Scientific Methods for Health Sciences: Fundamentals (HS550): Fall 2014
http://www.socr.umich.edu/people/dinov/2014/Fall/HS550/
Homework 3[1] Solutions

Problem 1

Do a two-sample test between the MMSE scores of the two groups of patients defined by Group0 and Group 1.

$$t_0 = \frac{\widehat{\mu_3} - \widehat{\mu_4}}{SE(\widehat{\mu_3} - \widehat{\mu_4})} = \frac{\hat{\mu} - \widehat{\mu_4}}{\sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}} = \frac{27.48069 - 26.6625}{\sqrt{\frac{2.539484^2}{48} + \frac{2.609656^2}{240}}} = 3.4558$$

$$df = \frac{\left(\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}\right)^2}{\frac{\left(\frac{s_0^2}{n_0}\right)^2}{n_0 - 1} + \frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1}} = 470.9973$$

The test statistic is 3.4558 and generates a p-value of 0.0005983. We have enough evidence to reject the null hypothesis of $H_0: \mu_0 = \mu_1$ at 5% level of significance, and claim that there are significant difference between the MMSE scores of the two groups of patients of group 0 and group 1 .

T-test result (using R-script below):

```
Welch Two Sample t-test

data:  g0$MMSCORE and g1$MMSCORE
t = 3.4558, df = 470.997, p-value = 0.0005983
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3529543 1.2834191
sample estimates:
mean of x mean of y
 27.48069  26.66250
```

RCODE:

```
# problem 1
# Save the data (http://wiki.socr.umich.edu/index.php/SOCR_Data_AD_BiomedBigMetadata) in a
local file /data_folder/Homework3.csv or C:\data_folder\Homework3.csv
biom <- read.csv('C:\\data_folder\\Homework3.csv'')
summary(biom)
attach(biom)
g0 <- subset(biom,GDTOTAL==0)
mu0 <- mean(g0$MMSCORE)
n0 <- dim(g0)[1]
s0 <- sd(g0$MMSCORE)
g1 <- subset(biom,GDTOTAL==1)
mu1 <- mean(g1$MMSCORE)
n1 <- dim(g1)[1]
```

[1] http://www.socr.umich.edu/people/dinov/2014/Fall/HS550/HWs.html

```
s1 <- sd(g1$MMSCORE)
df <- (s0^2/n0+s1^2/n1)^2/((s0^2/n0)^2/(n0-1)+(s1^2/n1)^2/(n1-1))
se <- sqrt(s0^2/n0+s1^2/n1)
t <- (mu0-mu1)/se  ## 3.455795
p <- 2*pt(-abs(t),df=df)  ## 0.0005983385
## Or use the t.test function
t.test(g0$MMSCORE,g1$MMSCORE)
```

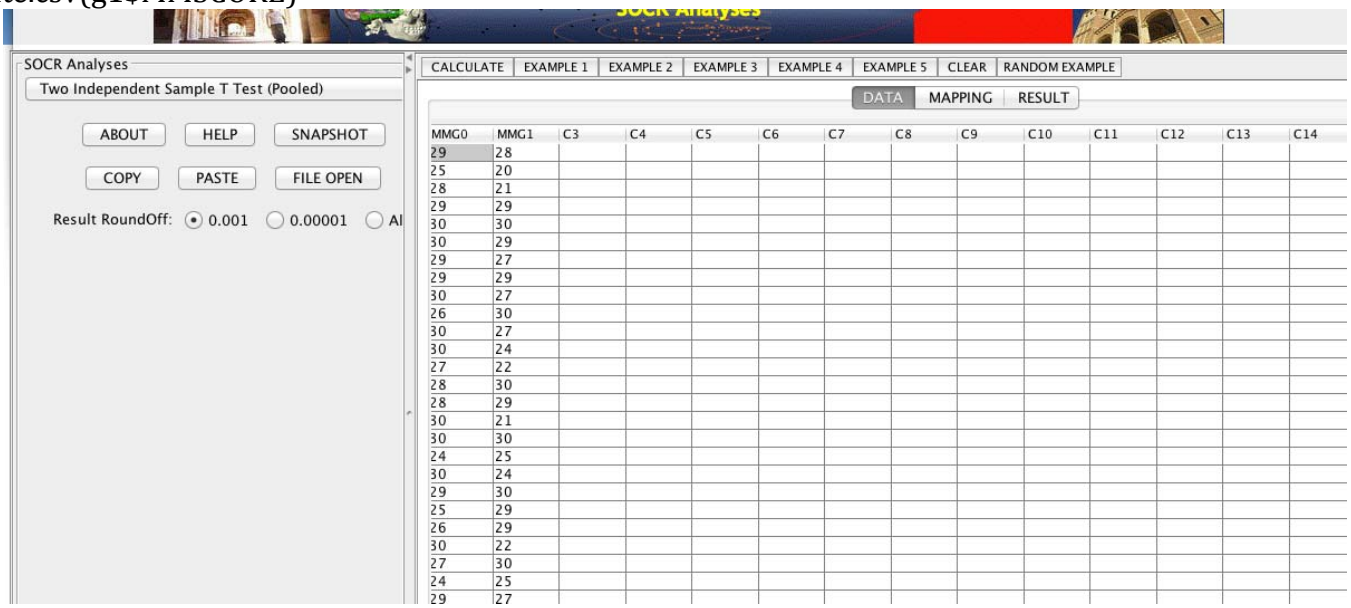
```
Welch Two Sample t-test

data:  g0$MMSCORE and g1$MMSCORE
t = 3.4558, df = 470.997, p-value = 0.0005983
```

