Drishti Bansal
Dr. MacDonald
DATA 5070

"Analysis of Covariance"

The aim for this paper is to determine whether there are significant *between groups* and *within subject* differences in each of the physiological variables (including Heart Rate, Blood Pressure, measured as systolic, diastolic, and mean arterial pressure) while controlling for potential confounding demographic and/or health variables (Gender, Age, Family History of Alcoholism, Smoking Habits, Drug Addiction, Alcoholism, Drinks Alcohol, Mental Illness, Race/Ethnicity).

Each person was given a form to determine if they have family history of alcoholism, using vape, smoking habits, drug addictions, mental health, alcohol consumption, alcoholism. Those are considered health variables that will be used as covariate for ANCOVA. There are race/ethnicity, gender and age for demographic variables, also will be used as potential covariates. They were also tested for their various measures of blood pressure before consumption of vaping and after.

ANCOVA (Analysis of Covariance) is used to test the effects of categorical variables on continuous dependent variable, controlling for the effects of selected other continuous variables which co-vary with the dependent. Its three purposes are: 1- to increase the sensitivity of the test of main effects and interactions by in turn, reducing the error term ; 2- to adjust the means on the dependent variable themselves on what they would be if all subjects scored equally on the covariates; and 3- the ANCOVA occurs where the researcher assesses on dependent variable after the adjustment of the other dependent variable that are treated as covariates (Textbook pg. 197*).

The dataset consists of 28 variables for a total of 148 participants. For analysis purpose, I deleted the presbp (Pre-Blood Pressure) and postbp (Post-Blood Pressure) variables because they are strings which cannot be used, and there are physiological variables that contain them already. For instance, Pre-Blood Pressure is at 108/88 for the first row, but prebpsys (Pre-Blood Pressure Systolic) is at 108 and a numerical data type that can be used for analysis. There are 75 participants who are vaper and 73 participants who are non-vaper in the grouping variable. There are other health variables as well and they were nominal types.

* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.

Between groups means the total variation between each group mean and their overall mean. If the Between group variation is high relative to the Within-group variation, then the F-statistic of the ANOVA will be higher and the corresponding p-value will be lower, which makes it more likely that we'll reject the null hypothesis that the group means are equal and that's how we will determine which variables will be covariates.

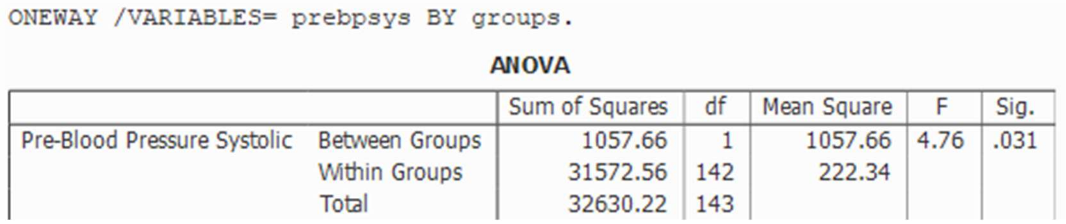
Within groups (repeated measures) means the total variation in the individual values in each group and their group mean.

There were 6, 1, 4, 5, 4, 6, 7 missing values in gender, famalco, smoker, addict, alcohol, mentalil, race respectively. There were four number of rows that have about 5-6 missing values for demographic/health variables, so we can delete those since we are focusing on the difference between the physiological and demographic/health variables.

Row number 16, 34, 108, 130 have been removed from the data. After cleaning, the data has now 26 variables for a total of 144 participants.

1. ***Selecting variables as Covariates***

The assumption is that if the means are different, then the variance within the groups should be small compared to the variance between groups. We look at ANOVA table below for prebpsys BY groups, mean square between groups of 1057.66 is much larger than that of within groups (222.34). The ratio of between groups to within groups is called F statistic and its value is 4.76 in this case. The larger the F is, the more the variability there is between groups than within groups, hence the means are presumed to be equal and there is no effect. However, there is a 0.031 for significant level which is lower than p-value (0.05) which would indicate that it would be quite unlikely to have F this large if there were no real differences among the means. Hence, we conclude that the means are not the same.



```
ONEWAY /VARIABLES= prebpsys BY groups.
```

ANOVA

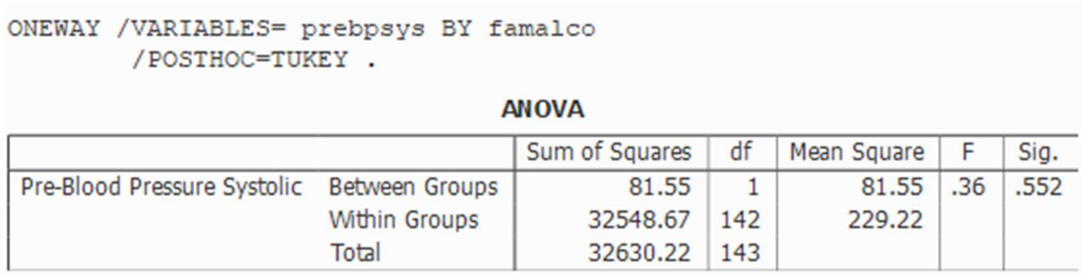
		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	1057.66	1	1057.66	4.76	.031
	Within Groups	31572.56	142	222.34		
	Total	32630.22	143			

Concurrently, for each demographic/health variables, [tables in appendix VII] with higher than p-value of 0.05, age, smoker (Smoking habits), addict (Drug Addiction), alcohol

* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.

(Alcoholism), drinkalc (drinks alcohol), mentalalil (Mental illness), race (Race/Ethnicity) are the variables that cannot be considered as covariates. This would mean that we cannot do any post-hoc analysis and look for differences between any particular pairs of means (Well, we can, but it is meaningless.)

However, there is one health variable, namely Family History of Alcoholism that has a very low F value although the p-value of 0.552 is higher than the level of significance 0.05. This indicates that there is not sufficient evidence to prove that the group means are likely to be equal. In other words, this is useless for ANCOVA.

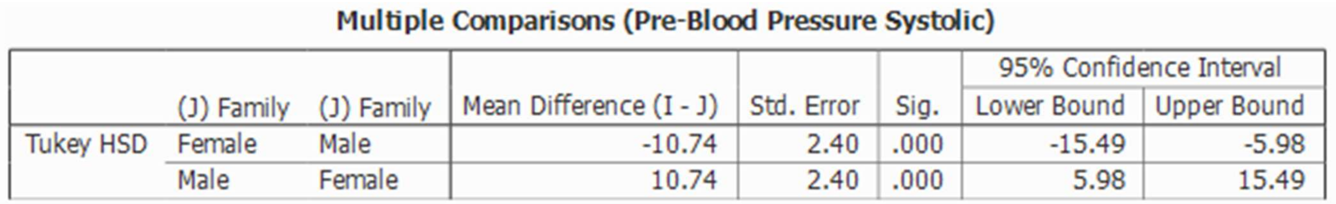


```
ONEWAY /VARIABLES= prebpsys BY famalco
       /POSTHOC=TUKEY .
```

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	81.55	1	81.55	.36	.552
	Within Groups	32548.67	142	229.22		
	Total	32630.22	143			

