Mi Le
BS 805
Project
Dr. Sarah Leatherman

Introduction

Investigators are interested in exploring the relationship between elevated plasma homocysteine (hcy) levels and poor cognition. They used data from 900 participants from the Framingham Heart Study to investigate this association and its implications for the development of Alzheimer's disease. Investigators collected a number of demographic, lab, and neurological data, as well as participant ID numbers. Demographic data included age, sex, education level, and cigarette smoking. Lab data included measures of hcy, plasma folate, plasma vitamin B12, and plasma vitamin B6. Neurological data included score on the Mini-Mental State Examination (MMSE) and whether the participant developed Alzheimer's at 7 years follow-up. There were a number of participants with missing values for the variable measuring whether there was Alzheimer's at follow-up, therefore, investigators wished to exclude these participants from the dataset. However, before removing them, we will conduct a number of statistical tests on a few variables to see how similar the group being excluded and the group not being excluded were to see if their removal will change the make up of our final dataset. In this new dataset of 663 participants, we will generate a number of descriptive statistics on our primary variables of interest. We will then assess the association between MMSE and age using three different models: simple linear regression, dummy variable regression, and piecewise linear regression. We will also examine the relationship between log-transformed hcy and MMSE using a simple linear regression model and we will also investigate whether sex acts as an effect modifier in this relationship. Lastly, we will attempt to fit a final model predicting MMSE from a pool of predictors we think would be important to test and include: log-transformed hcy, sex, education, age, and cigarette smoking. We will then compare our final model with a SAS-generated model using model selection techniques.

We are provided with three separate permanent SAS datasets: demog_ bs805_f22 (900 observations, 5 variables), labs_ bs805_f22 (900 observations, 5 variables), and neuro_ bs805_f22 (900 observations, 3 variables). We will horizontally merge these three datasets into one large temporary SAS dataset called project_temp (900 observations, 13 variables). Using this temporary project_temp dataset, we will copy its contents into a permanent SAS dataset called Project.fulldataset using a set statement and create a number of new variables. First, we will create a new variable called LHCY, which is the natural log of our pre-existing variable called HCY. Next, we will turn our continuous HCY variable into a dichotomous variable called HCYGE14, where 1 is for those with HCY values of at least 14 and 0 is when HCY is less than 14. We will also turn our continuous age variable into a categorical variable called AGEGRP where AGEGRP=0 represent those between 65-74, 1 represents those between 75-79, 2 represents those 80-84, and 3 represents those 85-89. Next, we will take our categorical variable for education, educg, and transform it into a binary dichotomous variable called HSDEG where 0 represents those with less than a high school degree (if education is less than 8 years, or greater than or equal to 8 years but no high school degree) and 1 represents those with a high school degree or higher (if high school degree but no college, or at least some college). There are a number of missing values in the ADIN7YRS variable, therefore we will create a new dichotomous variable called exclude where 1 represents those who are missing a ADIN7YRS value and 0 represents those who are not missing a ADIN7YRS value. Lastly, we will create a new set of dichotomous variables called MMSEF that flags cognitive deficits according to participants' MMSE and education level. For participants with less than 8 years of education, MMSEF will equal 1 if their MMSE is 22 or less and will equal 0 if MMSE is greater than 22. For participants with greater than or equal to 8 years of education but no high school degree, MMSEF will equal 1 if their MMSE is 24 or less and will equal 0 if MMSE is greater than 24. For participants with a high school degree but no college, MMSEF will equal 1 if their MMSE is 25 or less and will equal 0 if MMSE is greater than 25. For participants with at least some college, MMSEF will equal 1 if their MMSE is 26 or less and will equal 0 if MMSE is greater than 26. With these new additional variables our project_temp dataset now contains 900 observations and 19 variables.

Using this new dataset we just created, we will run a number of statistical tests to compare those who would be excluded based on the criteria established previously (missing ADIN7YRS values) and those who would not be excluded. First, we have want to compare the ages of those who would be excluded from the dataset and those who would not. Age is our continuous outcome variable, while exclusion is dichotomous, therefore, we will run a two sample t-test of means at the 0.05 significance level. Our null hypothesis is that the mean age between those who would be excluded from our dataset and those who are not excluded are the same. Our alternative hypothesis is that the mean age between those who would be excluded from the dataset and those who are not excluded are not the same. Before heading into the analysis, we first have to check whether the assumption of equal variance is met. Our null hypothesis is that the variance between participants who were exclude and participants who are not excluded are the same. Our alternative hypothesis is that the variance between participants who were exclude and participants who are not excluded are not the same. Based on the boxplot and the Equality of Variance test, we reject the null at a 0.05 significance level. There is significant evidence at the 0.05 significance level (n=900, F=1.29, df=236, 662, p-value=0.0162) to suggest that the variances are not the same, therefore we can assume unequal variance between participants who are excluded and participants who were not and refer to the Satterthwaite standard error and statistics.

For our t-test, we reject our null hypothesis at the 0.05 significance level, there is significant evidence (n=900, t=-4.36, df=375.13, p-value<0.0001) to suggest that the mean age among participants who are excluded is different from the mean age among participants who are not excluded. The mean age of those who are not excluded was 75.3152 (n=663, std=4.6064), while the mean age of those who are excluded was 76.9873 (n=237, std=5.2234). The difference in mean age was 1.6721 (se=0.3836), meaning that on average those who were not excluded had a mean age 1.6721 years lower than those who were excluded with a 95% confidence interval of (-2.4263, -0.9179). If we were to repeat this study many times with the same sample sizes, we would expect 95% of resulting confidence intervals to contain the true difference in mean age between those who are not excluded and those who are excluded. We hope that our confidence interval of (-2.4263, -0.9179) is a part of that 95% of confidence intervals that contain the true difference in mean age between those who are not excluded and those who are excluded.

Next, we want to compare the probability of being male in participants who were excluded compared to those who were not excluded. Both the exclusion and male variable were dichotomous variables, therefore, we will perform a chi-square test at a 0.05 significance level. Our null hypothesis is that the odds of being male in the group of participants who were excluded is the same as in the group of participants who were not excluded. Our alternative hypothesis is that the odds of being male in the group of participants who were excluded is not the same as in the group of participants who were not excluded. We reject the null hypothesis at the 0.05 significance level, there is significant evidence (n=900, chi-sq statistic=14.5864, df=1, p-value=0.0001) to suggest that the odds of being male in the group of participants who were excluded is not the same as in the group of participants who were not excluded. The proportion of participants who were male among those who were excluded was 0.4810 (114/237), while the proportion of participants who were male among those who were not excluded was 0.3409 (226/663). The odds ratio was 1.7921, meaning that those who were excluded had 1.7921 times the odds of being male compared to those who were not excluded. The odds ratio had a 95% confidence interval of (1.3261, 2.4220). If we were to repeat this study many times with the same sample sizes, we would expect 95% of resulting confidence intervals to contain the true odds ratio of being male between those who are excluded and those who are not excluded. We hope that our confidence interval of (1.3261, 2.4220) is a part of that 95% of confidence intervals that contain the true odds ratio of being male between those who are excluded and those who are not excluded.

Next, we want to compare the probability of having a high school degree or higher in participants who were excluded compared to those who were not excluded. Both the exclusion and high school degree variable were dichotomous variables, therefore, we will perform a chi-square test at a 0.05 significance level. Our null hypothesis is that the odds of having a high school degree or higher among participants who were excluded is the same as in the group of participants who were not excluded. Our alternative hypothesis is that the odds of having a high school degree or higher among participants who were excluded is not the same as in the group of participants who were not excluded. We fail to reject the null hypothesis at the 0.05 significance level, there is not significant evidence (n=885, chi-sq statistic=1.7046, df=1, p-value=0.1917) to suggest that the odds of having a high school degree or higher in the group of participants who were excluded is not the same as in the group of participants who were not excluded. The proportion of participants who had a high school degree or higher among those who were excluded was 0.6496 (152/234), while the proportion of participants who had a high school degree or higher among those who were not excluded was 0.6959 (453/651), The odds ratio was 0.8102, meaning that those who were excluded had 0.8102 times the odds of having a high school degree or higher compared to those who were not excluded. The odds ratio had a 95% confidence interval of (0.5906, 1.1116). If we were to repeat this study many times with the same sample sizes, we would expect 95% of resulting confidence intervals to contain the true odds ratio of having a high school degree or higher between those who are excluded and those who are not excluded. We hope that our confidence interval of (0.5906, 1.1116) is a part of that 95% of confidence intervals that contain the true odds ratio of having a high school degree or higher between those who are excluded and those who are not excluded.

Next, we want to compare the mean pack-years of cigarette smoking among those who would be excluded from the dataset and those who would not. Pack-years of cigarette smoking is our continuous outcome variable, while exclusion is dichotomous, therefore, we will run a two sample t-test of means at the 0.05 significance level. Our null hypothesis is that the mean pack-years of cigarette smoking between those who would be excluded from our dataset and those who are not excluded are the same. Our alternative hypothesis is that the mean pack-years of cigarette smoking between those who would be excluded from the dataset and those who are not excluded are not the same. Before heading into the analysis, we first have to check whether the assumption of equal variance is met. Our null hypothesis is that the variance between participants who were exclude and participants who are not excluded are the same. Our alternative hypothesis is that the variance between participants who were exclude and participants who are not excluded are not the same. Based on the boxplot and the Equality of Variance test, we fail to reject the null. There is not significant evidence at the 0.05 significance level (n=811, F=1.21, df=214, 595, p-value=0.0801) to suggest that the variances are not the same, therefore we can assume equal variance between those who are excluded and those who are not are roughly equal and refer to the pooled standard error and statistics. For our t-test, we fail to reject our null hypothesis at the 0.05 significance level, there is not significant evidence (n=811, t=-1.82, df=809, p-value=0.0687) to suggest that the mean pack-years of cigarette smoking among participants who are excluded is different from the mean pack-years of cigarette smoking among participants who are not excluded. The mean pack-years of cigarette smoking of those who are not excluded was 15.4109 (n=596, std=21.4253), while the mean pack-years of cigarette smoking of those who are excluded was 18.6036 (n=215, std=23.5911). The difference in mean pack-years of cigarette smoking was -3.1927 (std=22.0190), meaning that on average those who were not excluded had a mean 3.1927 pack-years of cigarette smoking less than those who were excluded with a 95% confidence interval of (-6.6312, 0.2457). If we were to repeat this study many times with the same sample sizes, we would expect 95% of resulting confidence intervals to contain the true difference in mean pack-years of cigarette smoking between those who are excluded and those who are not excluded. We hope that our confidence interval of (-6.6312, 0.2457) is a part of that 95% of confidence intervals that contain the true difference in mean pack-years of cigarette smoking between those who are excluded and those who are not excluded.

Next, we want to compare the mean cognitive status measured using the mini-mental state examination score (mmse) among those who would be excluded from the dataset and those who would not. Cognitive status is our continuous outcome variable, while exclusion is dichotomous, therefore, we will run a two sample t-test of means at the 0.05 significance level. Our null hypothesis is that cognitive status between those who would be excluded from our dataset and those who are not excluded are the same. Our alternative hypothesis is that cognitive status between those who would be excluded from the dataset and those who are not excluded are not the same. Before heading into the analysis, we first have to check whether the assumption of equal variance is met. Our null hypothesis is that the variance between participants who were exclude and participants who are not excluded are the same. Our alternative hypothesis is that the variance between participants who were exclude and participants who are not excluded are not the same. Based on the boxplot and the Equality of Variance test, we reject the null hypothesis; there is significant evidence at the 0.05 significance level (n=895, F=2.46, df=233, 660, p-value<0.0001) to suggest that the variance between those who are excluded and those who are not excluded are not the same, therefore will assume unequal variance and will refer to the Satterthwaite standard error and statistics. For our t-test, we reject our null hypothesis at the 0.05 significance level, there is significant evidence (n=895, t=3.86, df=302.8, p-value=0.0001) to suggest that the mean cognitive status among participants who are excluded is different from the mean cognitive status among participants who are not excluded. The mean cognitive status of those who are not excluded was 28.0953 (n=661, std=2.7538), while the mean cognitive status of those who are excluded was 26.9316 (n=234, std=4.3153). The difference in mean cognitive status was 1.1637 (se=0.3018) meaning that on average those who were not excluded had a mean 1.1637 cognitive status unit higher than those who were excluded with a 95% confidence interval of (0.5699, 1.7575). If we were to repeat this study many times with the same sample sizes, we would expect 95% of resulting confidence intervals to contain the true difference in mean cognitive status between those who are excluded and those who are not excluded. We hope that our confidence interval of (0.5699, 1.7575) is a part of that 95% of confidence intervals that contain the true difference in mean cognitive status between those who are excluded and those who are not excluded.

Next, we have want to compare the mean natural log of plasma homocysteine (hcy) among those who would be excluded from the dataset and those who would not. Log-transformed hcy is our continuous outcome variable, while exclusion is dichotomous, therefore, we will run a two sample t-test of means at the 0.05 significance level. Our null hypothesis is that the mean log-transformed hcy between those who would be excluded from our dataset and those who are not excluded are the same. Our alternative hypothesis is that mean log-transformed hcy between those who would be excluded from the dataset and those who are not excluded are not the same. Before heading into the analysis, we first have to check whether the assumption of equal variance is met. Our null hypothesis is that the variance between participants who were exclude and participants who are not excluded are the same. Our alternative hypothesis is that the variance between participants who were exclude and participants who are not excluded are not the same. Based on the boxplot and the Equality of Variance test, we fail to reject the null hypothesis; there is not significant evidence at the 0.05 significance level (n=900, F=1.10, df=236, 662, p-value=0.3822) to suggest that the variance between those who are excluded and those who are not excluded are not the same, therefore will assume equal variance and will refer to the pooled standard error and statistics. For our t-test, we reject our null hypothesis at the 0.05 significance level, there is significant evidence (n=900, t=-2.92, df=898, p-value=0.0036) to suggest that the mean log-transformed hcy among participants who are excluded is different from the mean log-transformed hcy among participants who are not excluded. The mean log-transformed hcy of those who are not excluded was 2.4272 (n=663, std=0.3824), while the mean log-transformed hcy of those who are excluded was 2.5126 (n=237, std=0.4003). The difference in log-transformed hcy was -0.0854 (std=0.3872) meaning that on average those who were not excluded had a mean log-transformed hcy 0.0854 units less than those who were excluded with a 95% confidence interval of (-0.1430, -0.0279). If we were to repeat this study many times with the same sample sizes, we would expect 95% of resulting confidence intervals to contain the true difference in mean log-transformed hcy between those who are excluded and those who are not excluded. We hope that our confidence interval of (-0.1430, -0.0279) is a part of that 95% of confidence intervals that contain the true difference in mean log-transformed hcy between those who are excluded and those who are not excluded.