Using SOCR Two independent sample t-test (pooled)

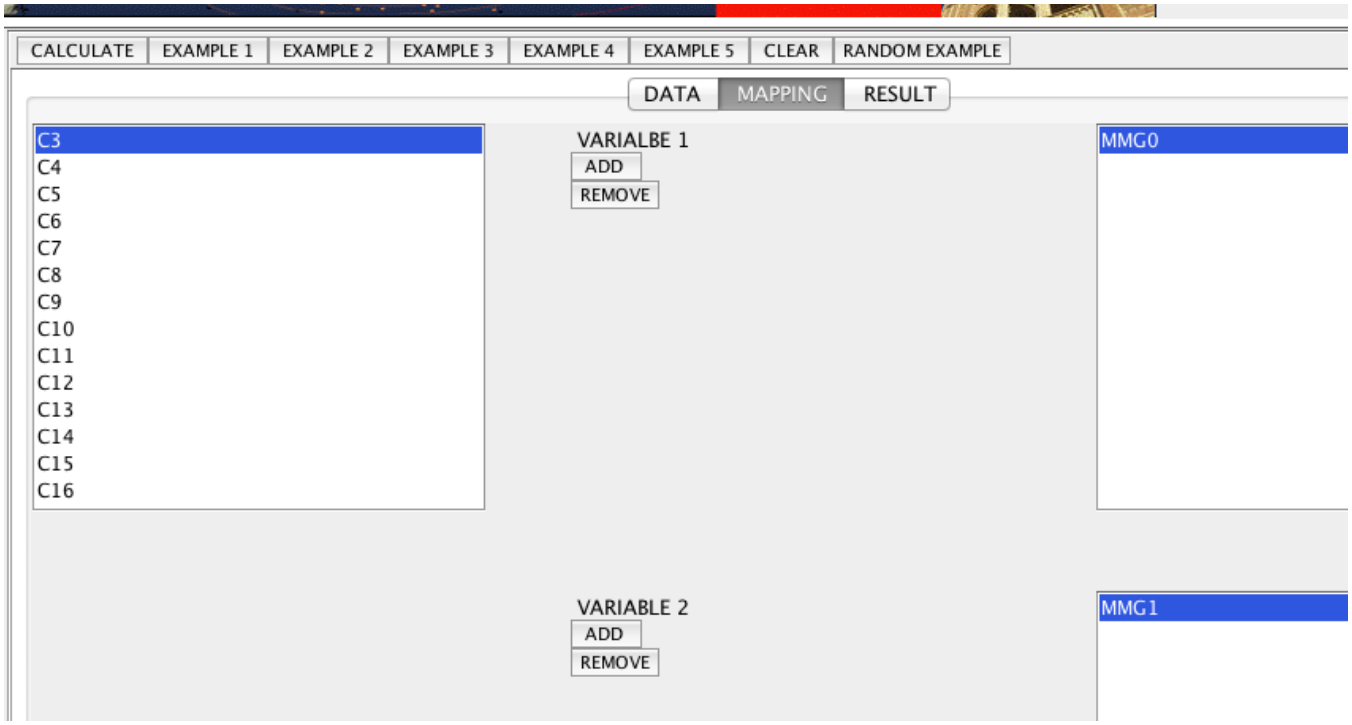
http://socr.ucla.edu/htmls/SOCR_Analyses.html

Step 1: Input the data of MMSCORES for Group 0 and Group 1, the data can be generated in R:

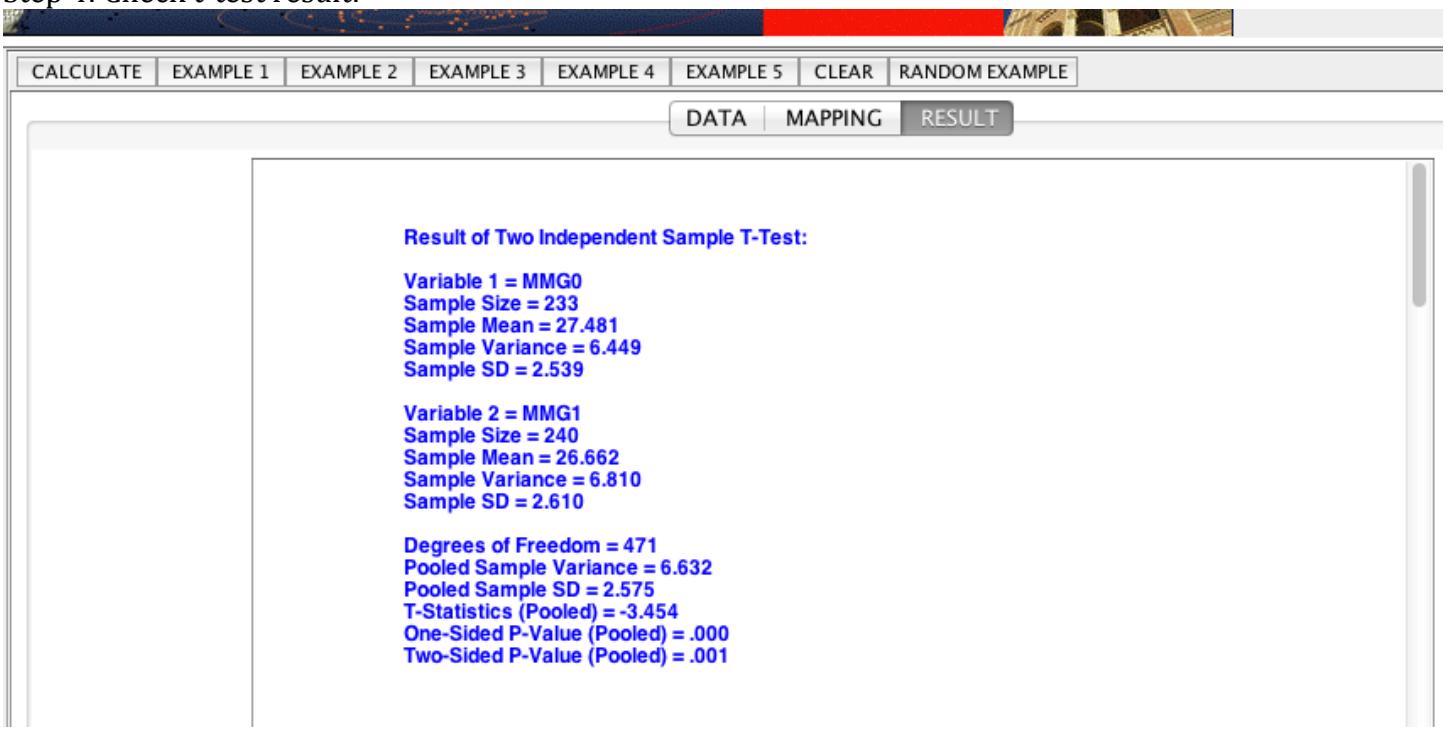
```
## SOCR data
write.csv(g0$MMSCORE)
write.csv(g1$MMSCORE)
```



Step 2: Mapping



Step 3: Calculate
Step 4: Check t-test result:



The result is very close to the one we got in R and the conclusion is also consistent, we reject the null hypothesis of no significant difference in the MMSCORE score in group 0 and group 1 at 5% level of significance and claim that the MMSCORE scores in group 0 and group 1 differ significantly.

Problem 2

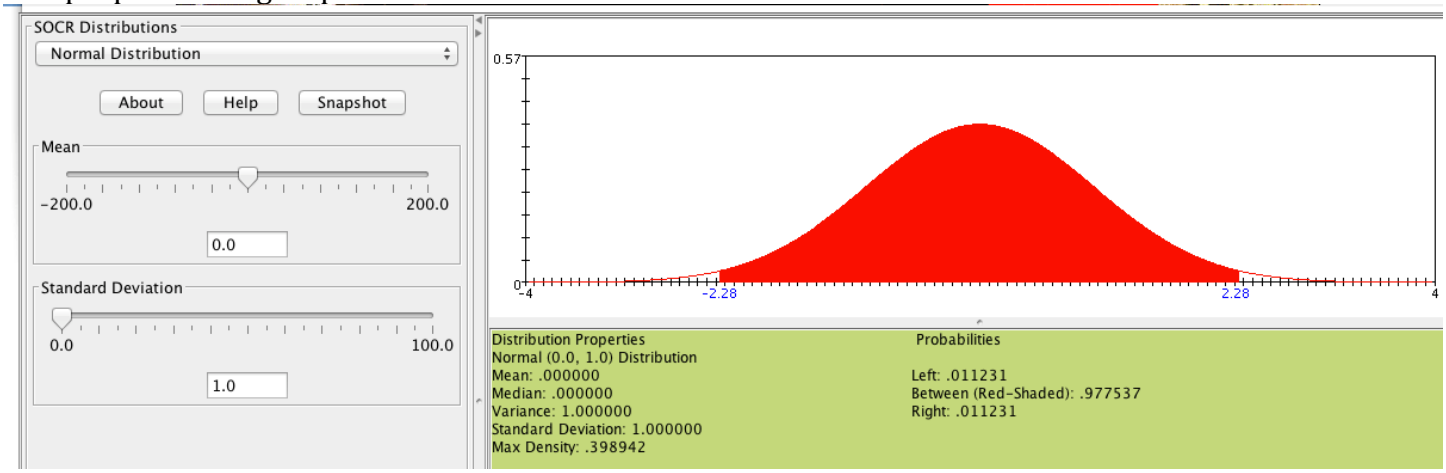
Do a test on the proportions of patients with {GDTOTAL >0} in two groups Group3 (CDGLOBAL=1) vs. Group4 (CDGLOBAL=0). Null hypothesis $P_3 = P_4$. Test statistics (t-test) for proportions in several groups without continuity correction (see Yates' continuity correction for details):

$$Z_0 = \frac{\widehat{p_3} - \widehat{p_4}}{SE(\widehat{p_3} - \widehat{p_4})} = \frac{\widehat{p_3} - \widehat{p_4}}{\sqrt{\frac{\widehat{p_3}(1-\widehat{p_3})}{n_3} + \frac{\widehat{p_4}(1-\widehat{p_4})}{n_4}}}$$

$$= \frac{\frac{39}{48} - \frac{471}{695}}{\sqrt{\frac{0.8125 * (1-0.8125)}{48} + \frac{0.6776978 * (1-0.6776978)}{695}}} = 2.282454$$

$$Z \sim N(0,1)$$

p value = 0.02246257, we have enough evidence to reject the null hypothesis of equal proportion at 5% level of significance and claim that there are significant difference between the two proportions, that is the proportion of patients with GDTOTAL >0 in group3 where CDGLOBAL=1 differs significantly from that proportion in group 4 where CDGLOBAL=0.

