gender for the dependent variable MMSE, a linear regression with the outcome variable ‘MMSE’ and three predictors (LHCY, MALE, and LHCY*MALE) was performed using PROC GLM with a class statement. In the dataset, there is 2 missing MMSE values and then there are 661 observations used in the model.

Table 8 multiple linear regression with interaction with a dummy variable for gender

	DF	F	P	N_obs	R^2
Model	3	4.81	0.0025	661	0.0215
Error	657				
LHCY*MALE	1	0.11	0.7349		
	Estimate	**Std Err**	**t**	**p**	
LHCY*MALE 0	0.2005	0.5918	0.34	0.7349	
LHCY*MALE 1	0.0000	.	.	.	

H_0: There is no association between MMSE and LHCY, MALE, and interaction of LHCY and MALE.

H_1: There is an association between MMSE and LHCY, MALE, and interaction of LHCY and MALE.

Results: Reject the global null hypothesis. We have significant evidence to reject H_0 (P=0.0025<0.05; F-value=4.81, df=(3,657), R^2=0.0215) and conclude that there is an association between MMSE and at least one of the three predictors (LHCY, MALE, and LHCY*MALE). The R^2=0.0215 for this model indicates that the three predictors account for 2.15% of the variability of the variable in MMSE.

Test for interaction between log homocysteine (LHCY) and sex (MALE): When testing the null hypothesis H_0 that there is no association between MMSE and the interaction effect ‘LHCY*MALE’, we **fail to reject** the null hypothesis (P=0.7349>0.05, df=(1, 659), F-value=0.11) and could conclude that there is no association between MMSE and the interaction effect of LHCY and MALE. The alternative hypothesis H_1 is that there is an association between MMSE and the interaction effect ‘LHCY*MALE’.

Test for interaction between log homocysteine (LHCY) and female (MALE 0): When testing the null hypothesis H_0 that there is no association between MMSE and the interaction effect ‘LHCY*MALE_0’, (male_0 indicates the subjects who is female), we fail to reject the null hypothesis (P=0.7349>0.05, df= 659, t-value=0.34, Std Error=0.5918). The alternative hypothesis H_1 is that there is an association between MMSE and the interaction effect ‘LHCY*MALE_0’.

In conclusion, there is no evidence to say the interaction effect exists between log homocysteine (LHCY) and gender (MALE) for the dependent variable MMSE. Therefore, gender would not modify the effect of LHCY on MMSE since there is no interaction effect in this model. There is no significant evidence of effect modification by gender on this relationship.

10. A linear regression model with MMSE and LHCY

A simple linear regression analysis was performed with outcome ‘MMSE’ and the ‘LHCY’ variable. For this analysis, there are 2 missing values for MMSE, so the total number of observations used are 661.

Table 9 linear regression model of MMSE with LHCY

	DF	**F**	**P**	**N_obs**	**R^2**		
Model	1	9.81	0.0018	661	0.0147		
Error	659						
	Beta estimate	**95% Lower CI**	**95% upper CI**	**Standard Error**	**Standard Estimate**	**t**	**p**
LHCY	-0.87151	-1.41785	-0.32517	0.27824	-0.12112	-3.13	0.0018

H_0: There is no linear association between MMSE and LHCY.

H_1: There is a linear association between MMSE and LHCY.

Results: In the linear regression analysis of the association of MMSE and LHCY, we found that there was a significant linear association at the 0.05 level (P=0.0018<0.05, F-value=9.81, df=(1, 659), R^2=0.0147). The β-estimated = -0.8715 (with P=0.0018<0.05, t-value= -3.13, df=659, SE= -0.1211, Std Error=0.2782; 95% confidence interval -1.4179, -0.3252). Hence, we have strong evidence to reject the null hypothesis *H_0* and conclude that there is a linear association between MMSE and LHCY. The R^2=0.0147 indicates that the variable ‘LHCY’ accounts for 1.47% of the variability of the variable ‘MMSE’ in the dataset. The β-estimated means that, for a one-unit increase in LHCY, on average, MMSE decreases by 0.8715.

To sum up, we have strong evidence to conclude that log homocysteine ‘LHCY’ has a linear relationship with cognitive function ‘MMSE’.