Next, we would like to examine plasma homocysteine (hcy) levels again using a slightly different measure. Previously we examined hcy as a log-transformed continuous variable, this time we transformed the hcy variable into a dichotomous variable where 1 equals hcy of 14 µmol/L or more, whereas 0 equals hcy values less than 14 µmol/L. We are interested in examining the relationship between this measure of hcy and exclusion group. More specifically, we are interested in comparing the probability of having a hcy of 14 µmol/L or more between participants who were excluded and those who were not excluded. We will conduct a chi-square test at a 0.05 significance level. Our null hypothesis is that the odds of having an hcy of 14 µmol/L or more among participants who were excluded is the same as in the group of participants who were not excluded. Our alternative hypothesis is that the odds of having a hcy of 14 µmol/L or more among participants who were excluded is not the same as in the group of participants who were not excluded. We reject the null hypothesis at the 0.05 significance level, there is significant evidence (n=900, chi-sq statistic=5.4773, df=1, p-value=0.0193) to suggest that the odds of having an hcy of 14 µmol/L or more in the group of participants who were excluded is not the same as in the group of participants who were not excluded. The proportion of participants who had an hcy of 14 µmol/L or more among those who were excluded was 0.3291 (78/237), while the proportion of participants who had an hcy of 14 µmol/L or more among those who were not excluded was 0.2504 (166/663), The odds ratio was 1.4687, meaning that those who were excluded had 1.4687 times the odds of having an hcy of 14 µmol/L or more compared to those who were not excluded. The odds ratio had a 95% confidence interval of (1.0634, 2.0286). If we were to repeat this study many times with the same sample sizes, we would expect 95% of resulting confidence intervals to contain the true odds ratio of having an hcy of 14 µmol/L or more between those who are excluded and those who are not excluded. We hope that our confidence interval of (1.0634, 2.0286) is a part of that 95% of confidence intervals that contain the true odds ratio of having an hcy of 14 µmol/L or more between those who are excluded and those who are not excluded.

In summary, when comparing those excluded to those not excluded on the basis of missing values for the adin7rs variable at a significance level of 0.05, we found that there were significant differences in mean age, MMSE score, and log-transformed hcy. There was not a significant difference in mean pack years of cigarette smoking between those excluded and those not excluded. Please refer to Table 1 for the results of these 2-sample t-test of means. Additionally, we found that there was a significant difference in the probability of being male and having a hcy greater than or equal to 14 µmol/L when comparing participants who were excluded and those who were not. We did not find a significant differnec ein the probability of having a high school degree or higher between the group that was excluded and the group that was not excluded. Please refer to Table 2 for the results of these chi-square tests.

Table 1. 2-Sample t-test of mean results comparing exclusion to non-exclusion group

Variable	**N**	**Non-exclusion group mean (sd)**	**Exclusion group mean (sd)**	**Difference in mean (sd/se)**	**t-statistic**	**p-value**
Age	900	75.3152 (4.6064)	76.9873 (5.2234)	-1.6721 (0.3836)	-4.36, df=375.13	< 0.0001
Cigarette smoking	811	15.4109 (21.4253)	18.6036 (23.5911)	-3.1927 (22.0190)	-1.82, df=809	0.0687
MMSE	895	28.0953 (2.7538)	26.9316 (4.1353)	1.1637 (0.3018)	3.86, df=302.8	0.0001
Log HCY	900	2.4272 (0.3824)	2.5126 (0.4003)	-0.0854 (0.3872)	-2.92, df=898	0.0036

Table 2. Chi-square test results comparing exclusion to non-exclusion group

Variable	**N**	**Non-exclusion group proportion**	**Exclusion group proportion**	**Odds Ratio (95% CI)**	**Chi-Square**	**p-value**
Being male	900	0.3409 (226/663)	0.4810 (114/237)	1.7921 (1.3261, 2.4220)	14.5864, df=1	0.0001
High school degree or higher	885	0.6959 (453/651)	0.6496 (152/234)	0.8102 (0.5906, 1.1116)	1.7046, df=1	0.1917
HCY ≥ 14	900	0.2504 (166/663)	0.3291 (78/237)	1.4687 (1.0634, 2.0286)	5.4773, df=1	0.0193

We will now create a new temporary SAS dataset called temp that is a copy of project.fulldataset (obs=900, 19 variables), however, we will be removing observations where there are missing values for the adin7yrs variable (ie. where the exclude variable is equal to 1). In this temp dataset we will also drop the exclude variable from the dataset. As a result, our final temp dataset has a total of 663 observations and 18 variables. We will be using this temp dataset for the remainder of our analyses.

We will now check the distribution of the following variables: pack years of cigarette smoking, untransformed hcy and log-transformed hcy, plasma folate, and MMSE. The distribution of pack years of cigarette smoking (N=597, missing=66) is heavily skewed right with roughly half of participants 1 or less pack years of cigarette smoking. The median of this distribution was 1 pack year, while the mean was 15.8608 pack years (sd=21.97598), which further supports our notion that the distribution is skewed. Furthermore, the first quartile (Q1) was 0 pack-years, the second (median) was 1, whereas the third quartile (Q3) jumps to 28.3864 again further supporting the fact that this distribution is skewed with a majority of participants concentrated around 0-1 pack year, with a number of participants with much higher pack-years, with the maximum number of pack-years being 122.5596. The mode was 0 pack-years, which indicates that most participants had 0 pack-years. Since this distribution is skewed, median and interquartile range (IQR) would best be used to describe the central tendency and variance of pack-years of cigarette smoking with a median of 1 pack-year and an IQR of 28.3864 (the same value as Q3 since Q1 was equal to 0).

The untransformed hcy variable is much more normally distributed (N=663) than pack-years of cigarette smoking. The mean (12.5450; sd=6.2488) and median (11.2000; IQR=5.3000) are fairly close to one another in value, which further supports the notion that untransformed hcy is normally distributed therefore we will use mean and std as our measures of central tendency and variation. The minimum value of untransformed hcy was 3.5 µmol/L, while the maximum value was 66.9 µmol/L. 75% of participants had an hcy measure between 3.5 and 14.3 µmol/L. We have a very high maximum that is much further away from the rest of the data (100th percentile was 66.9, while the 99th was 33.7). This indicates that we have few very high values in the distribution of untransformed hcy, that has dragged our mean slightly to the right, however, since our mean was not dragged very far and is still in close proximity to our median, our distribution remains normal and we must simply just be aware of the few high hcy values.

The distribution of log-transformed hcy (N=663) looks visually similar to the distribution of untransformed hcy, ie. their shapes are very similar, if not identical. The mean of log-transformed hcy was 2.4442 (sd=0.3942). while the median was also very close in value at 2.4159 (IQR=0.4630). Since hcy was log-transformed the scale of the distribution is much smaller, however, the shape of the distribution remain the same. The minimum was 1.2528 and the maximum was 4.2032. Similar to untransformed hcy, we see that a majority of participants are concentrated between a specific range; around 90% of the participants had log-transformed hcy values under 3 and the highest 10% of the sample ranged between 3 and 4.2032, while the bottom 90% ranged between a similarly sized range between 1.2528 and 2.9339. Similar to the untransformed hcy distribution, log-transformed hcy also has a few high values relative to the rest of the distribution.

Plasma folate is also relatively normal (N=658, missing=5), if not slightly skewed right. If we compare the mean to the median, the mean is 6.7327 (sd=6.6408), while the median is 4.4550 (IQR=5.6800) we see that the mean is slightly larger than the median, which may indicate that the distribution is slightly skewed to the right. If we look at the quartile distribution, there seems to be a relatively even normal spread up until the median (4.4550) or Q3 (8.3000). The highest 5% (20.1000-55.8600) has much larger plasma folate levels than the bottom 95% (0.8200-20.1000). Since this distribution is slightly skewed right, we may be best off using median and IQR as our measure of central tendency and variance.

MMSE (N=661, missing=2), on the other hand, is skewed left. Since the distribution is obviously skewed left, we will use median and IQR as our measures of central tendency and variation. The median of the distribution of MMSE is 29.0000, while the IQR is 3.0000. The scale of MMSE scores ranges between 0-30 and in the distribution of MMSE in this sample, we see that many participants score the highest possible score of 30. We see that the median is 29, which means that the upper 50% scored at least a 29 on the MMSE scale, which is where most of our participants are centered. The lower 50% ranges anywhere between 0 and 29.

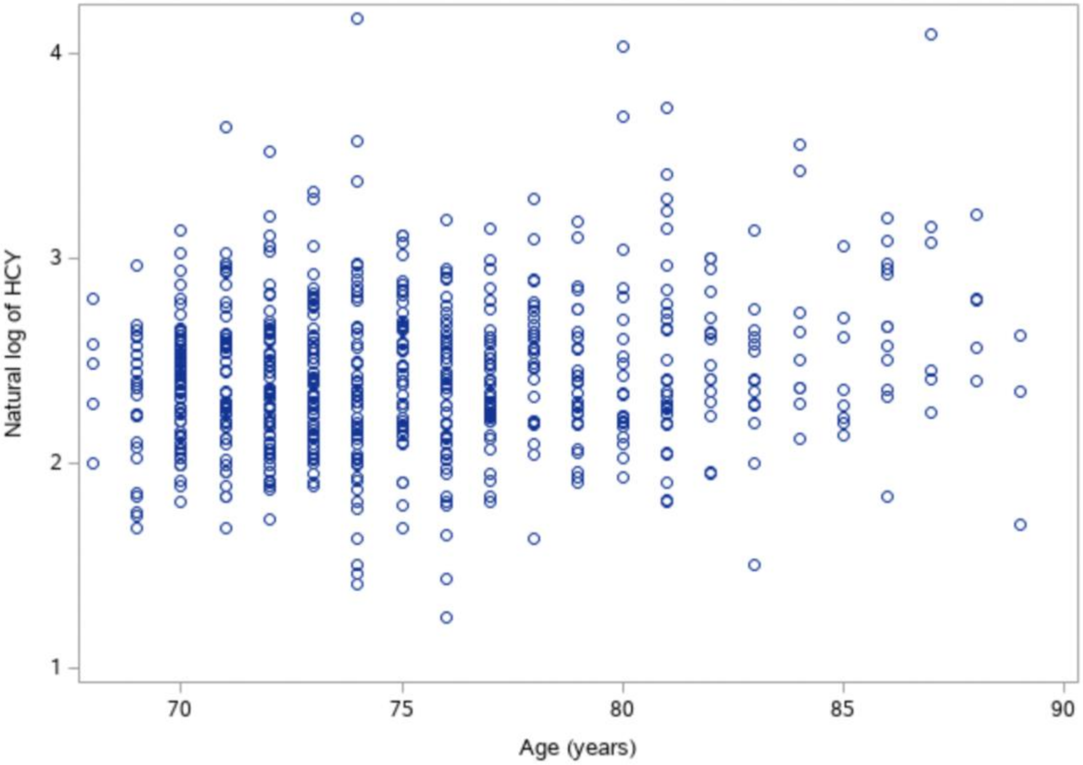
In summary, all our variables had some skew to them with cigarette smoking having the most drastic skew to the right and is very much not normally distributed. MMSE score also had a fairly obvious skew to the left and was not normally distributed. Both these variables being heavily skewed may be due to the fact of its measurement where cognitive function is heavily concentrated at higher levels with most participants scoring either 29 or 30 on MMSE, while a heavy majority of participants had 0 pack-years of cigarette smoking while those who have a history of cigarette smoking have very high pack years, although those participants are still in the minority. The two measure of hcy were a little more normally distributed than cigarette smoking or MMSE, although it still exhibited a slight skew to the right. Similarly, plasma folate was also slightly more normally distributed but also skewed slightly right. Please see Table 3 below for descriptive statistics of these variables where we see for a majority of these variables the mean is larger than the median, indicating a skew to the right, except for MMSE which exhibits a skew to the left and we see that its mean is smaller than its median.