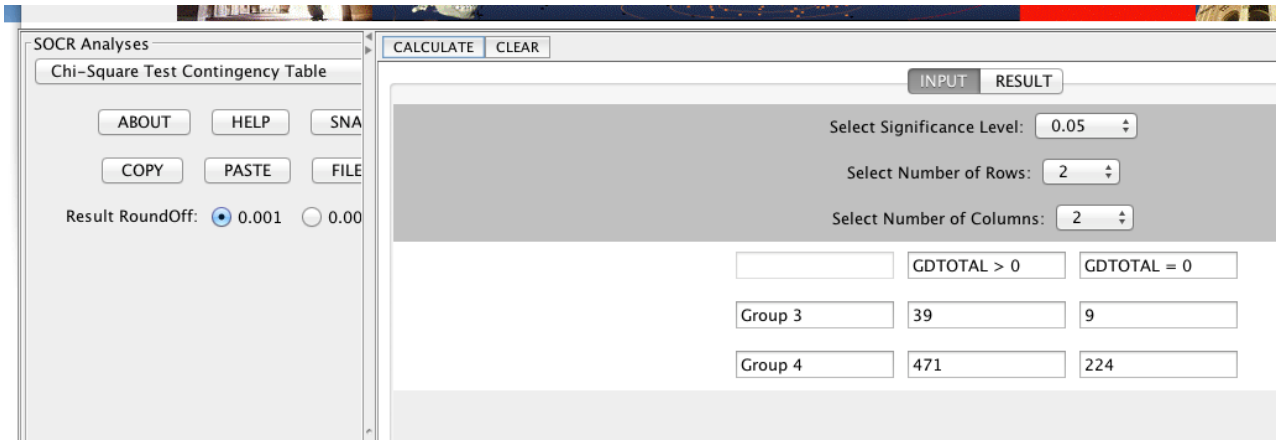


p value can also be observed from this, . $value = P(Z > Z_{0.025}) + P(Z < Z_{0.975}) = 0.11231 + 0.11231 \approx 0.22462$.

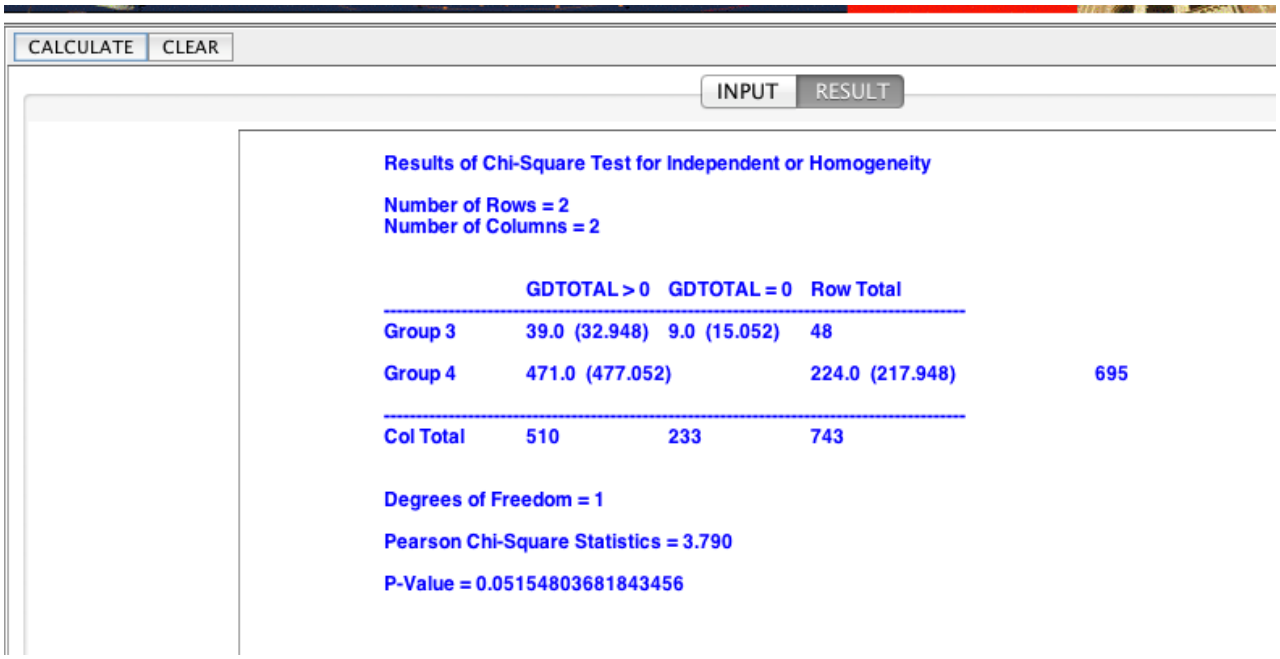
RCODE:

```
## problem 2
g3 <- subset(biom,CDGLOBAL==1)
n3 <- dim(g3)[1]
p3 <- sum(as.numeric(g3$GDTOTAL>0))/n3
g4 <- subset(biom,CDGLOBAL==0)
n4 <- dim(g4)[1]
p4 <- sum(as.numeric(g4$GDTOTAL>0))/n4
n4 <- dim(g4)[1]
se2 <- sqrt(p3*(1-p3)/n3+p4*(1-p4)/n4)
z2 <- (p3-p4)/se2
p2 <- 2*(1-pnorm(z2,0,1))
```