11. A full multiple linear regression model and Regression diagnostics

Considering that the variable ‘EDUCG’ and ‘AGEGRP’ (which we selected) are both categorical variables, a set of dummy variables (0,1) that code for education ‘EDUCG’ (EDUCG_1, EDUCG_2, EDUCG_3, which all using ‘4, if at least some college’ as the reference group) and code for age ‘AGEGRP’ (AGEGRP

_1, AGEGRP_2, AGEGRP_3, which all using ‘4, age of 85-89’ as the reference group) were created.

And the estimate coefficients (slopes) of the dummy variable ‘AGEGRP’ in the model denotes the average difference in the dependent variable (MMSE) with respect to the reference group (AGEGRP_4, or those who age from 85 to 89); the estimate coefficients (slopes) of the dummy variable ‘EDUCG’ in the model denotes the average difference in the dependent variable (MMSE) with respect to the reference group (EDUCG_4, or those who at least have college degree).

11.1 Full multiple linear regression analysis

A multiple linear regression analysis of the outcome of MMSE was performed with LHCY, MALE, PKYRS, dummy variables of EDUCG (EDUCG_1, EDUCG_2, EDUCG_3), and dummy variables of AGEGRP (AGEGRP_1, AGEGRP_2, AGEGRP_3). For this analysis, there are 69 missing values, so the total number of observations used are 594.

Table 10 full multiple linear regression model

	DF	**F**	**P**	**N_obs**	**R^2**	**Adj R^2**	
Model	9	11.50	<0.0001	594	0.1505	0.1375	
Error	584						
	Beta estimate	**95% Lower CI**	**95% upper CI**	**Standard Error**	**Standard Estimate**	**t**	**p**
LHCY	-0.36161	-0.86986	0.14664	0.25878	-0.05473	-1.40	0.1628
MALE	-0.36509	-0.79458	0.06439	0.21867	-0.06697	-1.67	0.0955
EDUCG_1	-3.43315	-4.63073	-2.23556	0.60976	-0.22245	-5.63	<0.0001
EDUCG_2	-1.48856	-1.99723	-0.97990	0.25899	-0.25713	-5.75	<0.0001
EDUCG_3	-0.47503	-0.93927	-0.01079	0.23637	-0.08923	-2.01	0.0449
AGEGRP_1	2.11072	1.18900	3.03244	0.46930	0.41006	4.50	<0.0001
AGEGRP_2	1.63251	0.69590	2.56911	0.47688	0.29335	3.42	0.0007
AGEGRP_3	1.54299	0.52884	2.55713	0.51636	0.20251	2.99	0.0029
PKYRS	0.00327	-0.00628	0.01281	0.00486	0.02720	0.67	0.5017

H_0: There is no linear association between MMSE and LHCY, MALE, dummy variables of EDUCG, dummy variables of AGEGRP, and PKYRS.

H_1: There is a linear association between MMSE and LHCY, MALE, dummy variables of EDUCG, dummy variables of AGEGRP, and PKYRS.

Results: Reject the null hypothesis. We have enough evidence to reject the null hypothesis H_0 (P<0.0001, F-value=11.50, df=(9,584), R^2=0.1505, adj-R^2=0.1375) and conclude that there is a linear association between MMSE and these predictors. The adjusted-R^2=0.1375 for this model indicates nearly 13.75% of the variability in MMSE was accounted for by the combination of LHCY, MALE, dummy variables of EDUCG, dummy variables of AGEGRP and PKYRS after adjusting for the number of model parameters.

Test for 'LHCY': When testing the null hypothesis that there is no linear association between MMSE and LHCY after adjusting for MALE, PKYRS, EDUCG and AGEGRP, we fail to reject H_0 (P=0.1628>0.05, t-value= -1.40, df=584, SE= -0.0547; β-estimated= -0.3616, 95% confidence interval -0.8699, 0.1466). For a one-unit change in MALE, on average, the MMSE decreases by 0.3616, after adjusting for other predictors.

Test for 'MALE': When testing the null hypothesis that there is no linear association between MMSE and MALE after adjusting for LHCY, PKYRS, EDUCG and AGEGRP, we fail to reject H_0 (P=0.0955>0.05, t-value= -1.67, df=584, SE= -0.0670; β-estimated= -0.3651, 95% confidence interval -0.7946, 0.0644). For a one-unit change in MALE, on average, the MMSE decreases by 0.3651, after adjusting for other predictors.

Test for 'PKYRS': When testing the null hypothesis that there is no linear association between MMSE and PKYRS after adjusting for LHCY, MALE, EDUCG and AGEGRP, we fail to reject H_0 (P=0.5017>0.05, t-value= 0.67, df=584, SE= 0.0272; β-estimated= 0.0033, 95% confidence interval -0.0063, 0.0128). For a one-unit change in PKYRS, on average, the MMSE increases by 0.0033, after adjusting for other predictors.