Table 3. Descriptive statistics for the following variables

Variable	N	Mean (sd)	Median (IQR)	Q1	Q3	Minimum	Maximum
Cigarette smoking (pack-years)	596	15.4109 (21.4253)	0.6092 (28.1979)	0	28.1979	0	122.5596
HCY (µmol/L)	663	12.2781 (5.9266)	11.000 (5.1000)	8.9000	14.000	3.5000	64.6000
Log HCY	663	2.4272 (0.3824)	2.3979 (0.4530)	2.1861	2.6391	1.2428	4.1682
Plasma folate (nmol/L)	652	6.2908 (6.0705)	4.2050 (5.3900)	2.5300	7.9200	0	53.240
MMSE	661	28.0953 (2.7538)	29.0000 (3.0000)	27.0000	30.0000	0	30.0000

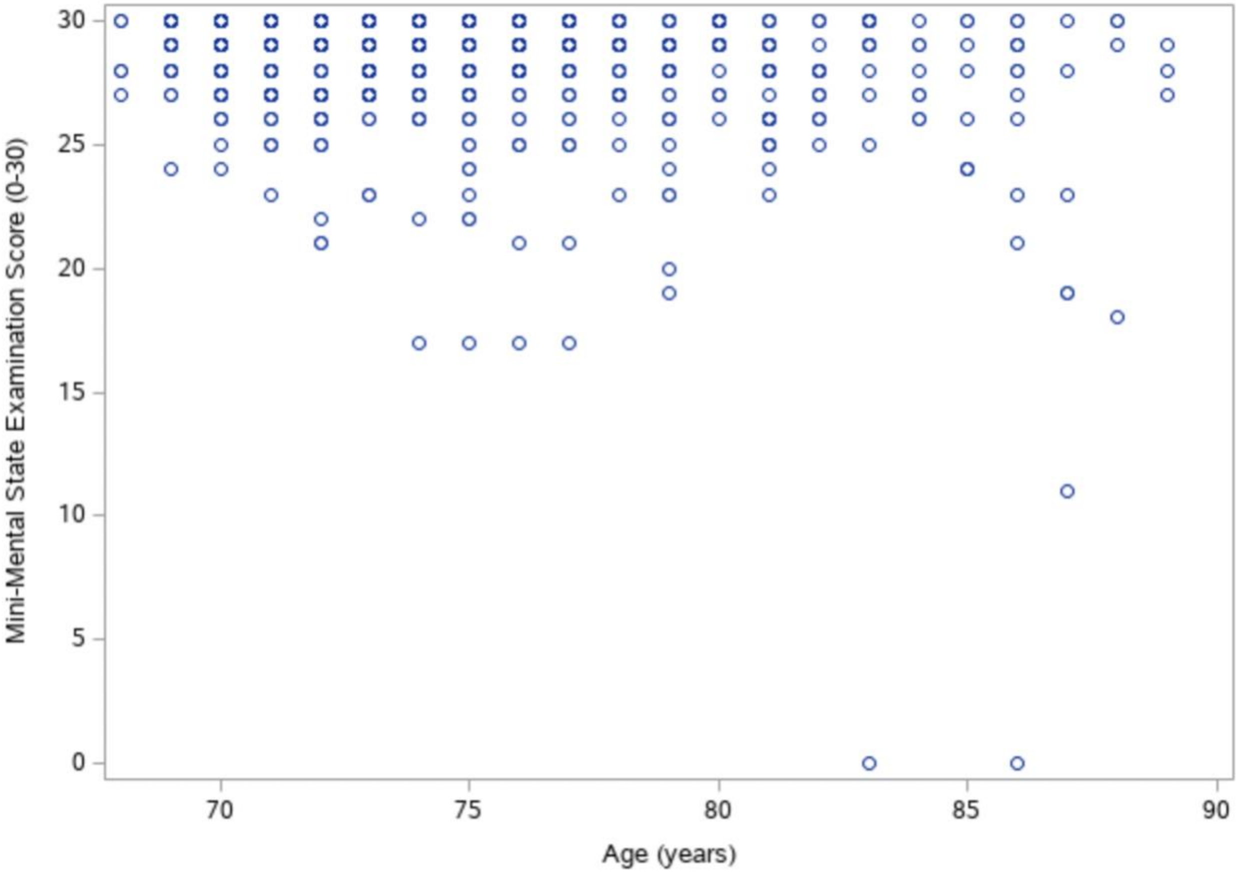
Investigators posit that hcy and cognitive function (MMSE) differs with age. We, therefore, are interested in whether there is a linear association between hcy and age and between MMSE and age, with age as our independent variable. For this analysis we will be using the log-transformed hcy variable. We will assume that the observations in this sample are independent from one another since there is no reason for us to assume otherwise. Before heading into our analysis, we want to also visually assess whether the association between log-transformed hcy and age look linear. If we refer to Figure 1, we can see that the association looks somewhat linear; there is no other discernable shape for this association other than linear. There seems to be a bit more scatter at higher values of age compared to lower values but this difference does not seem drastic, therefore we will assume there is relatively constant variance. We know from previously that log-transformed hcy is not particularly normal and is skewed right. The distribution for age is also similar in that it has a slight skew to the right, therefore normality may be an issue, but we will proceed with our simple linear regression with caution at a significance level of 0.05. Our null hypothesis is that there is no linear association between log-transformed hcy and age. Our alternative hypothesis is that there is a linear association between log-transformed hcy and age. We reject the null hypothesis at the 0.05 significance level; there is significant evidence (N=663, F=17.93, df=1,661, p-value<0.0001) to suggest that there is a linear association between log-transformed hcy and age. The R^2 for this model was 0.0264, meaning that 2.64% of the variability in log-transformed hcy can be explained by the variability in age. The regression coefficient for age was 0.0135 (0.0032), meaning that for every 1 year increase in age, on average we would expect a 0.0135 unit increase in log-transformed hcy. Looking at the diagnostic plots output by the reg procedure, particularly the normal quantile/QQ plot, we see that there is not an incredibly dramatic departure from normality. The predicted value vs. residual plot also confirms our suspicion that there may not be constant variance as we once again see that there is more scatter at higher predicted values compared to lower predicted values.

Figure 1. Scatterplot of age and log hcy



Next, we will assess whether MMSE is linearly associated with age. Again, we will assume that the observations in this sample are independent. The bar chart of the distribution of MMSE from before showed that the distribution of MMSE was skewed left, therefore it is not normally distributed, however, age is still roughly normal albeit slightly skewed righty. We will use a scatterplot to visually assess the association between MMSE and age, please refer to Figure 2, and we see that this relationship does not look particularly linear. There is also more variance at higher values of age than lower values of age, which will bring into question the constant variance assumption. We will proceed with our simple linear regression predicting MMSE from age with caution at a significance level of 0.05. Our null hypothesis is that there is not a linear association between MMSE and age. Our alternative hypothesis is that there is a linear association between MMSE and age. We reject our null hypothesis at the 0.05 significance level, there is significant evidence (N=661, F=38.02, df=1,659, p-value<0.0001) to suggest that there is a linear association between MMSE and age. The R^2 for this model is 0.0546, meaning that 5.46% of the variability in MMSE can be explained by the variability in age. The regression coefficient for age in the model predicting MMSE is -0.1406 (se=0.0228), meaning that for every 1 year increase in age, we can expect on average a decrease in MMSE of 0.1406 units. Looking at the diagnostic plots output by the reg procedure, particularly the normal quantile/QQ plot, we see that there is a noticeable departure from normality. The predicted value vs. residual plot also confirms our suspicion that there may not be constant variance as see that the variance changes throughout the plot with more variance at smaller predicted values than at higher predicted values. There are also seems to be more negative than positive residuals and this becomes more the case as we increase in predicted values, which gives further evidence that this association is not particularly linear.

Figure 2. Scatterplot of age and MMSE



In summary, age was found to be linearly associated with both log hcy and MMSE score, please refer to Tables 4 and 5 for the results of these simple linear regression models. Although both relationships were found to be significantly associated, we found that a number of assumptions may not have been met in order to properly run a simple linear regression. Specifically, in the relationship between age and log-transformed hcy we found that the constant variance assumption may not have been satisfied, while the normality assumption was met and the linearity assumption was dubiously met, for the most part. In the relationship between age and MMSE, none of these three assumptions were particularly satisfied. As a result, although our hypothesis test resulted in significant associations, we should be wary of the potential for bias as a simple linear regression model may not be the best model for these two relationships, particularly that between age and MMSE.

Table 4. Simple linear regression predicting log hcy with age

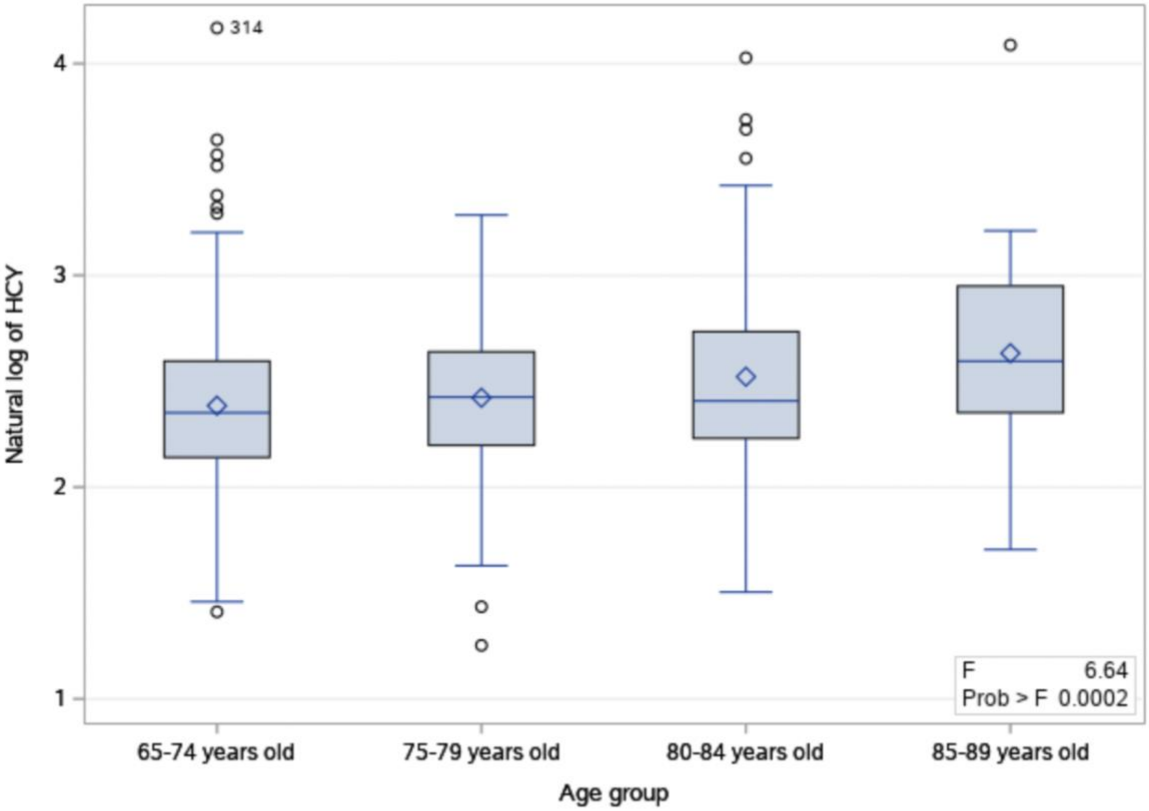
Variable	**N**	**F-value**	**p-value**	**Regression Coefficient (se)**	**R^2**
Age	663	17.93, df=1,661	< 0.0001	0.0135 (0.0032)	0.0264

Table 5. Simple linear regression predicting MMSE with age

Variable	**N**	**F-value**	**p-value**	**Regression Coefficient (se)**	**R^2**
Age	661	38.02, df=1,659	< 0.0001	-0.1406 (0.0228)	0.0546

We will now assess the relationship between log-transformed hcy and age once again, however, this time we will be using the categorical variable age called agegrp. Since log-transformed hcy is a continuous variable and age group is categorical, we will conduct a one-factor ANOVA test. We will assume that the data is derived from random samples, therefore assuming that the age groups are independent from one another. We will visually assess the normality of the distribution of log-transformed hcy among the different age groups and equality of variances through side-by-side boxplots, please refer to Figure 3. The boxplots show that the variances for log-transformed hcy among the different age groups are roughly equal; they are not drastically different from one another therefore we will assume equal variances. The distribution of log-transformed hcy within each age group also looks to be roughly normal, therefore we will assume normal distribution, although we should note that there seem to be a number of potential outliers in some of these age groups, therefore it is possible some of these age groups may have a skewed distribution of log-transformed hcy. With these assumptions met, we will move on to conduct our one-factor ANOVA with a significance level of 0.05. Our null hypothesis is that the mean log-transformed hcy are all equal for all age groups. Our alternative hypothesis is that the mean log-transformed hcy for the age groups are not all equal, at least one age group's log-transformed hcy is different.

Figure 3. Side-by-side boxplots of log hcy by age group



We reject the null hypothesis at the 0.05 significance level, there is significant evidence (N=663, F=6.64, df=3,659, p-value=0.0002) to suggest that the mean log-transformed hcy for the age groups are not all the same and that at least one age group's mean is different from the others. The mean log-transformed hcy for each group are as follows: 65-74 group = 2.3847 (N=337, sd=0.3664), 75-79 group =2.4218 (N=202, sd=0.3442), 80-84 group = 2.5212 (N=90, sd=0.4529), and 85-89 group = 2.6321 (N=34, sd=0.4543). In order to identify specific and large between group differences, we employed Tukey's post-hoc procedure at a significance level of 0.05. Our null hypotheses are as follows: (1) the mean log-transformed hcy measure for the 65-74 age group is the same as the mean for the 75-79 age group, (2) the mean log-transformed hcy measure for the 65-74 age group is the same as the mean for the 80-84 age group, (3) the mean log-transformed hcy measure for the 65-74 age group is the same as the mean for the 85-89 age group, (4) the mean log-transformed hcy measure for the 75-79 age group is the same as the mean for the 80-84 age group, (5) the mean log-transformed hcy measure for the 75-79 age group is the same as the mean for the 85-89 age group, and (6) the mean log-transformed hcy measure for the 80-84 age group is the same as the mean for the 85-89 age group. Our alternative hypotheses are as follows: (1) the mean log-transformed hcy measure for the 65-74 age group is not the same as the mean for the 75-79 age group, (2) the mean log-transformed hcy measure for the 65-74 age group is not the same as the mean for the 80-84 age group, (3) the mean log-transformed hcy measure for the 65-74 age group is not the same as the mean for the 85-89 age group, (4) the mean log-transformed hcy measure for the 75-79 age group is not the same as the mean for the 80-84 age group, (5) the mean log-transformed hcy measure for the 75-79 age group is not the same as the mean for the 85-89 age group, and (6) the mean log-transformed hcy measure for the 80-84 age group is not the same as the mean for the 85-89 age group. We also found that the R^2 for this model was 0.0293, meaning that 2.93% of the variability in log-transformed hcy can be explained by the variability in age group.

We reject the 2nd null hypotheses at a significance level of 0.05; there was significant evidence (N=663, t=-3.0467, df=659, p-value=0.0128) to suggest that the mean log-transformed hcy measure for the 65-74 age group is not the same as the mean for the 80-84 age group. We found that the 65-74 age group had on average a log-transformed hcy measure that was 0.1365 units less than the 80-84 age group. Similarly, we also reject the 3rd null hypothesis at a significance level of 0.05 and found that there was significant evidence (N=663, t=-3.6421, df=659, p-value=0.0017) to suggest that the mean log-transformed hcy measure for the 65-74 age group is the same as the mean for the 85-89 age group. We found that the 65-74 age group on average had a log-transformed hcy measure 0.2475 units less the 85-89 age group. Additionally, we also reject the 5th null hypothesis at a significance level of 0.05; there was significant evidence (N=663, t=-3.0045, df=659, p-value=0.0146) to suggest that the mean log-transformed found on average the 75-79 age group had a log-transformed hcy measure that was 0.2103 units less than the 85-89 age group.