Using SOCR: Chi-Square Test Contingency Table: http://socr.ucla.edu/htmls/SOCR_Analyses.html
Step 1: Input the contingency table



Step 2: Calculate
Step 3: Check the result:



The chi-square test has test statistics = 3.790, which is slightly smaller compared to the threshold of X^2=3.84. We don't have enough evidence to reject the null hypothesis of no significant difference in the proportions of GDTOTAL > 0 between group 3 and group 4 at 5% level of significance. The result is slightly different compared to the result concluded in R.

Problem 3

For the MCI-to-AD Converters (DX_Conversion) variable, the summary of the dataset suggests that there 1 missing value, 735 records with DX_Conversion=0 (No conversion or Reversion to NL/MCI), 7 records with DX_Conversion=1, that is has Conversion to NL/MCI, and 1 record with DX_Conversion =2, that is has Reversion to NL/MCI. To compare their performances in MMSCORE scores, I choose to compare two groups of Group with no conversion or Reversion to NL/MCI and group with only Conversion as well as

group with no conversion or reversion and group with either Conversion or Reversion and the t test result on H_0: $\mu_0 = \mu_1$ and another t test H_0: $\mu_0 = \mu_{12}$ are as following:

a. No vs. Conversion only

```
Welch Two Sample t-test

data:  DX0$MMSCORE and DX1$MMSCORE
t = 0.5758, df = 6.278, p-value = 0.5848
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.211888  1.968351
sample estimates:
mean of x mean of y
 26.80680  26.42857
```

Conclusion: there aren't any significant difference in the MMSCORE for group with no Conversion or Reversion to NL/MCI at 5% level of significance.

b. No vs. Conversion or Reversion:

```
Welch Two Sample t-test

data:  DX0$MMSCORE and DX12$MMSCORE
t = -0.0941, df = 7.265, p-value = 0.9276
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.769420  1.633026
sample estimates:
mean of x mean of y
  26.8068   26.8750
```

Conclusion: there aren't any significant difference in MMSCORE score between group with no Conversion or Reversion and group with either Conversion or Reversion to NL/MCI at 5% level of significance.

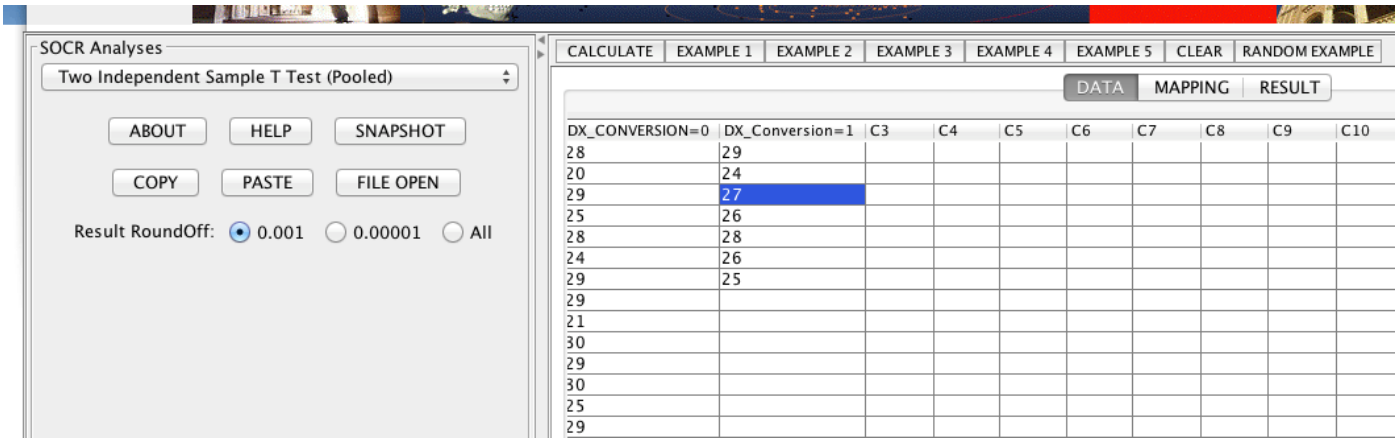
Hence, the fact whether patients have Conversion or Reversion to NL/MCI didn't have any significant influence on the MMSCORE scores. They aren't significantly associated.