Test for 'EDUCG_1': When testing the null hypothesis that there is no linear association between MMSE and EDUCG_1 after adjusting for LHCY, MALE, PKYRS and AGEGRP, we have strong evidence to reject H_0 (P<0.0001, t-value= -5.63, df=584, SE= -0.2225; β-estimated= -3.4332, 95% confidence interval -4.6307, -2.2356). For a one-unit change in EDUCG_1, on average, the MMSE decreases by 3.4332, after adjusting for other predictors. In other words, those who have education lesser than 8 years (EDUCG_1) will have 343.32% lower cognitive function than those who at least have a college degree (EDUCG_4).

Test for 'EDUCG_2': When testing the null hypothesis that there is no linear association between MMSE and EDUCG_2 after adjusting for LHCY, MALE, PKYRS and AGEGRP, we have strong evidence to reject H_0 (P<0.0001, t-value= -5.75, df=584, SE= -0.2571; β-estimated= -1.4886, 95% confidence interval -1.9972, -0.9799). For a one-unit change in EDUCG_2, on average, the MMSE decreases by 1.4886, after adjusting for other predictors. In other words, those who have education >= 8 years but with no high school degree (EDUCG_2) will have 148.86% lower cognitive function than those who at least have a college degree (EDUCG_4).

Test for 'EDUCG_3': When testing the null hypothesis that there is no linear association between MMSE and EDUCG_3 after adjusting for LHCY, MALE, PKYRS and AGEGRP, we have enough evidence to reject *H0* (P=0.0449<0.05, t-value= -2.01, df=584, SE= -0.0892; β-estimated= -0.4750, 95% confidence interval -0.9393, -0.0108). For a one-unit change in EDUCG_3, on average, the MMSE decreases by 0.4750, after adjusting for other predictors. In other words, those who have high school degree but with no college degree (EDUCG_3) will have 47.50% lower cognitive function than those who at least have a college degree (EDUCG_4).

Test for 'AGEGRP_1': When testing the null hypothesis that there is no linear association between MMSE and AGEGRP_1 after adjusting for LHCY, MALE, PKYRS and EDUCG, we have strong evidence to reject *H0* (P<0.0001, t-value= 4.50, df=584, SE=0.4101; β-estimated= 2.1107, 95% confidence interval 1.1890, 3.0324). For a one-unit change in AGEGRP_1, on average, the MMSE increases by 2.1107, after adjusting for other predictors. In other words, those who age from 65-74 (AGEGRP_1) will have 211.07% higher cognitive function than those who age from 85-89 (AGEGRP_4).

Test for 'AGEGRP_2': When testing the null hypothesis that there is no linear association between MMSE and AGEGRP_2 after adjusting for LHCY, MALE, PKYRS and EDUCG, we have strong evidence to reject *H0* (P=0.0007<0.05, t-value= 3.42, df=584, SE=0.2934; β-estimated = 1.6325, 95% confidence interval 0.6959, 2.5691). For a one-unit change in AGEGRP_2, on average, the MMSE increases by 1.6325, after adjusting for other predictors. In other words, those who age from 75-79 (AGEGRP_2) will have 163.25% higher cognitive function than those who age from 85-89 (AGEGRP_4).

Test for 'AGEGRP_3': When testing the null hypothesis that there is no linear association between MMSE and AGEGRP_3 after adjusting for LHCY, MALE, PKYRS and EDUCG, we have strong evidence to reject *H0* (P=0.0029<0.05, t-value= 2.99, df=584, SE=0.2025; β-estimated = 1.5430, 95% confidence interval 0.5288, 2.5571). For a one-unit change in AGEGRP_3, on average, the MMSE increases by 1.5430, after adjusting for other predictors. In other words, those who age from 80-84 (AGEGRP_3) will have 154.30% higher cognitive function than those who age from 85-89 (AGEGRP_4).

To sum up, we have enough evidence to reject the global null hypothesis and conclude that there is a linear association between MMSE and LHCY, MALE, dummy variables of EDUCG, dummy variables of AGEGRP, and PKYRS. And only the estimated individual slopes for LHCY, MALE, and PKYRS were not statistically significant because they all have p-values larger than 0.05. We could also conclude from the results that people who at least have a college degree would have a statistically significant better cognitive function than those who received lower education; and younger people would also have a statistically significant better cognitive function compared with older people who age from 85-89.