On the other hand we failed to reject the 1st null hypothesis at a significance level of 0.05, there was not significant evidence (N=663, t=-1.1059, df=659, p-value=0.6860) to suggest that the mean log-transformed hcy measure for the 65-74 age group is not the same as the mean for the 75-79 age group. The difference in mean log-transformed hcy between the 65-74 group and the 75-79 group was -0.03716. We also fail to reject the 4th null hypothesis at a significance level of 0.05, there was not significant evidence (N=663, t=-2.0760, df=659, p-value=0.1621) to suggest that the mean log-transformed hcy measure for the 75-79 age group is not the same as the mean for the 80-84 age group. The difference in mean log-transformed hcy between the 75-79 group and the 80-84 group was -0.0994. Lastly, we also fail to reject the 6th null hypothesis at a significance level of 0.05, there was not significant evidence (N=663, t=-1.4597, df=659, p-value=0.4626) to suggest that the mean log-transformed hcy measure for the 80-84 age group is not the same as the mean for the 85-89 age group. The difference in mean log-transformed hcy between the 80-84 group and the 85-89 group was -0.1110. These results somewhat support the notion that there is a linear relationship between age and log-transformed hcy. If we strictly look at the means, we see that as age category increases ordinally so too does the mean log-transformed hcy for that age group, ie. the 65-74 age group has the smallest mean, while the 75-79 age group have a mean higher than the 65-74 age group but lower than the 80-84 age group etc. The difference in mean from one age group to the next also seems somewhat uniform if we look strictly at the numbers. This indicates that there is somewhat of a linear relationship. However, when we look at the pairwise comparisons only some comparisons were significant while others were not; the non-significant comparisons indicate that some of the differences in log-transformed hcy between groups were not so different from each other, thus the means in some comparisons were essentially the same and not significantly different from one another. The comparisons between groups that border each other in terms of age – for example, 65-74 vs. 75-79 or 80-84 vs. 85-89 – were not found significant, which indicates that as we go up in age group, the difference in mean is not significant enough and that the two means are virtually the same, which does not indicate a linear relationship. However, if we compare the extremes, such as the youngest age group to the oldest, then we find a significant difference in mean. In conclusion, there may be a slight linear relationship between log-transformed hcy and age, however, it may be a weak one. Additionally if we look at the side-by-side boxplots we can see that there is a semblance of a silght linear relationship, however the variance for all groups are fairly large and indicate there is a lot of scatter and does not show a very obvious linear relationship.

In summary, we found the mean log-transformed hcy was not all the same for the age groups and that at least one of our age groups was different from the others, please refer to Table 6 for the results of the 1-factor ANOVA global test. In particular, we found that age group 65-74 has a significantly different mean log-transformed hcy than both the 80-84 age group and the 85089 age group. Additionally, the 75-79 age group had a significantly different mean log-transformed hcy than the 85-89 age group. Please refer to Table 7 for the differences, t-values, and p-values for the pairwise comparisons results. The results of this ANOVA does not seem to indicate a true linear relationship. Although the mean log-transformed hcy increases as we increase in our age group, we only found significant differences in mean log-transformed hcy between the age groups that were furthest away from each other with at least a 10 year difference between those age groups. This indicates that there may not be a uniform relationship between age and log-transformed hcy that applies to all ages and that each age group may have their own individual relationship with log-transformed hcy.

Table 6. 1-factor ANOVA Global test results

Global Test (N=663)	**F-value**	**p-value**	**R^2**
	6.64, df=3,659	0.0002	0.0293
Age group	**N**	**LSMean**	**SE**
65-74	337	2.3847	0.3664
75-79	202	2.4218	0.3442
80-84	90	2.5212	0.4529
85-89	34	2.6321	0.4543

Table 7. Multiple comparisons Tukey test for differences in mean log hcy between age groups

	65-74	75-79	80-84	85-89

65-74		Diff=-0.0372
		t=-1.1059
		p=0.6880
			Diff=-0.1365
			t=-3.0468
			p=0.0128
				Diff=-0.2475
				t=-3.6421
				p=0.0017

75-79	Diff=0.0372
	t=1.1059
	p=0.6880
			Diff=-0.0994
			t=-2.0760
			p=0.1621
				Diff=-0.2103
				t=-3.0045
				p=0.0146

80-84	**Diff=0.1365**
	t=3.0468
	p=0.0128
		Diff=0.0994
		t=2.0760
		p=0.1621
				Diff=-0.1110
				t=-1.4597
				p=0.4626

85-89	**Diff=0.2475**
	t=3.6421
	p=0.0017
		Diff=0.2103
		t=3.0045
		p=0.0146
			Diff=0.1110
			t=1.4597
			p=0.4626	

So far, we have twice evaluated the relationship between log-transformed hcy and age: once with our continuous age variable and secondly with our categorical age variable. Now, we are interested in fitting a piecewise linear model on age to predict log-transformed hcy. In order to perform this piecewise linear regression model, we created four new variables: age1 for those between the ages 65-74, age2 for those between the ages 75-79, age 3 for those between the ages 80-84, and age4 for those between the ages 85-89. We used a series of if-then statements to carry out this piecewise linear regression. We first told SAS that if age was at least 65 but less than 75 that age1 would simply equal the age of that observation, however, if age was 75 or greater then age1 would simply equal 75 – the upper bound of this first age group. Next, we told SAS that if dose was at least 65 but less than 75 then age2 would equal 75 (lower bound of this second age group). Then if dose fell within the bounds of age2 (at least 75 but less than 80), then age2 would simply equal the dose of that observation. And then if dose was equal or above the upper bound (80) of age group 2 then age2 would equal the upper bound (80). Next, we told SAS that if dose was at least 65 but less than 80 then age3 would equal 80 (lower bound of this third age group). Then if dose fell within the bounds of age3 (at least 80 but less than 85), then age3 would simply equal the dose of that observation. And then if dose was equal or above the upper bound (85) of age group 3 then age3 would equal the upper bound (85). Lastly, if age was at least 65 but less than 85, age4 would equal its lower bound of 85, but if dose falls within the bounds of age group 4 (85-89) then age4 would simply equal the age of that observation. We did not have any ages above 89 in this dataset, therefore it was not necessary to account for ages above 89.

Using these four new piecewise variables we created, we will run the piecewise linear model predicting log-transformed hcy at the 0.05 significance level. Our null hypothesis is that our set of piecewise variables (age1, age2, age3, and age4) are not associated with log-transformed hcy, adjusting for one another. Our alternative hypothesis is that our set of piecewise variables (age1, age2, age3, and age4) are associated with log-transformed hcy, adjusting for one another. We reject our null hypothesis at the 0.05 significance level, there is significant evidence (N=663, F=4.91, df=4,658, p-value=0.0007) to suggest that there is an association between our set of piecewise variables (age1, age2, age3, and age4) and log-transformed hcy, adjusting for one another. The adjusted R^2 for this model was 0.0231, meaning that 2.31% of the variability in log-transformed hcy can be explained by the variability in the piecewise variables of age.

We will now move on to individual predictor tests for age1-age4. Our null hypotheses are as follows: (1) there is no association between being between 65-74 years old and log-transformed hcy, (2) there is no association between being between 75-79 years old and log-transformed hcy, (3) there is no association between being between 80-84 years old and log-transformed hcy, and (4) there is no association between being between 85-89 years old and log-transformed hcy. Our alternative hypotheses are as follows: (1) there is an association between being between 65-74 years old and log-transformed hcy, (2) there is an association between being between 75-79 years old and log-transformed hcy, (3) there is an association between being between 80-84 years old and log-transformed hcy, and (4) there is an association between being between 85-89 years old and log-transformed hcy. We fail to reject all 4 null hypotheses at the 0.05 significance level. There was not significant evidence (N=663, t=0.51, df=658, p-value=0.6119) to suggest that there is an association between being between 65-74 years old and log-transformed hcy, thus age1 is not a significant predictor of log-transformed hcy. Its regression coefficient was 0.0047 (se=0.0093), meaning that for every 1 year increase in age among those ages 65-74, on average we would expect an increase in log-transformed hcy of 0.0047 units. There was also not significant evidence (N=663, t=1.30, df=658, p-value=0.1943) to suggest that there is an association between being between 75-79 years old and log-transformed hcy, thus age2 is not a significant predictor of log-transformed hcy. Its regression coefficient was 0.0153 (se=0.0118), meaning that for every 1 year increase in age among those ages 65-74, on average we would expect an increase in log-transformed hcy of 0.0153 units. There was also not significant evidence (N=663, t=1.23, df=658, p-value=0.2188) to suggest that there is an association between being between 80-84 years old and log-transformed hcy, thus age3 is not a significant predictor of log-transformed hcy. Its regression coefficient was 0.0235 (se=0.0191), meaning that for every 1 year increase in age among those ages 80-84, on average we would expect an increase in log-transformed hcy of 0.0235 units. Lastly, there was not significant evidence (N=663, t=0.46, df=658, p-value=0.6471) to suggest that there is an association between being between 85-89 years old and log-transformed hcy, thus age4 is not a significant predictor of log-transformed hcy. Its regression coefficient was 0.0203 (se=0.0443), meaning that for every 1 year increase in age among those ages 85-89, on average we would expect an increase in log-transformed hcy of 0.0203 units. In summary, although the overall model predicting log-transformed hcy with age as piecewise variables was significant, the individual piecewise variables were all found to not be significant predictors of log-transformed hcy. Next, we will compare these 4 regression coefficients to assess whether any are significantly different from one another.

Our null hypotheses are as follows: (1) the regression coefficient for those between 65-74 years old and 75-79 years old are equal, (2) the regression coefficient for those between 65-74 years old and 80-84 years old are equal, (3) the regression coefficient for those between 65-74 years old and 85-89 years old are equal, (4) the regression coefficient for those between 75-79 years old and 80-84 years old are equal, (5) the regression coefficient for those between 75-79 years old and 85-89 years old are equal, and (6) the regression coefficient for those between 80-84 years old and 85-89 years old are equal. Our null hypotheses are as follows: (1) the regression coefficient for those between 65-74 years old and 75-79 years old are not equal, (2) the regression coefficient for those between 65-74 years old and 80-84 years old are not equal, (3) the regression coefficient for those between 65-74 years old and 85-89 years old are not equal, (4) the regression coefficient for those between 75-79 years old and 80-84 years old are not equal, (5) the regression coefficient for those between 75-79 years old and 85-89 years old are not equal, and (6) the regression coefficient for those between 80-84 years old and 85-89 years old are not equal. Again, we fail to reject all 6 null hypotheses at the 0.05 significance level.

There is not significant evidence to suggest that the regression coefficient for those between 65-74 years old and 75-79 years old are not equal (N=663, F=0.32, df=1, 658, p-value=0.5697), nor that the regression coefficient for those between 65-74 years old and 80-84 years old are not equal (N=663, F=0.88, df=1, 658, p-value=0.3490), nor that the regression coefficient for those between 65-74 years old and 85-89 years old are not equal (N=663, F=0.12, df=1, 658, p-value=0.7327), nor that the regression coefficient for those between 75-79 years old and 80-84 years old are not equal (N=663, F=0.09, df=1, 658, p-value=0.7680), nor that the regression coefficient for those between 75-79 years old and 85-89 years old are not equal (N=663, F=0.01, df=1, 658, p-value=0.9101), and nor that the regression coefficient for those between 80-84 years old and 85-89 years old are not equal (N=663, F=0.00, df=1, 658, p-value=0.9561). In summary, not only was the overall model significant and none of the individual predictors were significant, but none of the individual slopes (regression coefficients) for the individual predictors were significantly different from one another.

In summary, although we found the overall piecewise model to be significant, when we looked at the significance of individual regression coefficients of our age groups, none of them were found to be significant, please refer to Table 6 for the results of the piecewise model. When we looked at the pairwise comparisons for the slopes, we also did not find any slopes that were significantly different from one another. We can, therefore, conclude that the individual 4 slopes for the 4 age groups are not much different from each other, therefore, there is no purpose in our using a model that has 4 different slopes. We could very well just use a model with one slope and that would have worked just as well since we found that the 4 individual slopes were not much different from one another.

Table 6. Piecewise linear regression results

Global Test	**N**	**Adjusted R^2**	**F-value**	**p-value**
	663	0.0231	4.91, df=4,658	0.0007
Age group	**Regression Coefficient**	**SE**	**t-value**	**p-value**
65-74	0.0047	0.0093	0.51, df=658	0.6119
75-79	0.0513	0.0118	1.30, df=658	0.1943
80-84	0.0235	0.0191	1.23, df=658	0.2188
85-89	0.0203	0.0443	0.46, df=658	0.6471

We have thus far evaluated the relationship between log-transformed hcy and age in three different models: one where we used a continuous variable for age, another where we used a categorical variable for age, and lastly a piecewise continuous variable for age. We would not opt to use the piecewise continuous variable for age for our multiple linear regression analysis moving forward. As we saw in our analysis, we found that as a whole the age variable was significantly associated with log-transformed hcy, however, none of the individual piecewise continuous variables were found to be significantly associated with log-transformed hcy nor were any of the individual slopes significantly different from one another, therefore there does not seem to be a purpose for running a piecewise model with individual slopes for each age group – one overall slope seems to be sufficient. Now between using a continuous age variable vs a categorical age variable, we would be best off going with the categorical form of age. In our model using categorical age we found there was a slightly higher R^2 value of 0.0293 compared to 0.0264 from the model using continuous age – this means that the model using categorical age can account for more of the variability in log-transformed hcy than the model using continuous age. Additionally, our post-hoc analysis when using categorical age highlighted that there were significant differences in mean log-transformed hcy between some groups and not others, particularly there being significant differences between the younger age groups compared to the more older age groups. With this in mind, it may be more beneficial for us to use categorical age that recognizes that different age groups can have different relationships with the outcome as compared to continuous that assumes a linear and constant relationship between age and log-transformed hcy. As we discussed previously when we assess the relationship between categorical age and log-transformed hcy, this model seems to indicate that there if there is a linear relationship between predictor and outcome, it would be fairly weak given the results of the post-hoc multiple comparisons and which comparisons were found to be significant and which were not. If we found the model using categorical age to be more strongly and overtly linear, then we would have gone with a continuous age variable since it would have then lent itself well to simply running a regression model, however, since we are not entirely convinced it is a linear model we will go with categorical age.