Consequently, *Gender and Grouping variables* are only variables with below the significance of level of 0.05 and, so become covariates. We can do Tukey post-hoc analysis on those two variables to determine which means are different and the size of the difference. Tukey post-hoc was selected because it is less restrictive while controlling for type one error. We can find ANOVA from Analyze Tools in PSPP:

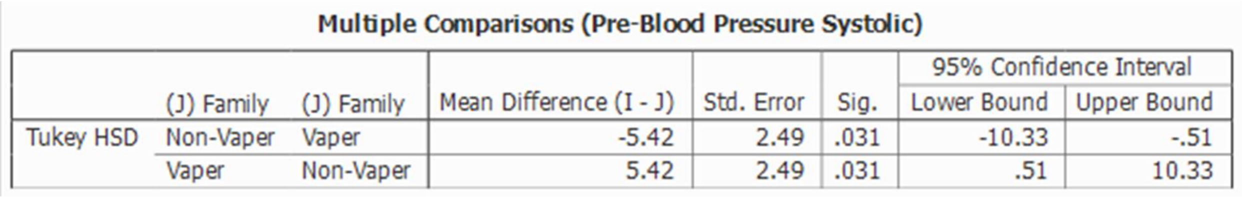


Multiple Comparisons (Pre-Blood Pressure Systolic)

						95% Confidence Interval	
	(J) Family	(J) Family	Mean Difference (I - J)	Std. Error	Sig.	Lower Bound	Upper Bound
Tukey HSD	Female	Male	-10.74	2.40	.000	-15.49	-5.98
	Male	Female	10.74	2.40	.000	5.98	15.49

The table above illustrates that the means of two groups of variables are not equal, but this doesn't mean that they are unequal. If you look at them closely, it is not that unreasonable. Here's why--- there is p-value for groups variable of 0.031 is below 0.05, but not that far from each other. There is almost 2% chance of committing a Type I error in rejecting the fact that the means of groups are not same.

* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.



Multiple Comparisons (Pre-Blood Pressure Systolic)

						95% Confidence Interval	
	(J) Family	(J) Family	Mean Difference (I - J)	Std. Error	Sig.	Lower Bound	Upper Bound
Tukey HSD	Non-Vaper	Vaper	-5.42	2.49	.031	-10.33	-.51
	Vaper	Non-Vaper	5.42	2.49	.031	.51	10.33

The table above shows the same thing as previous table, but for this gender variable, p-value of 0.000 is below 0.05, which indicates that there is really a statistically significant difference.

2. ***<u>Calculating correlations between demographic/health variables and physiological variables.</u>***

By calculating the bivariate correlation, the pre physiological variables are strongly correlated with each other except Pre-Blood Oxygen Saturation who is not correlated with most of pre-physiology variables [Appendix II]. As for the post physiology variables, most of them are correlated with each other, but Post- Respiratory Rate is not correlated with others except Post-Heart Rate, Post-Blood Sugar, and Post-Blood Oxygen Saturation [Appendix III].

Now, we can calculate the bivariate correlation between demographic/health and physiological variables, and we get the result that it is hard to conclude the correlation for all those variables due to too many variables [Appendix I]. Therefore, we can select the variables for covariates by running ANOVAs and we obtained two variables as covariates.

3. ***<u>Assumptions for running ANCOVAs</u>***

Assumptions to be met (Textbook, pg. 203*):

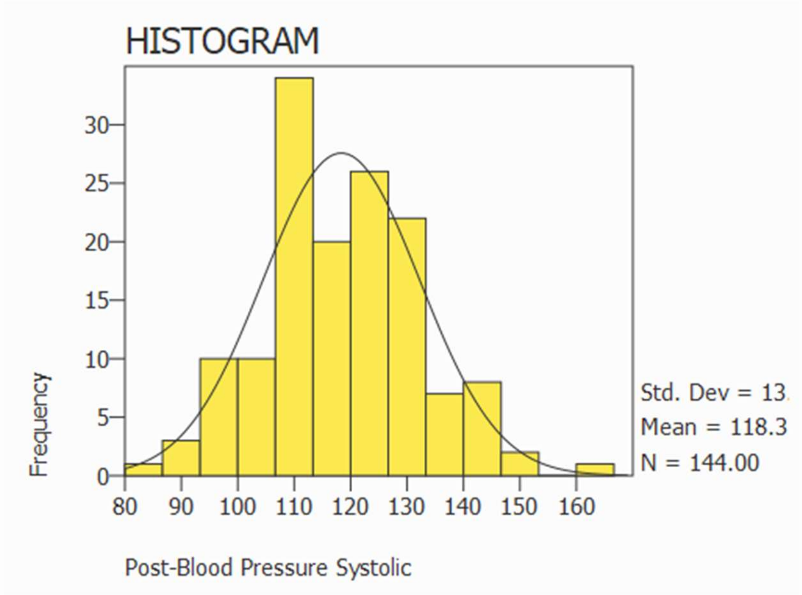
1) Absence of outliers
2) Absence of multicollinearity
3) Normality of sampling distribution
4) Homogeneity of Variance
5) Linearity
6) Homogeneity of Regression

<u>3.1 Making sure if the assumptions are met</u>

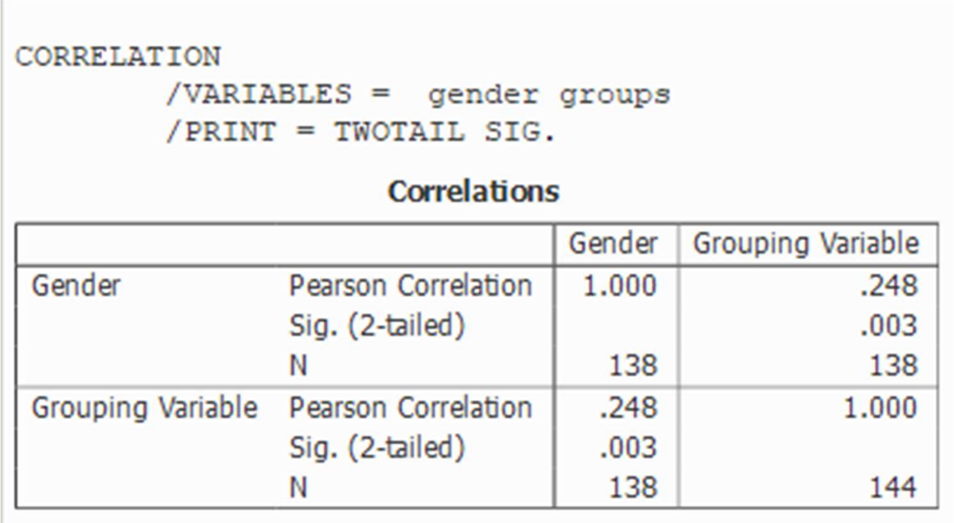
For the first assumption (absence of outliers), there should be normal curve in the graph for the data to be normally distributed. In this case, we can use histogram with frequency which

* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.

tells us the normal curve. We can select any physiological variable, so let's select postbpsy for a histogram and the result is below. We can tell that this graph looks plausible, and we don't see any skewness and kurtosis. Hence, the first assumption is met as the graph looks *reasonably* normally distributed and simultaneously, it also meets the third assumption which is normality of sampling distribution.