11.2 Regression diagnostics

According to the quantiles table and extreme values table of studentized and press residuals outputted by using PROC UNIVARIATE. The test for normality is significant (P<0.0001) for the studentized residual.

Table 11 Problematic observations

ID	studentized residual	press residual	Cook's Distance
254	-11.47	-27.67	0.235
350	-6.67	-16.27	0.179

To sum up, there are two problematic observations (DEMOGID = 254 and DEMOGID = 350) in data set. The two residual evaluations both suggested that DEMOGID=254 and DEMOGID=350 has a suspiciously lower measurement. The DEMOGID = 254 has a very low studentized residual of -11.47, a very low press residual of -27.67, and an extremely higher Cook's distance of 0.235, comparing with other observations. The DEMOGID = 350 has a relatively low studentized residual of -6.67, a low press residual of -16.27, and a higher Cook's distance of 0.179, comparing with other observations.

Multicollinearity Diagnostics

Table 12 The two-highest VIF and their tolerance

	Tolerance	Variance Inflation
AGEGRP_1	0.17498	5.71491
AGEGRP_2	0.19809	5.04821

The VIF values for AGEGRP_1 and AGEGRP_2 did not exceed 10 (VIF-value for AGEGRP_1 = 5.715, VIF-value for AGEGRP_2 = 5.048; Tolerance for AGEGRP_1 = 0.1750, Tolerance for AGEGRP_2 = 0.1981), which suggest that there is no clear indication of a collinearity problem related to these variables. However, since dummy variables 'AGEGRP_1' and 'AGEGRP_2' have very similar VIF and tolerance values, they may have similar effect in the model. The condition index (CI) for principal component 9 is approximately 4.8838 (<10), and over 90% (about 93.75% for AGEGRP_1 and 91.63% for AGEGRP_2) of the variance of AGEGRP_1 and AGEGRP_2 is explained by it.

Joint Confounding Diagnostics

$$\left|\frac{\hat{\beta}_{adjusted} - \hat{\beta}_{unadjusted}}{\hat{\beta}_{unadjusted}}\right| = \left|\frac{(-0.3616) - (-0.8715)}{-0.8715}\right| \approx 58.51\% > 10\%$$

From outcomes from Q10 and Q11, we could conclude that the results are different to those from the one-factor ANOVA in part II or the linear regression in part III. Without other factors in the model, the estimate slope for LHCY was about -0.8715 (P= 0.0018). Controlling for other factors in the model, the estimate slope for LHCY was about -0.3616 (P= 0.1628). Hence, based on the significant changes (58.51% > 10%) in estimates between unadjusted model and adjusted model, this confirmed there is joint confounding of the LHCY-to-MMSE relationship by other factors (MALE, EDUCG, AGEGRP, and PKYRS).

12. Model selection using LASSO with an AIC-based selection criterion

(1) use PROC GLMSELECT procedure with a CLASS statement (default reference groups)

In model selection procedure, we choose to use PROC GLMSELECT procedure with a CLASS statement for dummy variables MALE, EDUCG, and AGEGRP. The procedure chose EDUCG_3 (those who have a high school degree but no college) and AGEGRP_2 (age from 75 to 79) as the default reference groups.

After the selection procedures, the variable 'PKYRS' and one level 'AGEGRP_3' of the dummy variable AGEGRP were removed from the final model. In other words, the selected model only contains predictors 'LHCY', 'MALE', all levels ('EDUCG_1', 'EDUCG_2' and 'EDUCG_4') of dummy variable EDUCG, and two levels ('AGEGRP_1' and 'AGEGRP_4') of dummy variable AGEGRP. The final model generates a little higher adjusted R^2 (0.1379 > 0.1375) than the previous full multiple linear regression model in Q11, and the smallest AIC=1629.45908. The adjust R^2=0.1379 for the selected model indicates nearly 13.79% of the variability in MMSE was accounted for by the combination of selected factors. And the estimated slopes for all factors are shown in the follow table 13.