Next, we will perform a multiple linear regression model with interaction predicting MMSE from log-transformed hcy and sex (dummy variable) as our predictors. Our global null hypothesis is that there is no association between our set of predictors (log-transformed hcy and sex) and our outcome, MMSE. Our alternative hypothesis is that there is an association between our set of predictors (log-transformed hcy and sex) and our outcome, MMSE. We are also interested in whether there is evidence of effect modification by sex on this relationship, therefore we have included an interaction term and will run a significance test of that as well. Our interaction null hypothesis is that there is no interaction between sex and log-transformed hcy. Our alternative hypothesis is that there is interaction between sex and log-transformed hcy. We reject the global null hypothesis at the 0.05 significance level, there is significant evidence (n=661, F=4.81, df=3,657, p-value=0.0025) to suggest that there is an association between our set of predictors (log-transformed hcy and sex) and our outcome, MMSE. On the other hand, however, we fail to reject the interaction null hypothesis at the 0.05 significance level, there was not significant evidence (N=661, F=0.11, df=1,657, p-value=0.7349) to suggest that there is interaction between sex and log-transformed hcy. Since we did not find significant interaction between sex and log-transformed hcy, there is, therefore, not significant evidence of effect measure modification by sex in this relationship. The lack of effect measure modification indicates that sex does not substantially modify the relationship between MMSE and log-transformed hcy, i.e. that a participants' MMSE-log-transformed hcy relationship is not affected by their sex. Males and females, therefore, experience similar MMSE-log-transformed hcy relationships. Since there was not significant interaction, we will proceed to remove the interaction term and refit our model and run parameter estimate tests on our predictors.

Our global null hypothesis remains the same: there is no association between our set of predictors (log-transformed hcy and sex) and our outcome, MMSE. Our global alternative hypothesis is that there is an association between our set of predictors (log-transformed hcy and sex) and our outcome, MMSE. Our parameter estimates null hypotheses are as follows: (1) log-transformed hcy is not associated with MMSE, in the presence of sex, and (2) sex is not associated with MMSE, in the presence of log-transformed hcy. Our parameter estimates alternative hypotheses are as follows: (1) log-transformed hcy is associated with MMSE, in the presence of sex, and (2) sex is associated with MMSE, in the presence of log-transformed hcy. We reject our global null hypothesis at the 0.05 significance level, there is significant evidence (N=661, F=7.17, df=2,658, p-value=0.0008) to suggest that there is an association between our set of predictors (log-transformed hcy and sex) and our outcome, MMSE. We see that once we removed the interaction term, our model remains significant and has even increased in significance as evidenced by the higher F value with a lower df and lower p-value overall. For our parameter estimates, we reject our first null hypothesis at the 0.05 significance level; there is significant evidence (N=661, t=-3.01, df=658, p-value=0.0027) to suggest that log-transformed hcy is associated with MMSE, in the presence of sex. The regression coefficient for log-transformed hcy is -0.8363, meaning that for every 1 unit increase in log-transformed hcy, on average we can expect a 0.8363 unit decreased in MMSE, in the presence of sex. Lastly, we also reject our parameter estimates second null hypothesis, there is significant evidence (N=661, t=2.12, df=658, p-value=0.0347) to suggest that sex is associated with MMSE, in the presence of log-transformed hcy. The adjusted mean MMSE for female participants was 28.2575 (se=0.1309), while the adjusted mean MMSE for male participants was 27.7832 (se=0.1817). The regression coefficient for sex was 0.4743 and this represents the difference in mean MMSE between females and males, our reference group, in the presence of log-transformed hcy. In this sample, females, on average, had a mean MMSE 0.4743 units higher than males.

In summary, in this dummy variable multiple linear regression model, we did not find significant interaction between log-transformed hcy and sex in predicting MMSE, therefore, we can conclude that there is no effect measure modification by sex in this relationship. The model with interaction was found to be significant, however, the insignificant interaction term had an unfavorable effect on our predictors. Once we removed the interaction term, we found that both log-transformed hcy and sex were significant predictors of MMSE, adjusting for one another. Please refer to Tables 7 and 8 for the results of both models with and without interaction. Log-transformed hcy has an inverse relationship with MMSE; as log-transformed hcy increases the MMSE decreases – in this case, as log-transformed hcy increases by 1 unit, we would expect MMSE to decrease by 0.8363 units in this model. Additionally, in this model it is predicted that on average females have an MMSE score 0.4743 units higher than males.

Table 7. Dummy variable multiple linear regression predicting MMSE with interaction

Global Test	**N**	**R^2**	**F-value**	**p-value**
	661	0.0215	4.81, df=3,657	0.0025
Predictors	**Regression Coefficient**	**SE**	**t-value**	**p-value**
Log hcy	-0.9707	0.4846	-2.00, df=657	0.0456
Sex (ref=male)	-0.0155	1.4629	-0.01, df=657	0.9915

Log hcy*Male	0.2006	0.5918	0.34, df=657	0.7349

Table 8. Dummy variable multiple linear regression predicting MMSE without interaction

Global Test	**N**	**R^2**	**F-value**	**p-value**
	661	0.0213	7.17, df=2,658	0.0008
Predictors	**Regression Coefficient**	**SE**	**t-value**	**p-value**
Log hcy	-0.8363	0.4846	-3.01, df=658	0.0027
Sex (ref=male)	0.4743	1.4629	2.12, df=658	0.0347

Next, we will fit a simple linear regression model predicting MMSE with just log-transformed hcy as our singular predictor at a significance level of 0.05. First, it may be helpful if we constructed a scatterplot to visually assess what the relationship between MMSE and log-transformed hcy looks like. From the scatterplot, it is difficult to tell whether this is a linear relationship or not, with much of the scatter focused at the top of the plot with the higher values of MMSE. We also see two points with very low MMSE scores much further away from the majority of the other points. There does not visually seem to be a linear relationship, or any other discernable trend between MMSE and log-transformed hcy. Therefore, as we move into the simple linear regression we should interpret the results with caution. Our null hypothesis is that log-transformed hcy is not linearly associated with MMSE. Our alternative hypothesis is that log-transformed hcy is linearly associated with MMSE. We reject the null hypothesis at the 0.05 significance level, there is significant evidence (N=661, F=9.81, df=1,659, p-value=0.0018) to suggest that there is a linear association between MMSE and log-transformed hcy. The regression coefficient is -0.8715, meaning that for every 1 unit increase in log-transformed hcy, we can expect on average for MMSE score to decrease by 0.8715 units. The R^2 for this model was 0.0147, meaning that only 1.47% of the variability in MMSE can be explained by the variability in log-transformed hcy. In the regression diagnostics, we should be concerned that th normality assumption does not seem to be met according to the QQ plot. If we look at the predicted vs. residuals plot, we should be concerned with equal variance as well due to the presence of a few large negative residuals. We are already aware that MMSE has a skewed distribution, therefore, it will be difficult to ascertain whether this simple linear regression model is an accurate representation of the relationship between MMSE and log-tranformed hcy and even whether there is a linear relationship to begin with. Therefore, we should take the results presented here and in Table 9 with caution.

Table 9. Simple linear regression predicting MMSE with log hcy

Variable	**N**	**F-value**	**p-value**	**Regression Coefficient (se)**	**R^2**
Log hcy	661	9.81, df=1,659	0.0018	-0.8715 (0.2782)	0.0147

Next we will fit a multiple linear regression model predicting MMSE with log-transformed hcy, sex, education, age, and pack years of cigarette smoking. We will use the categorical age variable in this model, which we discussed previously. We will be using the reg procedure in SAS to fit this model in order for us to get regression diagnostic output, therefore, before we can fit our model, we have to create 2 sets of dummy variables to represent education and age. We will create a new temporary dataset called fullmodel where we will set this new dataset to the temp dataset (obs=663, 18 variables) we have primarily been using thus far and we will add in these new dummy variables and use this new dataset to fit our model. We will be creating 3 dummy variables for the 4 levels of education, where the group who has at least some college (where educg=4) is our reference group. In order to create these new dummy variables we will be using a series of if-then statements. When the value of educg equals 1 then our first dummy variable less8 will equal 1, else if educg equals anything but 1 then less8 will equal 0. When the value of educg equals 2 then our second dummy variable grt8noHS will equal 1, else if educg equals anything but 2 then grt8noHS will equal 0. Lastly, when the value of educg equals 3 then our third dummy variable HSnocol will equal 1, else if educg equals anything but 3 then HSnocol will equal 0. Now the remaining values in the educg variable that have not been accounted for in the if-then statements is when educg the value 4, which stands in for at least some college education. When educg is equal to 4, this will be denoted by all three dummy variables less8, grt8noHS, and HSnocol being equal to 0, thereby making at least some college education (educg=4) our reference group. Next, we will be creating 3 dummy variables for the 4 levels of age, where the youngest group between the ages 65-74 is our reference group. Using a series of if-then statements, when the value of agegrp equals 1 (ages 75-79) then our first dummy variable age2 will equal 1, else if agegrp equals anything but 1 then age2 will equal 0. When the value of agegrp equals 2 (ages 80-84) then our second dummy variable age3 will equal 1, else if agegrp equals anything but 2 then age3 will equal 0. Lastly, when the value of agegrp equals 3 (ages 85-89) then our third dummy variable age4 will equal 1, else if agegrp equals anything but 3 then age4 will equal 0. Now the remaining values in the agegrp variable that have not been accounted for in the if-then statements is when agegrp equals 0, which stands in for the 65-74 age gruop. When agegrp is equal to 0, this will be denoted by all three dummy variables age2, age3, and age4 being equal to 0, thereby making the 65-74 age group (agegrp=0) our reference group. Our new fullmodel dataset has a total of 663 observations and 24 variables.

Before proceeding into the multiple linear regression model, we will first visually assess the data for linearity, constant variance, normality, and potential problem points. Previously, we have already assessed that based on the hypothesis test, MMSE and log-transformed hcy are linearly associated, although we would proceed with caution as visually there did not seem to be a linear relationship. Similarly, we also found previously that MMSE and age are linearly associated. Visually, pack years of cigarette smoking and MMSE do not seem to have a linear relationship nor does it seem to look like what we would expect from a scatterplot, similarly to the relationship between MMSE and log-transformed hcy. We created a side-by-side boxplot in order to look at the variances and potential linearity of education. We found that education and MMSE do seem to be linearly related, where as education categorically increases (from lower levels of education to higher levels of education) so does the mean MMSE, however, there does not seem to be constant variance as the IQR (and even the raw range) for those with less than 8 years of education is much about double the size of the other education groups. In terms of constant variance, we also assess side-by-side boxplots for sex and found that between males and females the variance is equal enough; neither was particularly much larger or smaller than the other. Side-by-side boxplots for age indicate that there are varying differences between the youngest and oldest age groups, however, generally speaking the variances are not wildly different therefore we will assume equal variance for age. We will also create side-by-side boxplots for our continuous predictors: log-transformed hcy and pack years of cigarette smoking to assess constant variance. We will use the rank procedure in SAS to turn these continuous variables into quintiles and then output side-by-side boxplots. According to these side-by-side boxplots, log-transformed lhcy and pack years of cigarette smoking have roughly equal variance, where for both variables the IQRs for the quintiles are roughly equal, although for both variables there are a fair amount of potential outliers among the quintiles. When it comes to normality, we already know that MMSE is clearly not normal and skewed left. When we created vertical bar charts earlier, we established that pack years of cigarette smoking was also not normal and heavily skewed right, log-transformed hcy was relatively normal, and age was relatively normally distributed although has a slight skew right. All in all, some of the assumptions were met for some of the variables, however, there were a number of assumptions that were not met for other variables, in which case we will continue to fit this model predicting MMSE with the predictors that were requested, but we should be cautious.

Now we will head into our multiple linear regression model predicting MMSE with log-transformed hcy, education, age, and pack years of cigarette smoking using a significance level of 0.05. Our null hypothesis is that our set of predictors (log-transformed hcy, sex, education, age, and pack years of cigarette smoking) are not associated with our outcome, MMSE. Our alternative hypothesis is that our set of predictors (log-transformed hcy, sex, education, age, and pack years of cigarette smoking) are associated with our outcome, MMSE; that at least one of our predictors is associated with the outcome. We reject the null hypothesis at a significance level of 0.05, there is significant evidence (N=594, F=11.50, df=9, 584, p-value < 0.0001) to suggest that our set of predictors (log-transformed hcy, sex, education, age, and pack years of cigarette smoking) are associated with our outcome, MMSE. The adjusted R^2 value in this model is 0.1375, meaning that 13.75% of the variability in MMSE can be accounted for by the variability in our set of predictors. Since we found our set of predictors to be significant, we will now conduct individual predictor tests to see which variables were significant predictors of MMSE. Our null hypotheses are as follows: (1) log-transformed hcy is not associated with MMSE, in the presence of the other predictors – sex, education, age, and pack years of cigarette smoking, (2) sex is not associated with MMSE, in the presence of the other predictors – log-transformed hcy, education, age, and pack years of cigarette smoking, (3) education is not associated with MMSE, in the presence of the other predictors – log-transformed hcy, sex, age, and pack years of cigarette smoking, (4) age is not associated with MMSE, in the presence of the other predictors – log-transformed hcy, sex, education, and pack years of cigarette smoking, and (5) pack years of cigarette smoking is not associated with MMSE, in the presence of the other predictors – log-transformed hcy, sex, age, education. Our alternative hypotheses are as follows: (1) log-transformed hcy is associated with MMSE, in the presence of the other predictors – sex, education, age, and pack years of cigarette smoking, (2) sex is associated with MMSE, in the presence of the other predictors – log-transformed hcy, education, age, and pack years of cigarette smoking, (3) education is associated with MMSE, in the presence of the other predictors – log-transformed hcy, sex, age, and pack years of cigarette smoking, (4) age is associated with MMSE, in the presence of the other predictors – log-transformed hcy, sex, education, and pack years of cigarette smoking, and (5) pack years of cigarette smoking is associated with MMSE, in the presence of the other predictors – log-transformed hcy, sex, age, education.