For the second assumption, we have two covariates in this analysis, so we can use them to determine if they are highly correlated with each other. We can select Bivariate Correlation in Analyze Tool in PSPP and the result is that they are about 0.25 correlated which is not that high with a 0.003 significance of level which is below the p-value of 0.05. It means that the second assumption is met.



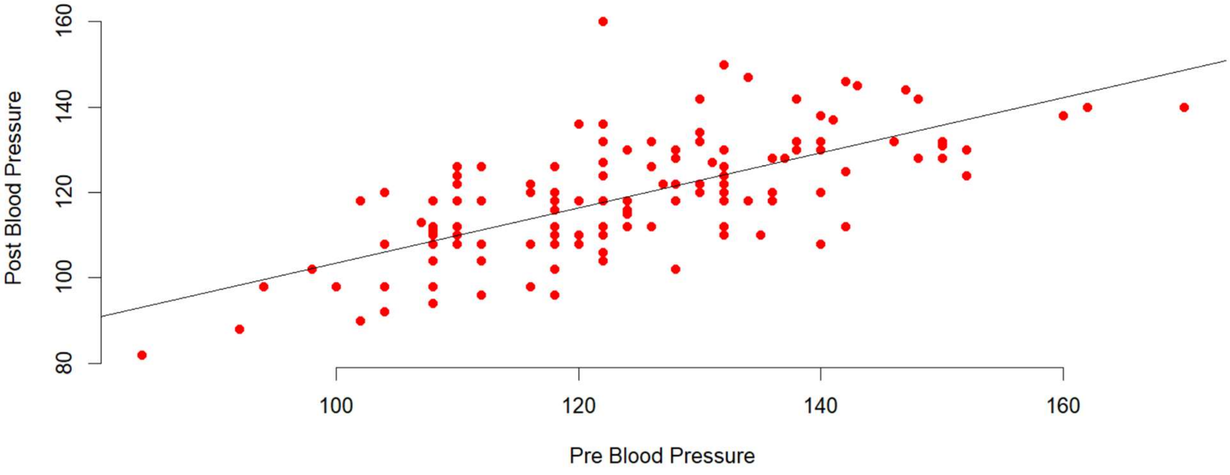
```
CORRELATION
        /VARIABLES =  gender groups
        /PRINT = TWOTAIL SIG.
```

Correlations

		Gender	Grouping Variable
Gender	Pearson Correlation	1.000	.248
	Sig. (2-tailed)		.003
	N	138	138
Grouping Variable	Pearson Correlation	.248	1.000
	Sig. (2-tailed)	.003	
	N	138	144

* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.

For the fifth assumption, we can use pre- physiological variables and groups as a covariate by computing regression line in Rstudio and we get the output from running the computation of regression line:

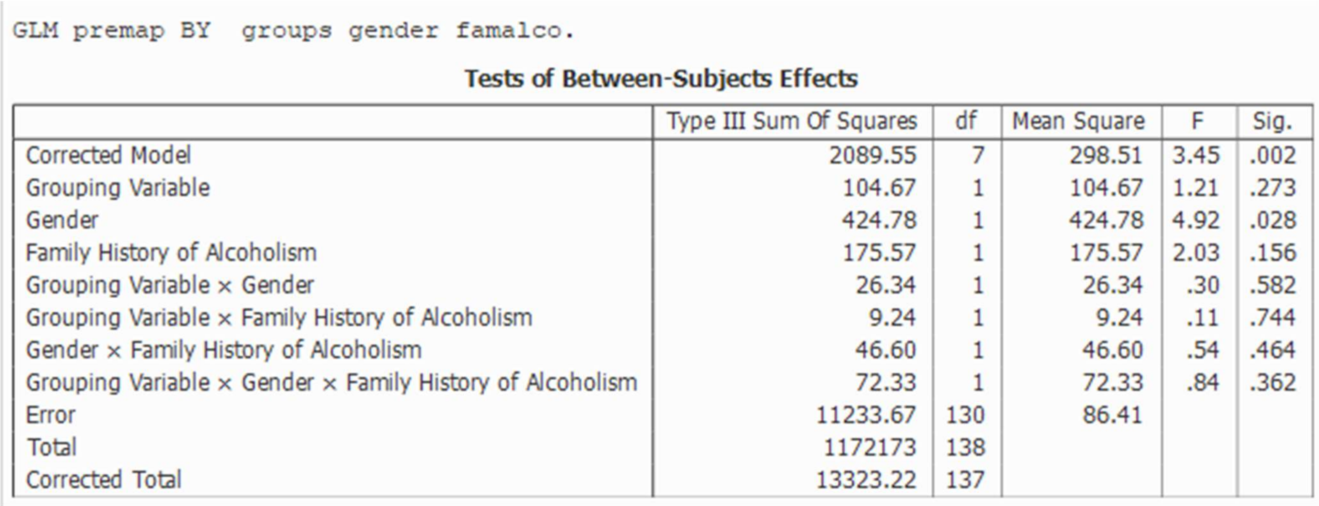


This seems to be meeting the fifth assumption and since the R square [V] of this is about 0.50 which seems to meet the assumptions of homogeneity of regression slopes and linearity.

4. __Completing a total of mixed factorial ANCOVAs__

We are using groups as the between groups independent variable and pre-to-post variable measurements as within the groups to complete a total mixed analysis of covariance. We have average of pre and post physiological variables (Pre-Test Mean Arterial Pressure, and Post-Test Mean Arterial Pressure, respectively), so we can use that as a repeated measure variable instead of running an analysis on each pre and post physiological variable. We researchers want to evaluate the effects of grouping variable on pre-to-post physiological variables after adjusting the covariate variable which is gender in this case.

* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.

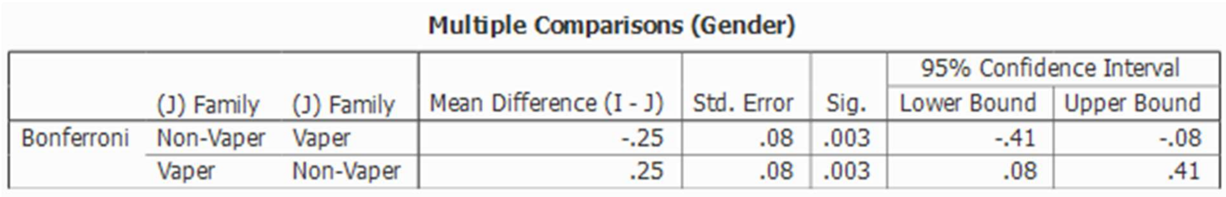


```
GLM premap BY  groups gender famalco.
```

Tests of Between-Subjects Effects

	Type III Sum Of Squares	df	Mean Square	F	Sig.
Corrected Model	2089.55	7	298.51	3.45	.002
Grouping Variable	104.67	1	104.67	1.21	.273
Gender	424.78	1	424.78	4.92	.028
Family History of Alcoholism	175.57	1	175.57	2.03	.156
Grouping Variable × Gender	26.34	1	26.34	.30	.582
Grouping Variable × Family History of Alcoholism	9.24	1	9.24	.11	.744
Gender × Family History of Alcoholism	46.60	1	46.60	.54	.464
Grouping Variable × Gender × Family History of Alcoholism	72.33	1	72.33	.84	.362
Error	11233.67	130	86.41		
Total	1172173	138			
Corrected Total	13323.22	137			

The table above shows that there was a significant difference (F[7,130] = 4.92, p = 0.028) in gender variable. If main ANOVA is significant, post hoc test using Bonferroni correction is carried out to see which groups differ.