RCODE:

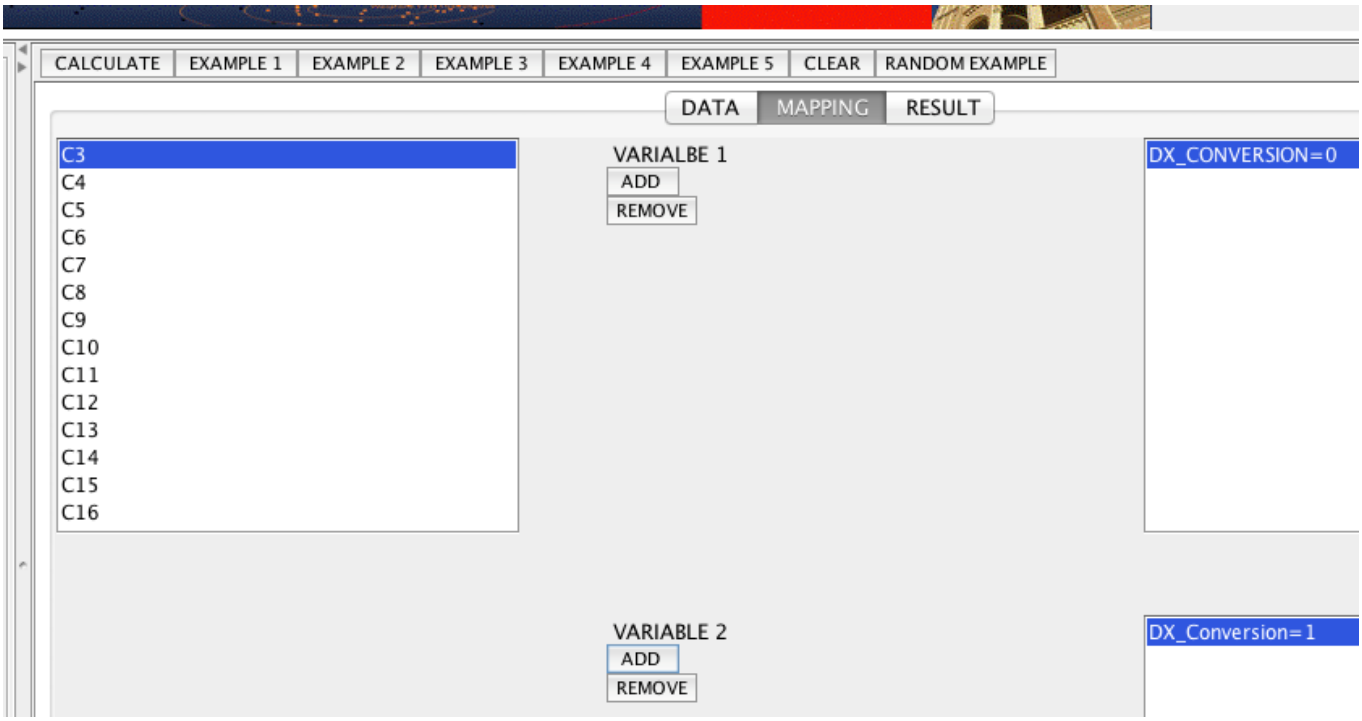
```
## problem 3
summary(DX_Conversion)
## .   0   1   2
## 1 735   7   1
data3 <- subset(biom,DX_Conversion!='.')
summary(data3$DX_Conversion)
DX0 <- subset(data3,data3$DX_Conversion==0)
DX1 <- subset(data3,data3$DX_Conversion==1)
DX12 <- subset(data3,data3$DX_Conversion!=0)
t.test(DX0$MMSCORE,DX1$MMSCORE)
t.test(DX0$MMSCORE,DX12$MMSCORE)
```

Using SOCR: Two Independent Sample T-test Pooled: http://socr.ucla.edu/htmls/SOCR_Analyses.html

Step 1: Input data:

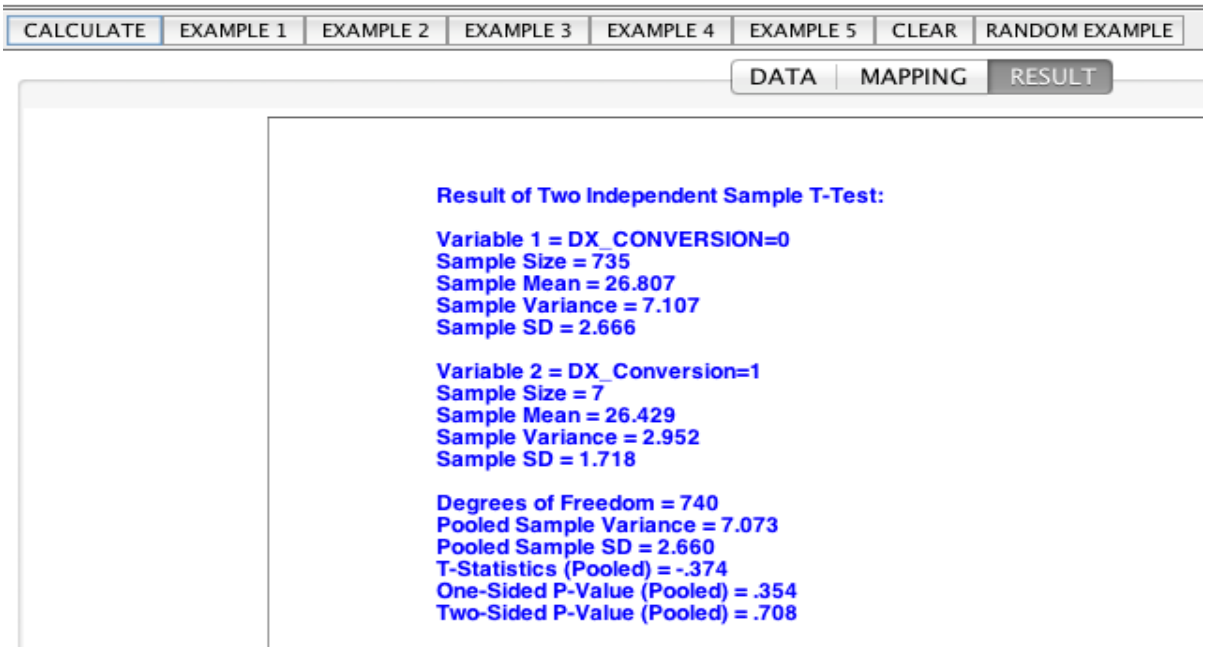


Step 2: Mapping

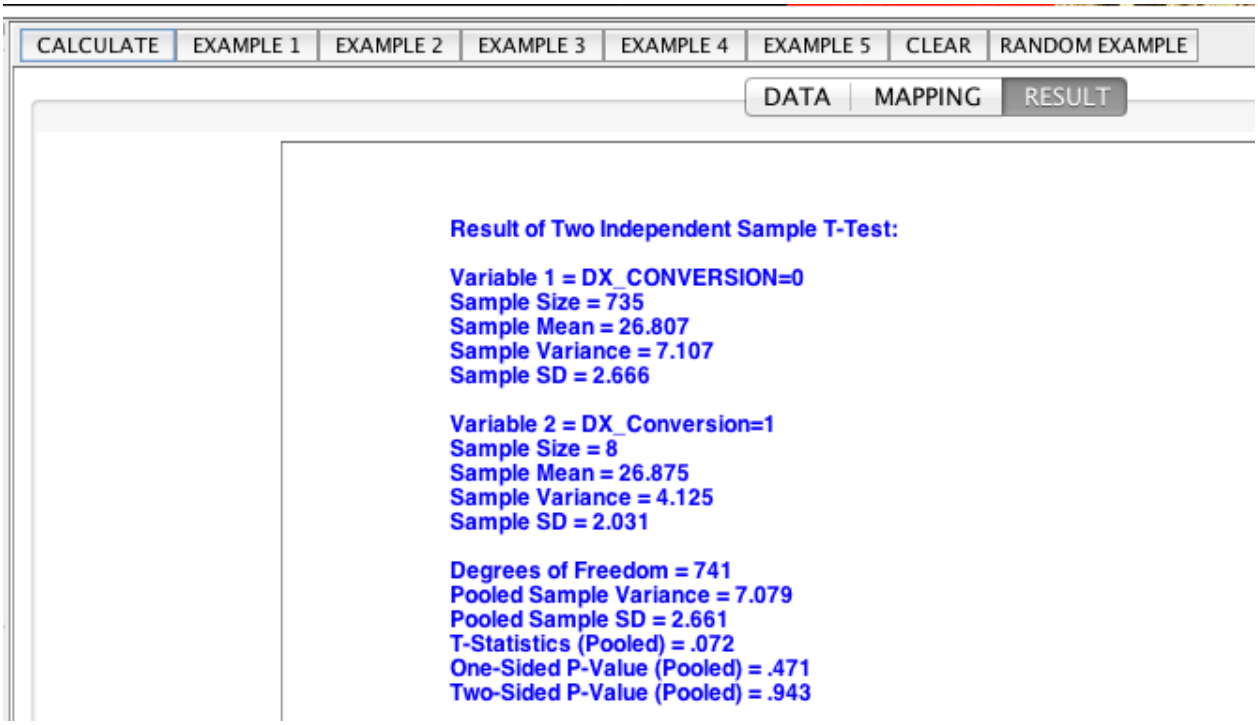


Step 3: Calculate

Step 4: Check result:



Similar for comparison of MMSCORE score between group with DX_Conversion =0 and group with DX_Conversion ≠ 0 (which has one more point with DX_Conversion=2 and MMSCORE = 30 compared to the group with DX_Conversion=1):
Result:



The result is consistent with the conclusion from R, we reject the null hypothesis at 5% level of significance and claim that whether patients have Conversion or Reversion to NL/MCI didn't have any significant influence on the MMSCORE scores. They aren't significantly associated.

Problem 4

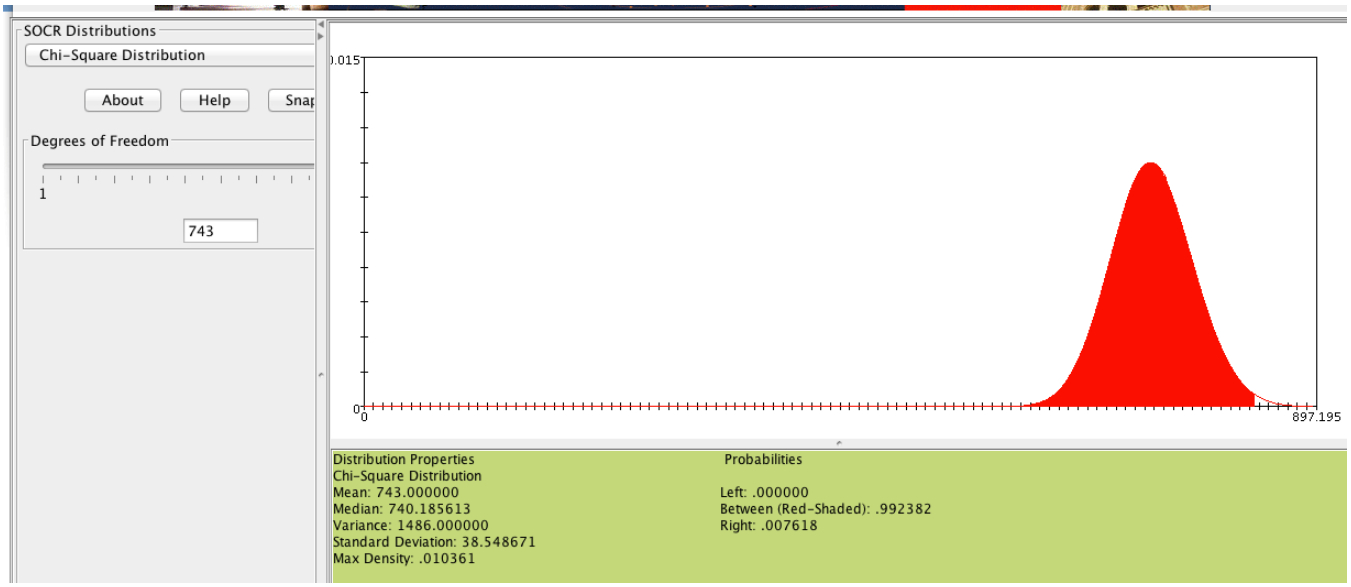
Do a Chi-square test on the standard deviation of MMSCORE with the null hypothesis of $H_0: \sigma_0^2 = 2.5^2$.
The test statistic $\sim X^2_{df=n-1}$