We reject 3rd null hypothesis at the 0.05 significance level, there is significant evidence (N=594, F=18.67, df=3,584, p-value < 0.0001) to suggest that education is associated with MMSE, in the presence of the other predictors – log-transformed hcy, sex, age, and pack years of cigarette smoking. Our reference group is the education group whom had at least some college education and we are interested in comparing the mean MMSE from other education groups to the reference group in order to assess which education groups are significantly different from one another. Our null hypotheses are as follows: (1) those with less than 8 years of education have a mean MMSE score equal to those with at least some college education, adjusting for log-transformed hcy, sex, age, and pack years of cigarette smoking, (2) those with 8 or more years of education but no high school degree have a mean MMSE score equal to those with at least some college education, adjusting for log-transformed hcy, sex, age, and pack years of cigarette smoking, and (3) those with a high school degree, but no college education have a mean MMSE score equal to those with at least some college education, adjusting for log-transformed hcy, sex, age, and pack years of cigarette smoking. Our alternative hypotheses are as follows: (1) those with less than 8 years of education have a mean MMSE score not equal to those with at least some college education, adjusting for log-transformed hcy, sex, age, and pack years of cigarette smoking, (2) those 8 or more years of education but no high school degree have a mean MMSE score not equal to those with at least some college education, adjusting for log-transformed hcy, sex, age, and pack years of cigarette smoking, and (3) those with a high school degree, but no college education have a mean MMSE score not equal to those with at least some college education, adjusting for log-transformed hcy, sex, age, and pack years of cigarette smoking. We reject all three null hypotheses at the 0.05 significance level, there is significant evidence to suggest that those with less than 8 years of education have a mean MMSE score different from those with at least some college education (N=594, t=-5.63, df=584, p-value < 0.0001), adjusting for log-transformed hcy, sex, age, and pack years of cigarette smoking; that those with 8 years or more of education but no high school degree have a mean MMSE score different from those with at least some college education (N=594, t=-5.75, df=584, p-value < 0.0001), adjusting for log-transformed hcy, sex, age, and pack years of cigarette smoking; and that those with a high school degree, but no college education have a mean MMSE score not equal to those with at least some college education (N=594, t=-2.01, df=584 p-value=0.0449), adjusting for log-transformed hcy, sex, age, and pack years of cigarette smoking. The mean MMSE for each education group are as follows: less than 8 years = 25.7368 (sd=4.2537), 8 or more years, but no high school degree = 27.0169 (sd=3.8991), high school degree, but no college education = 28.4174 (sd=2.1119), and at least some college = 28.8571 (sd=1.3620). On average, those with less than 8 years of education had a mean MMSE score 3.4332 (se=0.6098) units less compared to those with at least some college education with a 95% confidence interval of (-4.6307, -2.2356), adjusted for the other predictors – log-transformed hcy, sex, age, and pack years of cigarette smoking. The standardized regression coefficient for this dummy variable predictor was -0.2225, meaning that when all predictors are standardized to the same scale, those with less than 8 years of education would on average have MMSE scores 0.2225 standardized units less than those with at least some college education. Alternatively, this standardized regression coefficient may indicate that it is one of the stronger predictors of MMSE score as it is the second largest magnitude standardized coefficient. On average, those with 8 or more years of education, but no high school degree had a mean MMSE score 1.4886 (se=0.2590) units less compared to those with at least some college education with a 95% confidence interval of (-1.9972, -0.9799), adjusted for the other predictors – log-transformed hcy, sex, age, and pack years of cigarette smoking. The standardized regression coefficient for this dummy variable predictor was -0.2571, meaning that when all predictors are standardized to the same scale, those with 8 or more years of education, but no high school degree would on average have MMSE scores 0.2571 standardized units less than those with at least some college education. Alternatively, this standardized regression coefficient may indicate that it is the strongest predictor of MMSE score as it has the largest magnitude out of all the standardized coefficients. On average, those with a high school degree, but no college education had a mean MMSE score 0.4750 (se=0.2364) units less compared to those with at least some college education with a 95% confidence interval of (-0.9393, -0.0108), adjusted for the other predictors – log-transformed hcy, sex, age, and pack years of cigarette smoking. The standardized regression coefficient for this dummy variable predictor was -0.0892, meaning that when all predictors are standardized to the same scale, those with a high school degree, but no college education would on average have MMSE scores 0.0892 standardized units less than those with at least some college education. If we were to repeat this study many times with the same sample size (N=594), we would expect 95% of resulting confidence intervals to contain the true mean difference in MMSE score between those with less than 8 years of education, those with 8 or more years, but no high school degree, and those with a high school degree, but no college education, all compared to those with at least some college education (our reference group), all adjusted for the other predictors – log-transformed hcy, sex, education, and pack years of cigarette smoking. We hope that our confidence intervals of (-4.6307, -2.2356), (-1.9972, -0.9799), and (-0.9393, -0.0108) are all part of the 95% of confidence intervals that contain the true mean difference in MMSE between those with less than 8 years of education compared to those with at least some college education, those 8 or more years of education, but no high school degree compared to those with at least some college education, and those a high school degree, but no college education compared to those with at least some college education, respectively and all adjusted for the other predictors – log-transformed hcy, sex, age, and pack years of cigarette smoking.

We also reject the 4th null hypothesis at the 0.05 significance level, there is significant evidence (N=594, F=7.42, df=3,584, p-value < 0.0001) to suggest that age is associated with MMSE, in the presence of the other predictors – log-transformed hcy, sex, education, and pack years of cigarette smoking. Our reference group is 65-74 age group and we are interested in comparing the mean MMSE from the other age groups to the reference group in order to assess which age groups are significantly different from one another. Our null hypotheses are as follows: (1) those in the 75-79 age group have a mean MMSE score equal to those in the 65-74 age group, adjusting for log-transformed hcy, sex, education, and pack years of cigarette smoking, (2) those in the 80-84 age group have a mean MMSE score equal to those in the 65-74 age group, adjusting for log-transformed hcy, sex, education, and pack years of cigarette smoking, and (3) those in the 85-89 age group have a mean MMSE score equal to those in the 65-74 age group, adjusting for log-transformed hcy, sex, education, and pack years of cigarette smoking. Our alternative hypotheses are as follows: (1) those in the 75-79 age group have a mean MMSE score not equal to those in the 65-74 age group, adjusting for log-transformed hcy, sex, education, and pack years of cigarette smoking, (2) those in the 80-84 age group have a mean MMSE score not equal to those in the 65-74 age group, adjusting for log-transformed hcy, sex, education, and pack years of cigarette smoking, and (3) those in the 85-89 age group have a mean MMSE score not equal to those in the 65-74 age group, adjusting for log-transformed hcy, sex, education, and pack years of cigarette smoking. We reject the first and the third null hypotheses at the 0.05 significance level, there is significant evidence to suggest that those in the 75-79 age group have a mean MMSE score different from those in the 65-74 age group (N=594, t=-2.11, df=584, p-value=0.0353), adjusting for log-transformed hcy, sex, education, and pack years of cigarette smoking and that those in the 85-89 age group have a mean MMSE score different from those in the 65-74 age group (N=594, t=-4.50, df=584, p-value < 0.0001), adjusting for log-transformed hcy, sex, education, and pack years of cigarette smoking. On the other hand, we fail to reject the second null hypothesis at the 0.05 significance level, there is not significant evidence (N=594, t=-1.82, df=584, p-value=0.0688) to suggest that those in the 80-84 age group have a mean MMSE score different from those in the 65-74 age group, adjusting for log-transformed hcy, sex, education, and pack years of cigarette smoking. The mean MMSE for each age group are as follows: 65-74 age group (N=337) = 28.5549 (sd=1.7856), 75-79 age group (N=202) = 27.9257 (sd=2.4996), 80-84 age group (N=89) = 27.7640 (sd=3.4510), and 85-89 age group (N=33) = 25.3333 (sd=6.3525). In this sample, on average, those in the 75-79 age group had a mean MMSE score 0.4782 (se=0.2267) units less compared to those in the 65-74 age group with a 95% confidence interval of (-0.9235, -0.0330), adjusting for log-transformed hcy, sex, education, and pack years of cigarette smoking. The standardized regression coefficient for this dummy variable predictor was -0.0859, meaning that when all predictors are standardized to the same scale, those with less than 8 years of education would on average have MMSE scores 0.0895 standardized units less than those with at least some college education. Alternatively, this standardized regression coefficient may indicate that it is not one of the strongest predictors of MMSE score as it contributes very little to MMSE score, compared to, for example, the having 8 or more years of education, but no high school degree dummy which contributes a larger magnitude to MMSE score of -0.2751. On average, those in the 85-89 age group had a mean MMSE score 2.1107 (se=0.4693) units less compared to those in the 65-74 age group with a 95% confidence interval of (-3.0324, -1.1890), adjusted for the other predictors – log-transformed hcy, sex, education, and pack years of cigarette smoking. The standardized regression coefficient for this dummy variable predictor was -0.1796, meaning that when all predictors are standardized to the same scale, those with 8 or more years of education, but no high school degree would on average have MMSE scores 0.1796 standardized units less than those with at least some college education. Alternatively, this standardized regression coefficient may indicate that it is a moderate predictor of MMSE score as it has one of the larger magnitudes out of all the standardized coefficients. Although not significant, on average, those in the 80-84 age group had a mean MMSE score 0.5677 (se=0.3115) units less compared to those in the 65-74 age group with a 95% confidence interval of (-1.1794, 0.0440), adjusted for the other predictors – log-transformed hcy, sex, education, and pack years of cigarette smoking. The standardized regression coefficient for this dummy variable predictor was -0.0745, meaning that when all predictors are standardized to the same scale, those with a high school degree, but no college education would on average have MMSE scores 0.0745 standardized units less than those with at least some college education. If we were to repeat this study many times with the same sample size (N=594), we would expect 95% of resulting confidence intervals to contain the true mean difference in MMSE score between those with less than 8 years of education, those with 8 or more years, but no high school degree, and those with a high school degree, but no college education, all compared to those with at least some college education (our reference group), all adjusted for the other predictors – log-transformed hcy, sex, education, and pack years of cigarette smoking. We hope that our confidence intervals of (-0.9235, -0.0330), (-1.1794, 0.0440), and (-3.0324, -1.1890) are all part of the 95% of confidence intervals that contain the true mean difference in MMSE between the 75-79 age group compared to those in the 65-74 age group, those in the 80-84 age group compared to those in the 65-74 age group, and those in the 85-89 age group compared to those in the 65-74 age group, respectively and all adjusted for the other predictors – log-transformed hcy, sex, education, and pack years of cigarette smoking.

We fail to reject our 1st null hypothesis at the 0.05 significance level, there is not significant evidence (N=594, t=-1.40, df=584, p-value=0.1628) to suggest that log-transformed hcy is associated with MMSE, in the presence of the other predictors – sex, education, age, and pack years of cigarette smoking. The regression coefficient for age was -0.3616 (se=0.2588), meaning that for every 1 unit increase in log-transformed hcy, we would expect on average for MMSE score to decrease by 0.3616 units, in the presence of the other predictors – sex, education, age, and pack years of cigarette smoking. The standardized regression coefficient for log-transformed hcy was - 0.0547, meaning that when all predictors were standardized to the same scale, for every 1 standardized unit increase in log-transformed hcy on average we would expect MMSE score to decrease by 0.0547 units. This standardized coefficient is much smaller than the standardized coefficient of having less than 8 years of education dummy (as compared to some college education) variable, which was -0.2224 , indicating that log-transformed hcy does not have a large effect on MMSE score. One log-transformed hcy standardized unit on average decreases MMSE by 0.0548 units, while having less than 8 years of education decreases MMSE by 0.2224 units – a much larger impact. This regression coefficient had a 95% confidence interval of (-0.8699, 0.1466). If we were to repeat this study many times with the same sample size (N=594), 95% of the resulting confidence intervals would contain the true regression coefficient for log-transformed hcy. We hope that our confidence interval of (-0.8699, 0.1466) is a part of that 95% that contains the true regression coefficient for log-transformed hcy, adjusting for the other predictors – sex, education, age, and pack years of cigarette smoking.

Additionally, we also fail to reject the 2nd null hypothesis at the 0.05 significance level, there is not significant evidence (N=594, t=-1.67, df=584, p-value=0.0955) to suggest that sex is not associated with MMSE, in the presence of the other predictors – log-transformed hcy, education, age, and pack years of cigarette smoking. The mean MMSE score for males was 27.7566 (sd=2.7613) and the mean MMSE for females was 28.2713. The regression coefficient for sex (a 0/1 dummy variable) was -0.3651, meaning that on average, males had a mean MMSE score 0.3651 units less than females with a 95% confidence interval of (-0.7946, 0.0644) adjusting for log-transformed hcy, education, age, and pack years of cigarette smoking. The standardized regression coefficient for sex hcy was -0.0670, meaning that when all predictors were standardized to the same scale, males had a MMSE score on average 0.0680 units less than females. This standardized coefficient is much smaller than the standardized coefficient of having less than 8 years of education dummy (as compared to some college education) variable, which was -0.2224, indicating that sex does not have a large effect on MMSE score. If we were to repeat this study many times with the same sample size (N=594), 95% of the resulting confidence intervals would contain the true difference in mean MMSE score between males and females, adjusting for log-transformed hcy, education, age, and pack years of cigarette smoking. We hope that our confidence interval of (-0.7946, 0.0644) are a part of the 95% of confidence intervals that contain the true difference in mean MMSE score between males and females, adjusting for log-transformed hcy, education, age, and pack years of cigarette smoking.

Lastly, we fail to reject the 5th null hypothesis at the 0.05 significance level, there is not significant evidence (N=594, t=0.67, df=584, p-value=0.5017) to suggest pack years of cigarette smoking is associated with MMSE, in the presence of the other predictors – log-transformed hcy, sex, age, education. The regression coefficient for pack years of cigarette smoking was 0.0033 (se=0.0049), meaning that for every 1 pack-year increase in cigarette smoking, we would expect on average for MMSE score to increase by 0.0033 units, in the presence of the other predictors – log-transformed hcy, sex, education, and age. The standardized regression coefficient for log-transformed hcy was 0.0272, meaning that when all predictors were standardized to the same scale, for every 1 standardized unit increase in pack years of cigarette smoking on average we would expect MMSE score to increase by 0.0282 units. This standardized coefficient has a smaller magnitude than the standardized coefficient of having less than 8 years of education dummy (as compared to some college education) variable, which was -0.2224, indicating that pack years of cigarette smoking does not have a large effect on MMSE score. One pack years of cigarette smoking standardized unit on average increases MMSE by a mere 0.0282 units, while having less than 8 years of education compared to at elast some college education would decrease MMSE by 0.1744 – a much larger impact on MMSE than pack years. This regression coefficient had a 95% confidence interval of (-0.0063, 0.0128). If we were to repeat this study many times with the same sample size (N=594), 95% of the resulting confidence intervals would contain the true regression coefficient for pack years of cigarette smoking. We hope that our confidence interval of (-0.0063, 0.0128) is a part of that 95% that contains the true regression coefficient for pack years of cigarette smoking, adjusting for the other predictors – log-transformed hcy, sex, education, and age.