Multiple Comparisons (Gender)

						95% Confidence Interval	
	(J) Family	(J) Family	Mean Difference (I - J)	Std. Error	Sig.	Lower Bound	Upper Bound
Bonferroni	Non-Vaper	Vaper	-.25	.08	.003	-.41	-.08
	Vaper	Non-Vaper	.25	.08	.003	.08	.41

There is not a significant difference between non vaper and vaper (p = 0.003, both).

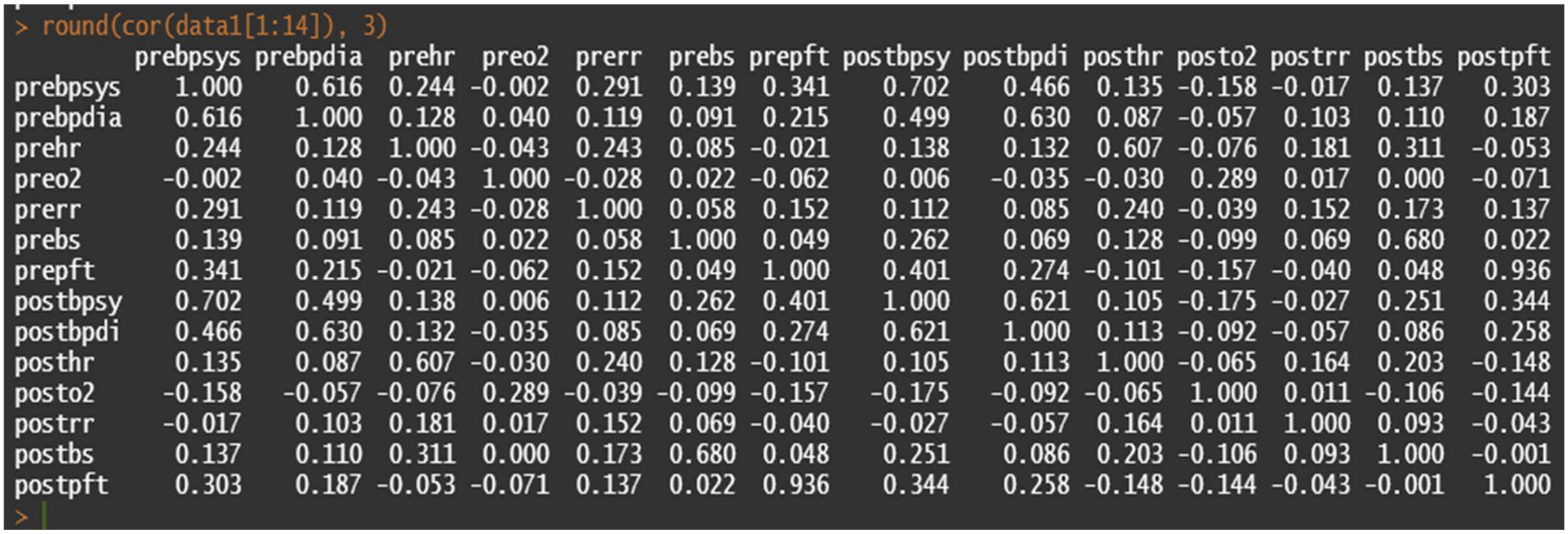
In conclusion, ANCOVA was conducted to compare the effectiveness of grouping variable while controlling for the covariate. There were tests that were carried out and assumptions met.

(PS I am struggling to understand how the mixed factorial ancova is conducted in PSPP and thereby, failing to fulfil the objective for this assignment. I would need help with that)

* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.

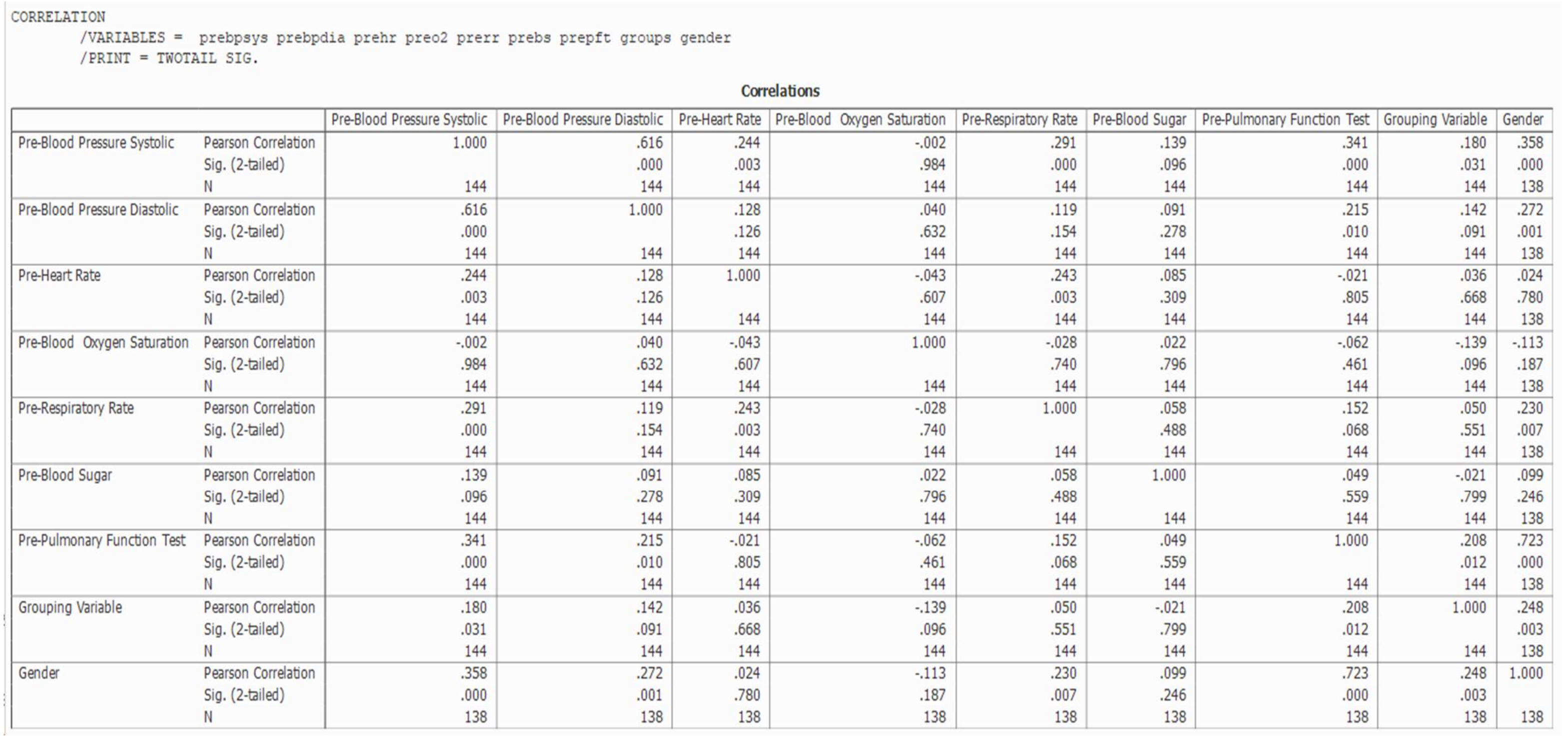
Appendix:

I. Correlation between pre and post physiological variables.



```
> round(cor(data1[1:14]), 3)
          prebpsys prebpdia prehr  preo2  prerr  prebs prepft postbpsy postbpdi posthr posto2 postrr postbs postpft
prebpsys     1.000    0.616 0.244 -0.002  0.291  0.139  0.341    0.702    0.466  0.135 -0.158 -0.017  0.137   0.303
prebpdia     0.616    1.000 0.128  0.040  0.119  0.091  0.215    0.499    0.630  0.087 -0.057  0.103  0.110   0.187
prehr        0.244    0.128 1.000 -0.043  0.243  0.085 -0.021    0.138    0.132  0.607 -0.076  0.181  0.311  -0.053
preo2       -0.002    0.040 -0.043 1.000 -0.028  0.022 -0.062    0.006   -0.035 -0.030  0.289  0.017  0.000  -0.071
prerr        0.291    0.119 0.243 -0.028  1.000  0.058  0.152    0.112    0.085  0.240 -0.039  0.152  0.173   0.137
prebs        0.139    0.091 0.085  0.022  0.058  1.000  0.049    0.262    0.069  0.128 -0.099  0.069  0.680   0.022
prepft       0.341    0.215 -0.021 -0.062 0.152  0.049  1.000    0.401    0.274 -0.101 -0.157 -0.040  0.048   0.936
postbpsy     0.702    0.499 0.138  0.006  0.112  0.262  0.401    1.000    0.621  0.105 -0.175 -0.027  0.251   0.344
postbpdi     0.466    0.630 0.132 -0.035  0.085  0.069  0.274    0.621    1.000  0.113 -0.092 -0.057  0.086   0.258
posthr       0.135    0.087 0.607 -0.030  0.240  0.128 -0.101    0.105    0.113  1.000 -0.065  0.164  0.203  -0.148
posto2      -0.158   -0.057 -0.076  0.289 -0.039 -0.099 -0.157   -0.175   -0.092 -0.065  1.000  0.011 -0.106  -0.144
postrr      -0.017    0.103 0.181  0.017  0.152  0.069 -0.040   -0.027   -0.057  0.164  0.011  1.000  0.093  -0.043
postbs       0.137    0.110 0.311  0.000  0.173  0.680  0.048    0.251    0.086  0.203 -0.106  0.093  1.000  -0.001
postpft      0.303    0.187 -0.053 -0.071 0.137  0.022  0.936    0.344    0.258 -0.148 -0.144 -0.043 -0.001   1.000
>
```

II. Correlation between pre-physiology variables



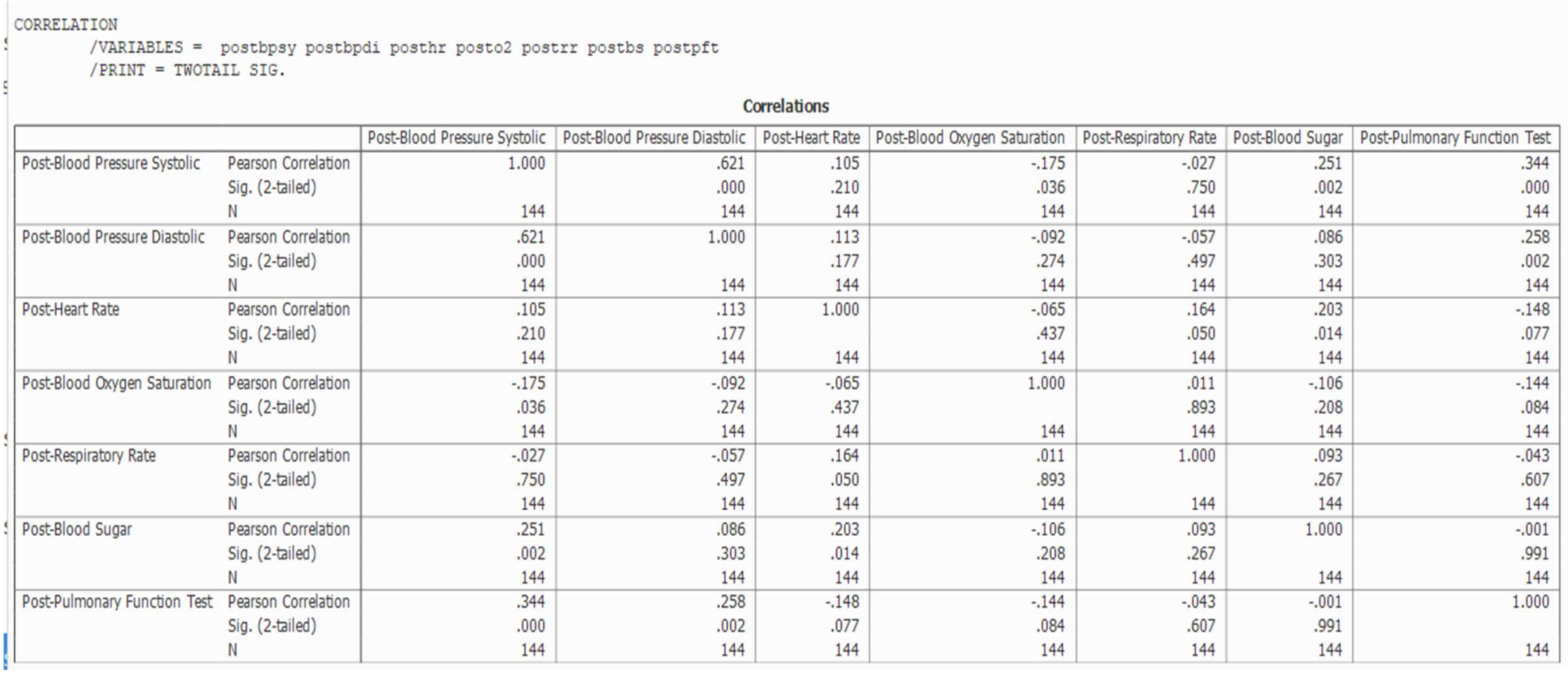
```
CORRELATION
       /VARIABLES =  prebpsys prebpdia prehr preo2 prerr prebs prepft groups gender
       /PRINT = TWOTAIL SIG.
```