Table 13 The estimated slopes for the selected factors in final model

SELECTED FACTORS	LHCY	MALE_0	EDUCG_1	EDUCG_2	EDUCG_4	AGEGRP_1	AGEGRP_4
ESTIMATED SLOPES	**-0.2681**	**0.1895**	**-2.6326**	**-0.9467**	**0.4440**	**0.4384**	**-1.4738**

(2) use PROC GLMSELECT procedure with dummy variables created in Q11 (no CLASS statement)

In model selection procedure, we choose to use PROC GLMSELECT procedure dummy variables created in Q11: LHCY, MALE, PKYRS, dummy variables of EDUCG (EDUCG_1, EDUCG_2, EDUCG_3), and dummy variables of AGEGRP (AGEGRP_1, AGEGRP_2, AGEGRP_3). The EDUCG_4 (those who at least have a college degree) and AGEGRP_4 (age from 80 to 84) as the default reference groups. After the selection procedures, no variables were removed from the final model. And the final model are totally identical with the model we created in Q11.

So, the reason for the difference between the two model selection is that the reference groups for EDUCG and AGEGRP are different in this two procedure.

Conclusion

In this study, our main interest is to evaluate the relationship between cognitive function ‘MMSE’, which as the dependent variable, with log homocysteine ‘LHCY’, which as the main independent variable. First, we compare the distribution between those excluded (whose ADIN7YRS variable is missing) to those not excluded (whose ADIN7YRS variable is non-missing) and then remove the subjects which have missing values for AD in 7 years of follow-up. Then, we recognized the relative charts and other descriptive statistics for the main independent variables remained in our dataset. Later, we examine the relationship between both LHCY and MMSE with continuous age ‘AGE’ and concluded from the results that log homocysteine and cognitive function are linearly associated with continuous age, respectively. Then, we test the mean LHCY difference between four age groups and conclude that there exist at least one age group differs from the others with respect to mean LHCY. To be specify, the mean log homocysteine for age group 3 (age 80-84) and age group 4 (age 85-89) are statistically identical; whereas the mean log homocysteine for age group 1 (age 65-74) and age group 4 (age 85-89) are statistically different, and the mean log homocysteine for age group 2 (age 75-79) and age group 4 (age 85-89) are also statistically different. The result also supports a linear relationship between age groups and log homocysteine. Later, the piecewise linear model was performed, and we concluded that though the overall model is statistically significant at 0.05 confidence level, the estimated individual slopes for age1, age2, age3, and age4 were not statistically significant. According to the previous outcome, I chose the categorical variable ‘AGEGRP’ (in Q7, the mean LHCY differences analysis by the dummy regression using PROC GLM with a class statement) for further analysis later, because it has higher R^2 (0.0293 with P=0.0002, which is larger than 0.0264 with P<0.0001 for simple linear regression and larger than 0.0290 with P=0.0007 for piecewise regression) compared with other two regression models.

A multiple linear regression with interaction with a dummy variable for gender to assess the relationship of LHCY to MMSE (dependent), and the results suggested that there is no significant evidence of effect modification by gender on this relationship. Then a linear regression analysis of the association of MMSE and LHCY was performed, and we have strong evidence to conclude that log homocysteine ‘LHCY’ has a linear relationship with cognitive function ‘MMSE’. A full multiple linear regression and regression diagnostics were performed later, and we have enough evidence to conclude there is a linear association between MMSE and LHCY, MALE, dummy variables of EDUCG, dummy variables of AGEGRP, and PKYRS. And people who at least have a college degree would have a statistically significant better cognitive function than those who received lower education; and younger people would also have a statistically significant better cognitive function compared with older people who age from 85-89. Then, according to the regression diagnostics, there are two problematic observations (DEMOGID = 254 and DEMOGID = 350) in this dataset. And according to multicollinearity analysis, there seems no clear indication of a collinearity problem related to these variables because even the two highest VIF-values are lower than 10. And based on the significant changes (58.51% > 10%) in estimates between unadjusted model and adjusted model, there is joint confounding of the LHCY-to-MMSE relationship by other factors

(MALE, EDUCG, AGEGRP, and PKYRS).

Last, in the model selection, the procedure chose EDUCG_3 (those who have a high school degree but no college) and AGEGRP_2 (age from 75 to 79) as the default reference groups, which is different reference groups compared with our previous performed models. Then, the variable ‘PKYRS’ and one level ‘AGEGRP_3’ of the dummy variable AGEGRP were removed from the final model. But if we choose to use PROC GLMSELECT procedure dummy variables created in Q11, we will find that no variables were removed from the final model. And the final model is totally identical with the model we created in Q11. The reason for the difference between the two model selection is that the reference groups for EDUCG and AGEGRP are different in this two procedure.