Table 10. Full model multiple linear regression predicting MMSE results

Variable	N	F-value/t-value	p-value	Regression Coefficient (se)	Adj-R^2
Global Test	594	11.50, df=9,584	< 0.0001	N/A	0.1375
Log hcy		-1.40, df=584	0.1628	-0.3616 (0.2588)	
Sex		-1.67, df=584	0.0955	-0.3651 (0.2187)	

Education		18.67, df=3,584	< 0.0001	N/A	
Age		7.42, df=3,584	< 0.0001	N/A	
Cigarette smoking		0.67, df=584	0.5017	0.0033 (0.0049)	

In summary, education and age were found to be significant predictors of MMSE, while log-transformed hcy, sex, and pack years of cigarette smoking were not. Therefore, we will proceed to remove log-transformed hcy, sex, and pack years of cigarette smoking and refit our model. Our final model will now predict MMSE from education and age. Our null hypothesis is that there is no association between our predictors, education and age, and our outcome MMSE. Our alternative hypothesis is that there is an association between our predictors, education and age, and our outcome MMSE. We reject the null hypothesis at the 0.05 significance level, our model remains significant and there is significant evidence (N=661, F=17.48. df=6, 654, p-value < 0.0001) to suggest that there is an association between our predictors, education and age, and our outcome MMSE. The adjusted R^2 for this model was 0.1303 meaning that 13.03% of the variability in MMSE can be accounted for by the variability in our predictors – education and age. Moving on to our individual predictor tests, our null hypotheses are as follows: (1) education is not associated with MMSE, in the presence of age, and (2) age is not associated with MMSE, in the presence of education. Our alternative hypotheses are as follows: (1) education is associated with MMSE, in the presence of age, and (2) age is associated with MMSE, in the presence of education.

We reject the first null hypothesis at the 0.05 significance level, there is significant evidence (N=661, F=17.85, df=3,654, p-value < 0.0001) to suggest that education is associated with MMSE, in the presence of age. Again, our reference group is the education group whom had at least some college education and we are interested in comparing the mean MMSE from other education groups to the reference group in order to assess which education groups are significantly different from one another. Our null hypotheses are the same as previously: (1) those with less than 8 years of education have a mean MMSE score equal to those with at least some college education, adjusting for age, (2) those with 8 or more years of education but no high school degree have a mean MMSE score equal to those with at least some college education, adjusting for age, and (3) those with a high school degree, but no college education have a mean MMSE score equal to those with at least some college education, adjusting for age. Our alternative hypotheses are as follows: (1) those with less than 8 years of education have a mean MMSE score not equal to those with at least some college education, adjusting for age, (2) those 8 or more years of education but no high school degree have a mean MMSE score not equal to those with at least some college education, adjusting for age, and (3) those with a high school degree, but no college education have a mean MMSE score not equal to those with at least some college education, adjusting for age. This time around, we only reject the first and second null hypotheses at the 0.05 significance level, there is significant evidence to suggest that those with less than 8 years of education have a mean MMSE score different from those with at least some college education (N=661, t=-4.65, df=654, p-value < 0.0001), adjusting for age, and that those with 8 years or more of education but no high school degree have a mean MMSE score different from those with at least some college education (N=661, t=-6.15, df=654, p-value < 0.0001), adjusting for age. On the other hand, we fail to reject the third null hypothesis at the 0.05 significance level, there is not significant evidence (N=661, t=-1.74, df=654, p-value=0.0820) to suggest that those with a high school degree, but no college education have a mean MMSE score not equal to those with at least some college education, adjusting for age. The mean MMSE for each education group are as follows: less than 8 years (N=19) = 25.7368 (sd=4.2537), 8 or more years, but no high school degree (N=178) = 27.0169 (sd=3.8991), high school degree, but no college education (N=242) = 28.4174 (sd=2.1119), and at least some college (N=210) = 28.8571 (sd=1.3620). On average, those with less than 8 years of education had a mean MMSE score 2.8575 (se=0.6151) units less compared to those with at least some college education with a 95% confidence interval of (-4.0653, -1.6497), adjusted for age. The standardized regression coefficient for this dummy variable predictor was -0.1735, meaning that when all predictors are standardized to the same scale, those with less than 8 years of education would on average have MMSE scores 0.1735 standardized units less than those with at least some college education. Alternatively, this standardized regression coefficient may indicate that it is one of the stronger predictors of MMSE score as it is the third largest magnitude standardized coefficient. On average, those with 8 or more years of education, but no high school degree had a mean MMSE score 1.6031 (se=0.2605) units less compared to those with at least some college education with a 95% confidence interval of (-2.1145, -1.0916), adjusted for age. The standardized regression coefficient for this dummy variable predictor was -0.2584, meaning that when all predictors are standardized to the same scale, those with 8 or more years of education, but no high school degree would on average have MMSE scores 0.2584 standardized units less than those with at least some college education. Alternatively, this standardized regression coefficient may indicate that it is the strongest predictor of MMSE score/has the highest influence as it has the largest magnitude out of all the standardized coefficients. Although not significant, on average, those with a high school degree, but no college education had a mean MMSE score 0.4178 (se=0.2399) units less compared to those with at least some college education with a 95% confidence interval of (-0.8888, 0.0532), adjusted for age. The standardized regression coefficient for this dummy variable predictor was -0.0732, meaning that when all predictors are standardized to the same scale, those with a high school degree, but no college education would on average have MMSE scores 0.0732 standardized units less than those with at least some college education. Compared to the other education groups, this is not one of the strongest predictors of MMSE. If we were to repeat this study many times with the same sample size (N=594), we would expect 95% of resulting confidence intervals to contain the true mean difference in MMSE score between those with less than 8 years of education, those with 8 or more years, but no high school degree, and those with a high school degree, but no college education, all compared to those with at least some college education (our reference group), all adjusted for age. We hope that our confidence intervals of (-4.0653, -1.6497), (-2.1145, -1.0916), and (-0.8888, 0.0532) are all part of the 95% of confidence intervals that contain the true mean difference in MMSE between those with less than 8 years of education compared to those with at least some college education, those 8 or more years of education, but no high school degree compared to those with at least some college education, and those a high school degree, but no college education compared to those with at least some college education, respectively and all adjusted for age.

We also reject the second null hypothesis at the 0.05 significance level, there is significant evidence (N=661, F=12.14, df=3, 654, p-value < 0.0001) to suggest that age is associated with MMSE, in the presence of education. Again, our reference group is 65-74 age group and we are interested in comparing the mean MMSE from the other age groups to the reference group in order to assess which age groups are significantly different from one another. Our null hypotheses remain the same as previously and are as follows: (1) those in the 75-79 age group have a mean MMSE score equal to those in the 65-74 age group, adjusting for education, (2) those in the 80-84 age group have a mean MMSE score equal to those in the 65-74 age group, adjusting for education, and (3) those in the 85-89 age group have a mean MMSE score equal to those in the 65-74 age group, adjusting for education. Our alternative hypotheses are as follows: (1) those in the 75-79 age group have a mean MMSE score not equal to those in the 65-74 age group, adjusting education, (2) those in the 80-84 age group have a mean MMSE score not equal to those in the 65-74 age group, adjusting for education, and (3) those in the 85-89 age group have a mean MMSE score not equal to those in the 65-74 age group, adjusting for education. This time around, we reject all three null hypotheses at the 0.05 significance level, there is significant evidence to suggest that those in the 75-79 age group have a mean MMSE score different from those in the 65-74 age group (N=661, t=-2.21, df=654, p-value=0.0275), adjusting for education, that those in the 80-84 age group have a mean MMSE score different from those in the 65-74 age group (N=661, t=-1.96, df=654, p-value=0.0499), adjusting education, and that those in the 85-89 age group have a mean MMSE score different from those in the 65-74 age group (N=661, t=-5.87 df=654, p-value < 0.0001), adjusting for education. The mean MMSE for each age group remains as follows: 65-74 age group (N=337) = 28.5549 (sd=1.7856), 75-79 age group (N=202) = 27.9257 (sd=2.4996), 80-84 age group (N=89) = 27.7640 (sd=3.4510), and 85-89 age group (N=33) = 25.3333 (sd=6.3525). In this sample, on average, those in the 75-79 age group had a mean MMSE score 0.5088 (se=0.2304) units less compared to those in the 65-74 age group with a 95% confidence interval of (-0.9611, -0.0565), adjusting for education. The standardized regression coefficient for this dummy variable predictor was -0.0852, meaning that when all predictors are standardized to the same scale, those with less than 8 years of education would on average have MMSE scores 0.0852 standardized units less than those with at least some college education. Alternatively, this standardized regression coefficient may indicate that it is not one of the strongest predictors of MMSE score as it contributes very little to MMSE score, compared to some of the education dummy variables discussed previously. On average, those in the 80-84 age group had a mean MMSE score 0.6073 (se=0.3092) units less compared to those in the 65-74 age group with a 95% confidence interval of (-1.2144, 0.0002), adjusted for education. The standardized regression coefficient for this dummy variable predictor was -0.0753, meaning that when all predictors are standardized to the same scale, those with a high school degree, but no college education would on average have MMSE scores 0.0753 standardized units less than those with at least some college education. On average, those in the 85-89 age group had a mean MMSE score 2.7762 (se=0.4727) units less compared to those in the 65-74 age group with a 95% confidence interval of (-3.7043, -1.8480), adjusted for education. The standardized regression coefficient for this dummy variable predictor was -0.2197, meaning that when all predictors are standardized to the same scale, those with 8 or more years of education, but no high school degree would on average have MMSE scores 0.2197 standardized units less than those with at least some college education. This standardized regression coefficient indicates a stong predictor of MMSE score as it has one of the larger magnitudes out of all the standardized coefficients. If we were to repeat this study many times with the same sample size (N=661), we would expect 95% of resulting confidence intervals to contain the true mean difference in MMSE score between those with less than 8 years of education, those with 8 or more years, but no high school degree, and those with a high school degree, but no college education, all compared to those with at least some college education (our reference group), all adjusted for education. We hope that our confidence intervals of (-0.9611, -0.0565), (-1.2144, 0.0002), and (-3.7043, -1.8480) are all part of the 95% of confidence intervals that contain the true mean difference in MMSE between the 75-79 age group compared to those in the 65-74 age group, those in the 80-84 age group compared to those in the 65-74 age group, and those in the 85-89 age group compared to those in the 65-74 age group, respectively and all adjusted for education.

Table 11. Final model multiple linear regression predicting MMSE results

Variable	**N**	**F-value/t-value**	**p-value**	**Regression Coefficient (se)**	**Adj-R^2**
Global Test	661	17.48, df=6,654	< 0.0001	N/A	0.1303
Education (ref=at least some college)		17.85, df=3,654	< 0.0001	N/A	
< 8 years		-4.65, df=654	< 0.0001	-2.8575 (0.6151)	
≥ 8 years, no HS degree		-6.15, df=654	< 0.0001	-1.6031 (0.2605)	
HS degree, no college		-1.74, df=654	0.0820	-0.4178 (0.2399)	
Age (ref= 65-74)		12.14, df=3,654	< 0.0001	N/A	
75-79		-2.21, df=654	0.0275	-0.5088 (0.2304)	
80-84		-1.96, df=654	0.0499	-0.6073 (0.3092)	
85-89		-5.87, df=654	< 0.0001	-2.7762 (-0.4727)	

We will run a goodness of fit measure using C(p) to see how well our final model predicting MMSE using education and age holds up. If we look at Figure 4, we see that C(p) is at its lowest/equals the number of parameters – in this case, there are 10 parameters in our full model – when there are somewhere around 8 or 9 predictors in our model. When there are 8 predictors in the model that is also when we have the highest adjusted R^2 value of 0.1383, which has a corresponding C(p) statistic of 8.4519. SAS has included in this model log-transformed hcy, sex, and all dummy variables for education and age, ie. the only predictor SAS has dropped from the model was pack years of cigarette smoking. Our final model that only includes the dummy variables for education and age (which corresponds to the model that has 6 parameters that SAS output) is also comparable to the model with 8 parameters in it. Our model that predicts MMSE with just education and age has an adjusted R^2 of 0.1346 and a C(p) statistic of 8.9123, which still fits the goodness of fit criteria for C(p) statistic which is that C(p) should be as equal to or less than the number of parameters, in this case 10. Please see table 12. for the different models SAS outputted using the R^2 selection method.