Correlations

		Pre-Blood Pressure Systolic	Pre-Blood Pressure Diastolic	Pre-Heart Rate	Pre-Blood Oxygen Saturation	Pre-Respiratory Rate	Pre-Blood Sugar	Pre-Pulmonary Function Test	Grouping Variable	Gender
Pre-Blood Pressure Systolic	Pearson Correlation	1.000	.616	.244	-.002	.291	.139	.341	.180	.358
	Sig. (2-tailed)		.000	.003	.984	.000	.096	.000	.031	.000
	N	144	144	144	144	144	144	144	144	138
Pre-Blood Pressure Diastolic	Pearson Correlation	.616	1.000	.128	.040	.119	.091	.215	.142	.272
	Sig. (2-tailed)	.000		.126	.632	.154	.278	.010	.091	.001
	N	144	144	144	144	144	144	144	144	138
Pre-Heart Rate	Pearson Correlation	.244	.128	1.000	-.043	.243	.085	-.021	.036	.024
	Sig. (2-tailed)	.003	.126		.607	.003	.309	.805	.668	.780
	N	144	144	144	144	144	144	144	144	138
Pre-Blood Oxygen Saturation	Pearson Correlation	-.002	.040	-.043	1.000	-.028	.022	-.062	-.139	-.113
	Sig. (2-tailed)	.984	.632	.607		.740	.796	.461	.096	.187
	N	144	144	144	144	144	144	144	144	138
Pre-Respiratory Rate	Pearson Correlation	.291	.119	.243	-.028	1.000	.058	.152	.050	.230
	Sig. (2-tailed)	.000	.154	.003	.740		.488	.068	.551	.007
	N	144	144	144	144	144	144	144	144	138
Pre-Blood Sugar	Pearson Correlation	.139	.091	.085	.022	.058	1.000	.049	-.021	.099
	Sig. (2-tailed)	.096	.278	.309	.796	.488		.559	.799	.246
	N	144	144	144	144	144	144	144	144	138
Pre-Pulmonary Function Test	Pearson Correlation	.341	.215	-.021	-.062	.152	.049	1.000	.208	.723
	Sig. (2-tailed)	.000	.010	.805	.461	.068	.559		.012	.000
	N	144	144	144	144	144	144	144	144	138
Grouping Variable	Pearson Correlation	.180	.142	.036	-.139	.050	-.021	.208	1.000	.248
	Sig. (2-tailed)	.031	.091	.668	.096	.551	.799	.012		.003
	N	144	144	144	144	144	144	144	144	138
Gender	Pearson Correlation	.358	.272	.024	-.113	.230	.099	.723	.248	1.000
	Sig. (2-tailed)	.000	.001	.780	.187	.007	.246	.000	.003	
	N	138	138	138	138	138	138	138	138	138

* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.

III. Correlation between post-physiology variables

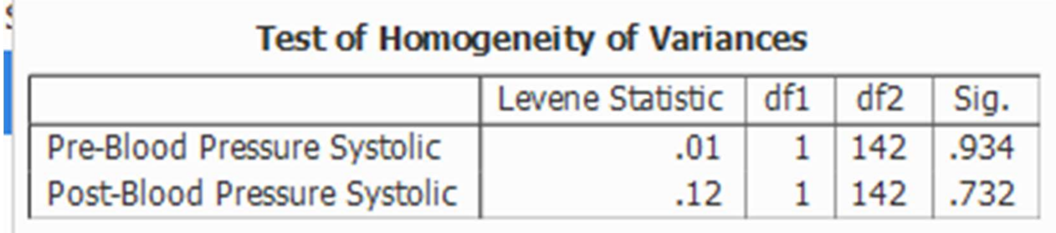


```
CORRELATION
        /VARIABLES =  postbpsy postbpdi posthr posto2 postrr postbs postpft
        /PRINT = TWOTAIL SIG.
```

Correlations

		Post-Blood Pressure Systolic	Post-Blood Pressure Diastolic	Post-Heart Rate	Post-Blood Oxygen Saturation	Post-Respiratory Rate	Post-Blood Sugar	Post-Pulmonary Function Test
Post-Blood Pressure Systolic	Pearson Correlation	1.000	.621	.105	-.175	-.027	.251	.344
	Sig. (2-tailed)		.000	.210	.036	.750	.002	.000
	N	144	144	144	144	144	144	144
Post-Blood Pressure Diastolic	Pearson Correlation	.621	1.000	.113	-.092	-.057	.086	.258
	Sig. (2-tailed)	.000		.177	.274	.497	.303	.002
	N	144	144	144	144	144	144	144
Post-Heart Rate	Pearson Correlation	.105	.113	1.000	-.065	.164	.203	-.148
	Sig. (2-tailed)	.210	.177		.437	.050	.014	.077
	N	144	144	144	144	144	144	144
Post-Blood Oxygen Saturation	Pearson Correlation	-.175	-.092	-.065	1.000	.011	-.106	-.144
	Sig. (2-tailed)	.036	.274	.437		.893	.208	.084
	N	144	144	144	144	144	144	144
Post-Respiratory Rate	Pearson Correlation	-.027	-.057	.164	.011	1.000	.093	-.043
	Sig. (2-tailed)	.750	.497	.050	.893		.267	.607
	N	144	144	144	144	144	144	144
Post-Blood Sugar	Pearson Correlation	.251	.086	.203	-.106	.093	1.000	-.001
	Sig. (2-tailed)	.002	.303	.014	.208	.267		.991
	N	144	144	144	144	144	144	144
Post-Pulmonary Function Test	Pearson Correlation	.344	.258	-.148	-.144	-.043	-.001	1.000
	Sig. (2-tailed)	.000	.002	.077	.084	.607	.991	
	N	144	144	144	144	144	144	144

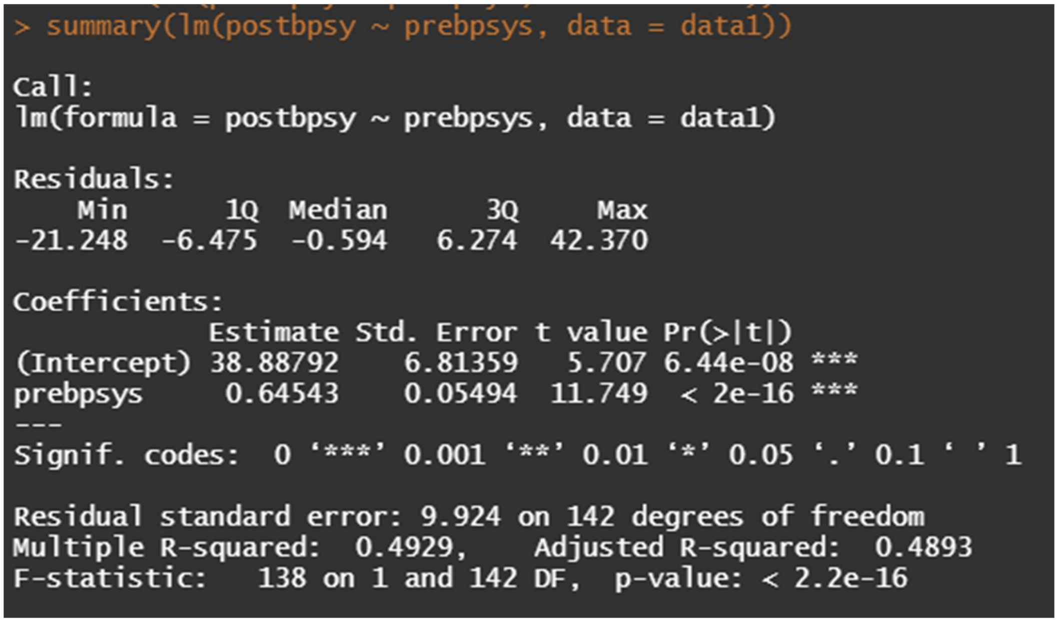
IV. Test of Homogeneity of Variance



Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
Pre-Blood Pressure Systolic	.01	1	142	.934
Post-Blood Pressure Systolic	.12	1	142	.732

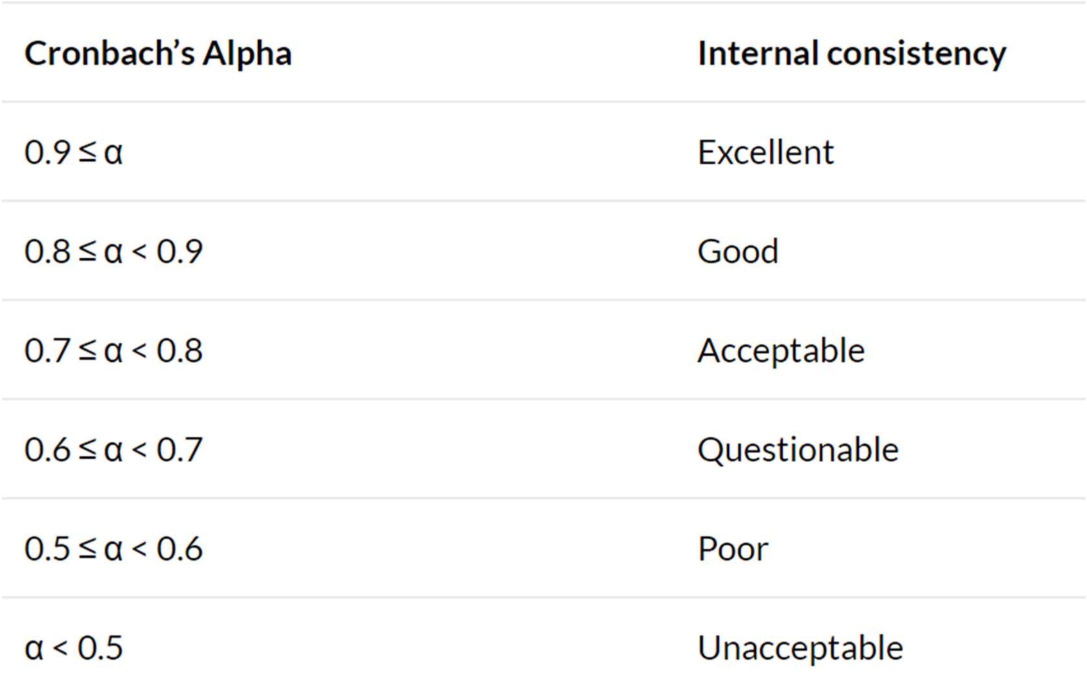
V. Summary of regression



```
> summary(lm(postbpsy ~ prebpsys, data = data1))

Call:
lm(formula = postbpsy ~ prebpsys, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.248  -6.475  -0.594   6.274  42.370 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 38.88792    6.81359   5.707 6.44e-08 ***
prebpsys     0.64543    0.05494  11.749  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.924 on 142 degrees of freedom
Multiple R-squared:  0.4929,	Adjusted R-squared:  0.4893 
F-statistic:   138 on 1 and 142 DF,  p-value: < 2.2e-16
```

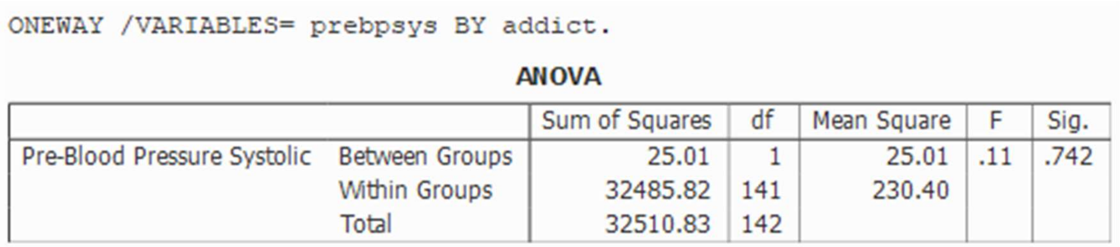
* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.

VI. CronBach’s Alpha table



Cronbach’s Alpha	Internal consistency
0.9 ≤ α	Excellent
0.8 ≤ α < 0.9	Good
0.7 ≤ α < 0.8	Acceptable
0.6 ≤ α < 0.7	Questionable
0.5 ≤ α < 0.6	Poor
α < 0.5	Unacceptable

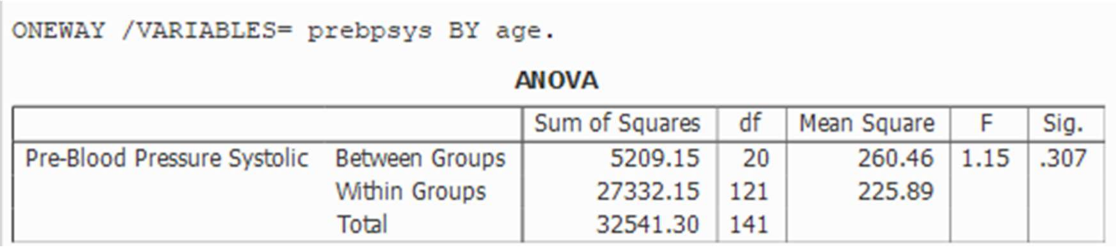
VII. ANOVA for a pre-physiological variable by each demographic and health variables to determine which variables are considered as covariates.



```
ONEWAY /VARIABLES= prebpsys BY addict.
```

ANOVA

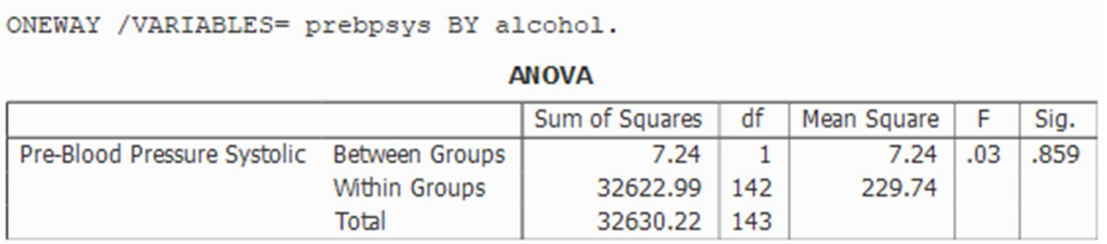
		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	25.01	1	25.01	.11	.742
	Within Groups	32485.82	141	230.40		
	Total	32510.83	142			



```
ONEWAY /VARIABLES= prebpsys BY age.
```

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	5209.15	20	260.46	1.15	.307
	Within Groups	27332.15	121	225.89		
	Total	32541.30	141			