Figure 4. C(p) vs. number of parameters plot

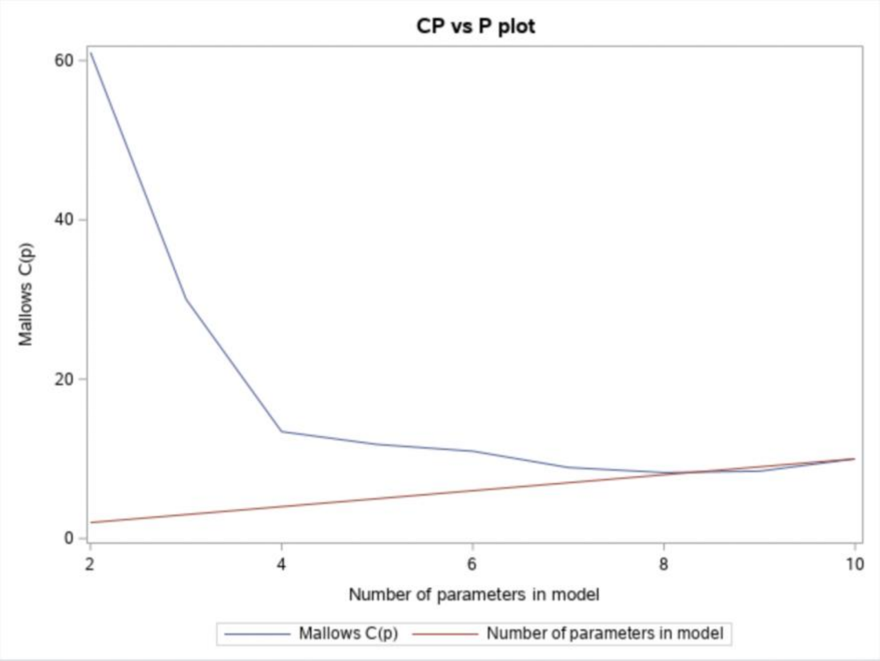
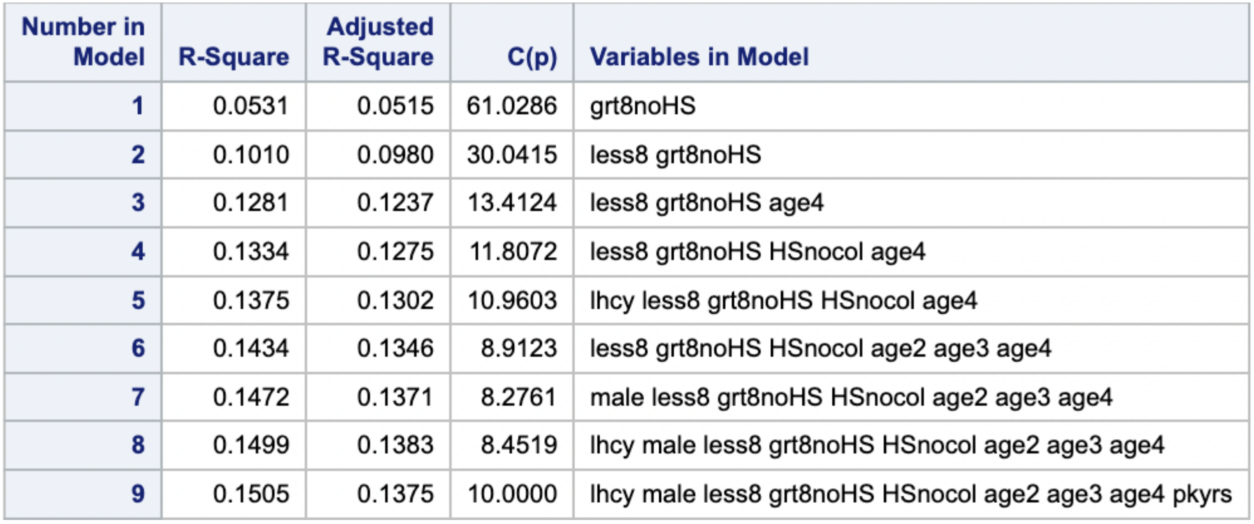


Table 12. Models using R^2 selection method



Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	0.0531	0.0515	61.0286	grt8noHS
2	0.1010	0.0980	30.0415	less8 grt8noHS
3	0.1281	0.1237	13.4124	less8 grt8noHS age4
4	0.1334	0.1275	11.8072	less8 grt8noHS HSnocol age4
5	0.1375	0.1302	10.9603	lhcy less8 grt8noHS HSnocol age4
6	0.1434	0.1346	8.9123	less8 grt8noHS HSnocol age2 age3 age4
7	0.1472	0.1371	8.2761	male less8 grt8noHS HSnocol age2 age3 age4
8	0.1499	0.1383	8.4519	lhcy male less8 grt8noHS HSnocol age2 age3 age4
9	0.1505	0.1375	10.0000	lhcy male less8 grt8noHS HSnocol age2 age3 age4 pkyrs

We will now assess whether there is potential collinearity in this new final model of our predicting MMSE with education and age. We will use a variance inflation factor (VIF) cut off of 10 to indicate which factors may be collinear. According to our output, none of the predictors have a VIF close to 10 and all the predictors have VIFs between 1-2 – with the highest VIF being 1.3384 for the high school degree, but no college education dummy variable – and none of the VIFs are considerably different from the others, therefore, we conclude that there is no collinearity within our final model.

Now we will move on to assessing potential problem points, particularly influence points and outliers. Outliers are points where the observation does not seem in line with the rest of the model, ie. this point will exhibit a large residual, therefore to identify outliers we will pay particular attention to studentized residuals with magnitudes larger than 2 or 3 – as this will indicate residuals that are 2 or 3 standard deviations away from the rest of the residuals, which have been fitted to a normal model – and PRESS residuals. Influence points are observations that have a large effect on the model, therefore we will be paying particular attention to predicted values that are abnormally large or small, so we will be looking at predicted values, Cook's distance, and PRESS residuals. We will be asking SAS to output a new dataset called influence using an output statement in the same reg procedure we used to model our final multiple linear regression model predicting MMSE using education and age. In this output statement we will ask SAS to populate and create new variables this new dataset influence with columns for predicted values of MMSE, normal residuals for MMSE, studentized residuals for MMSE, and press residuals for MMSE with an id statement set to demogid to make it easier for us to tie observations back to participant ids. The influence dataset now has 663 observations and 28 variables – it originally had 24 variables from the fullmodel dataset, and now has an additional 4 variables for the predicted, residual, studentized residual, and PRESS residuals we asked SAS to calculate.

We will first start off by looking at the distribution of MMSE values and see if there are any particularly large or small observations not in line with the majority of observations. MMSE score is maxed out at 30, therefore, we do not see any particularly large observations as the maximum in this distribution is 30. When we look at the distribution histogram we can see it is heavily skewed left so we would expect most observations to have high values closer to 30 than to 0. We discussed this earlier on in the report, but it is clear that a majority of observations are close to 30 as the median in this distribution is 29, meaning that half of the sample has MMSE between 29-30, while 75% of the sample have MMSE scores above 27 (Q1=27). We do, however, see a number of very small observations not in line with the rest of the observations. Looking at the lowest extreme observations, we have 2 observations with a MMSE of 0 (demogid = 231 and 254), 1 with an MMSE of 11 (demogid = 350), and 2 with an MMSE of 17 (demogid = 499 and 697). This indicates that we do potentially have a number of problem points. Next we will look at the distribution of predicted values of MMSE to identify any potential influence points.

The distribution of predicted MMSE values is slightly less skewed than the distribution of MMSE values. We do not see particularly small predicted values that is very far from the rest of the distribution. The lowest predicted value we see is 23.5043 at demogid's 886, 543, and 534. When we look at the quantiles, there is a much more even spread compared to the quantile spread in the MMSE values distribution, although much of the points are still concentrated at the higher end of MMSE. There are potentially a few influence points here, but the evidence is lacking. We will take a look later at PRESS residuals to see if there is any additional evidence of influence points.

We will now move on to look at studentized residuals to parse out any potential outliers. We see a number of very low studentized residuals much further away from the rest of the distribution of studentized residuals. The lowest extremes were: -10.5614 (demogid=254), -9.8026 (demogid=231), -5.902 (demogid=350), -4.3834 (demogid=499), and -3.9214 (demogid=697). All these extremes are greater in magnitude than 3 therefore all these observations have residuals that are greater than 3 standard deviations away from the majority of residuals if the residuals were fit to a normal curve. We would consider these observations – participants 254, 231, 350, 499, and 697 – to be outliers. These observations are also the same observations we flagged earlier when we were looking for points with MMSE values much different from the majority of the distribution. We can now confirm that those points were in fact outliers, but may not necessarily have been influence points, based on the predicted MMSE distribution we examined earlier. If we examine the distribution of PRESS residuals, we see a similar pattern where the lowest extreme PRESS residuals are the same observations that gave us the lowest extreme studentized residuals. This would usually tells us that these points could be outliers and/or influence points, but since out examination of the predicted values did not flag these same observations (254, 231, 350, 499, and 697), we can conclude that they are simply outliers and not high influence points.

Our preliminary analysis of influence points did not yield compelling results, so now we will assess Cook's Distance to identify high influence points. Looking at Cook's Distance, where we have a general criteria of being over 4/n indicates high influence, there are a handful of observations that would be deemed high influence points, in particular observation 108 (Cook's D=0.233), observation 350 (Cook's D=0.176), and observation 231 (Cook's D=0.456), which stood out by having Cook's Distance much greater than the other observations that were flagged for having a Cook's Distance that met the criteria of being greater than 4/n. Please refer to Table 13 for a full list of observations that were flagged for being outliers and/or influence points based on studentized residuals and Cook's Distance, respectively. There were a total of 29 high influence points identified based on Cook's Distance and a total of 7 outliers as determined by the studentized residuals.

Table 13. Influential Points and Outliers based on Cook's Distance and Studentized Residuals, respectively

Demogid	Studentized Residual	Cook's Distance	Demogid	Studentized Residual	Cook's Distance
20	1.294	0.014	350	**-5.920**	0.176
42	1.490	0.018	448	-2.280	0.025
69	1.442	0.011	452	-2.748	0.010
83	1.490	0.018	460	**-3.515**	0.104
102	1.607	0.013	497	-2.713	0.062
129	1.442	0.011	499	**-4.383**	0.023
134	1.607	0.013	543	1.824	0.040
139	1.442	0.011	616	**-3.717**	0.114
152	1.607	0.013	697	**-3.921**	0.020

186	1.490	0.018	773	-1.314	0.014
231	**-9.803**	0.465	775	1.488	0.011
234	-2.676	0.035	797	1.679	0.014
254	**-10.56**	0.233	856	-2.280	0.025
268	1.442	0.011	900	1.089	0.010
321	1.490	0.018			

* All observations listed are influential points. Bolded studentized residuals represent observations that are outliers and exhibit high influence.

We will also assess whether there is any joint confounding on the relationship between log-transformed hcy and MMSE. Previously, we were asked to run a simple linear regression on this relationship the regression coefficient for log-tranformed hcy was -0.8715 (unadjusted). We were also asked to run a full multiple linear regression model log-transformed hcy and MMSE with additional pedictors of sex, education, age, and pack years of cigarette smoking and found the regression coefficient for log-transformed hcy to be -0.3616 (adjusted). Using these two regression coefficients for log-transformed hcy, we will assess whether there was potential joint confounding using the change in estimates criterion that states that a greater than 10% change in the regression coefficient of log-transformed hcy (or any other estimate) would indicate joint confounding. With an unadjusted slope of -0.8715 and an adjusted slope of -0.3616, this resulted in a 58.51% change, which is greater than 10% therefore confirming that there was joint-confounding by sex, education, age, and pack years of cigarette smoking on the relationship between log-transformed hcy and MMSE. In our final model, we found that adjusted for all these other factors, log-transformed hcy did not remain a significant predictor of MMSE and was ultimately dropped from our final model.

In summary, our full model with all 5 specified predictors was found to be overall significant, however, there were a number of predictors that individually were not significant predictors once we adjusted for the other predictors in the model. Log-transformed hcy, sex, and pack years of cigarette smoking were not found to be significant predictors of MMSE, while education and age were significant predictors. This full model had an adjusted R^2 of 0.1375. Thus, in our final model we removed log-transformed hcy, sex, and pack years of cigarette smoking from our model and so our final model ended up predicting MMSE using education and age and resulted in an adjusted R^2 of 0.1303. In our final model both education and age remained significant and a majority of the levels in both variables were found to be significant against its reference group. Please refer to Tables 10 and 11 for the results of both the full and final models. Similar to many of the other models predicting MMSE we have ran throughout this report, there remain concerns around whether the assumptions of linearity, constant variance, and normality for linear regression are met as a result of the skewness of MMSE and a few of the other predictors, therefore, we should interpret these models with caution. We found a number of problem points with large handful of high influential points (29 in total) and 7 outliers (which were also deemed high influence points) based on studentized residuals and Cook's distance. We did not find any evidence of collinearity in either the full or final model; our VIFs were all low with none of them in either model being larger than 2. We did, however, find joint confounding by the predictors on the relationship between log transformed hcy and

MMSE with a 58.51% change in the slope of log-transformed hcy once the additional predictors were added in and adjusted for in the full model as compared to the simple linear regression predicting MMSE from only log-transformed hcy.

Lastly, we will use the glmselect procedure and LASSO selection method with an AIC-based selection criterion to have SAS output the best model for predicting MMSE using the following predictors: log-transformed hcy, sex, education, age, and pack years of cigarette smoking. SAS selected a model with a total of 7 predictors (the point at which AIC was the lowest at 1629.4591): log-transformed hcy; sex; having 8 or more years of education, but no high school degree, having less than 8 years of education; at least some college; 85-89 age group; and 65-74 age group. This is quite different from our final model where we predicted MMSE with just the education and age group variables having dropped lhcy as it was no longer a significant predictor once adjusted for all other predictors. Interestingly enough, for both our categorical predictors (education and age) SAS chose to keep the reference group levels in the model while not including the last remaining level for each predictor, even though we specified the reference group when we called in the categorical predictors. SAS also chose to keep in the log-transformed hcy predictor even though, again, we originally removed it from our model after not retaining significance after adjusting for all other predictors in our model. These different models have have been the result of a difference in how we went about creating these models. SAS slowly introduced predictors into our model one by one based on a AIC-based criteria, presumably to see how significance changes based on the addition of predictors one by one. We, on the other hand, simply put all our predictors in at one time and then removed non-significant predictors as they were all adjusted for all other predictors. Our way of model selection is less robust and less nuanced than presumably SAS's model was, however, SAS's model is confusing based on the fact that it includes our reference groups for education and age into our model. Perhaps the unmet linear regression assumptions played a role in negatively impacting the SAS-based model selection introducing a form a bias.

Conclusion

Investigators are interested in the relationship between hcy levels and cognitive function measured by MMSE and its implications for the development of Alzheimer's disease in a sample of 900 participants from the Framingham Heart Study, although this sample size was reduced to 663 when we removed participants with missing data on whether they developed Alzheimer's at 7 years follow up. Interestingly, our final model for predicting MMSE does not include log-transformed hcy, which was our primary predictor of interest. When building out our final model we found that in a simple linear regression log-transformed hcy was found to be a significant predictor, however, once we started to adjust for other factors – such as education, age, sex, and cigarette smoking – in our full model (N=594, F=11.50, df=9, 584, p-value < 0.0001, R^2=0.1375), log-transformed hcy was no longer a significant predictor of MMSE. Our final model (N=661, F=17.48. df=6, 654, p-value < 0.0001, R^2=0.1303) predicting MMSE only included the predictors for education and age. A common theme throughout our analysis was how many of the assumptions for linear regression were not met for many of our multiple and simple linear regression models. Very early on when we ran descriptive statistics and distributions for a few variables, particularly MMSE, hcy, log-transformed hcy, and cigarette smoking, we found that these variables were not normally distributed and some were heavily skewed, some more than others. This could have played a role in our not meeting the assumptions necessary to run these regression models. We also found a number of influence points and outliers in our dataset, which did not help with our modeling of this relationship. These factors may have then influenced our results from our SAS model selection as the model SAS recommended us did not match quite well with our final model. Other more advanced modeling or transformation techniques may have been better suited for this analysis.