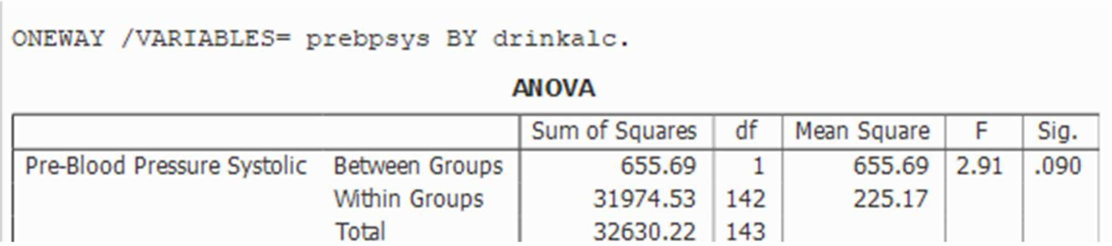


```
ONEWAY /VARIABLES= prebpsys BY alcohol.
```

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	7.24	1	7.24	.03	.859
	Within Groups	32622.99	142	229.74		
	Total	32630.22	143			

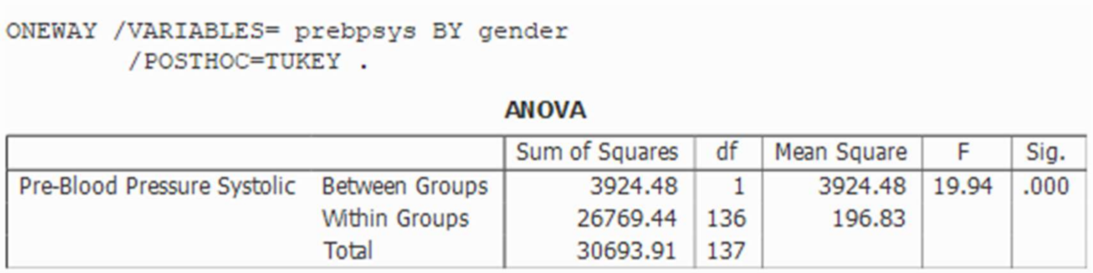
* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.



```
ONEWAY /VARIABLES= prebpsys BY drinkalc.
```

ANOVA

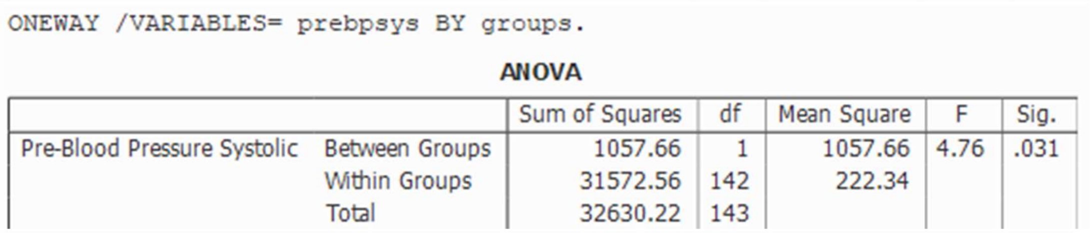
		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	655.69	1	655.69	2.91	.090
	Within Groups	31974.53	142	225.17		
	Total	32630.22	143			



```
ONEWAY /VARIABLES= prebpsys BY gender
       /POSTHOC=TUKEY .
```

ANOVA

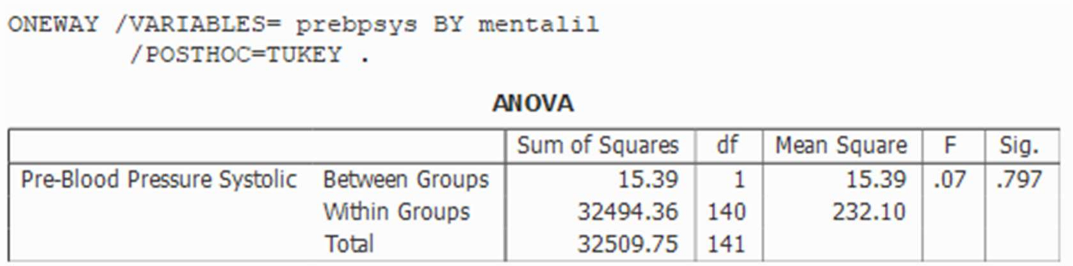
		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	3924.48	1	3924.48	19.94	.000
	Within Groups	26769.44	136	196.83		
	Total	30693.91	137			



```
ONEWAY /VARIABLES= prebpsys BY groups.
```

ANOVA

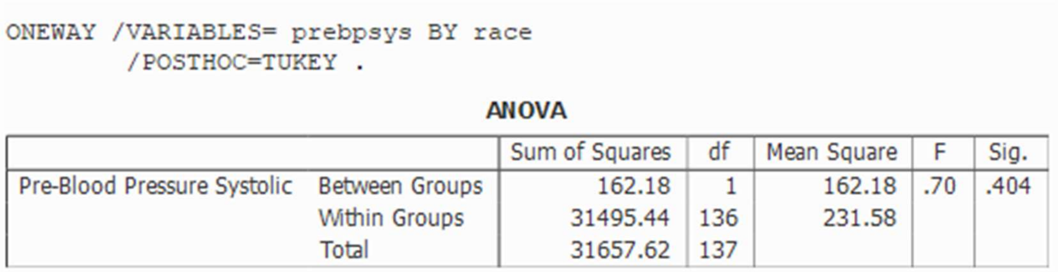
		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	1057.66	1	1057.66	4.76	.031
	Within Groups	31572.56	142	222.34		
	Total	32630.22	143			



```
ONEWAY /VARIABLES= prebpsys BY mentalil
       /POSTHOC=TUKEY .
```

ANOVA

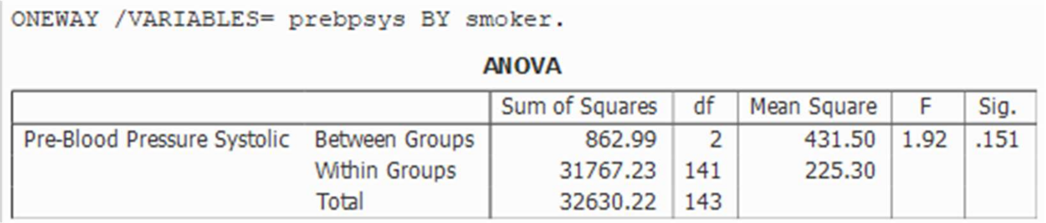
		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	15.39	1	15.39	.07	.797
	Within Groups	32494.36	140	232.10		
	Total	32509.75	141			



```
ONEWAY /VARIABLES= prebpsys BY race
       /POSTHOC=TUKEY .
```

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	162.18	1	162.18	.70	.404
	Within Groups	31495.44	136	231.58		
	Total	31657.62	137			



```
ONEWAY /VARIABLES= prebpsys BY smoker.
```

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Pre-Blood Pressure Systolic	Between Groups	862.99	2	431.50	1.92	.151
	Within Groups	31767.23	141	225.30		
	Total	32630.22	143			

* Refers to the chapter 6- Analysis of Covariance provided by Dr. MacDonald on Blackboard.