$$X_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(744-1)*2.657866^2}{2.5^2} = 839.7985.$$

$$df = n - 1 = 743$$

The corresponding p value is 0.007617524, so we reject the null hypothesis at 5% level of significance and claim that the standard deviation of MMSCORE scores are significantly different from 2.5.

To check on the p-value of the chi-square test: http://socr.ucla.edu/htmls/SOCR_Distributions.html



By selecting the degree of freedom of 743 and roughly a test score at around 839.7985 gives a p.value at around 0.0076175 ($p.value = P\left(X^2_{df=n-1} > X^2_o\right) = 0.0076175$), which suggest that we have enough evidence to reject the null hypothesis of $\sigma_0^2 = 2.5^2$ at 5% level of significance and claim that the standard deviation of the MMSCORE score is significantly different from 2.5.

RCODE:

```
MM.std <- sd(MMSCORE)
n <- length(MMSCORE)
chi.test <- (n-1)*MM.std^2/2.5^2
p.value <- pchisq(chi.test,df=n-1,lower.tail=F)
```

Problem 5

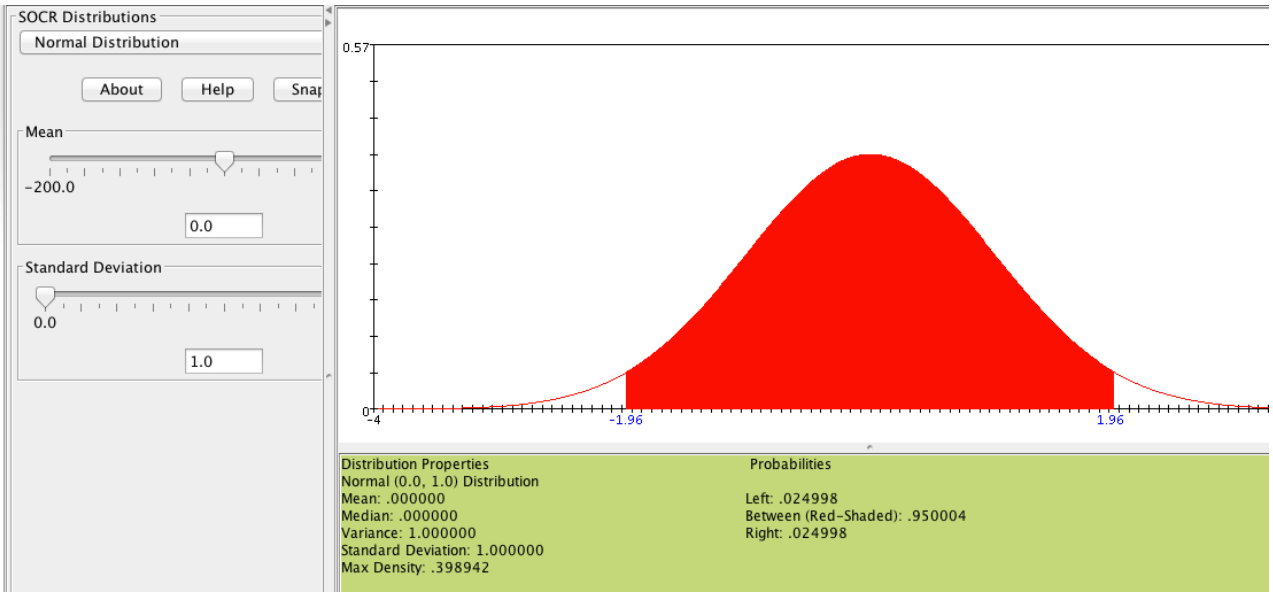
Correlation between systolic and diastolic blood pressure within group 3 and group 4 are 0.4052743 and 0.439872 respectively.

Using Fisher's transformation to test for comparing the two correlations using Normal distribution on null hypothesis $H_0: r_1 = r_2$, transform the two correlations into $r_{11} = \frac{1}{2}\ln||\frac{1+r_1}{1-r_1}||$ and $r_{22} = \frac{1}{2}\ln||\frac{1+r_2}{1-r_2}||$, the test statistic follows a standard normal distribution $N(0,1)$:

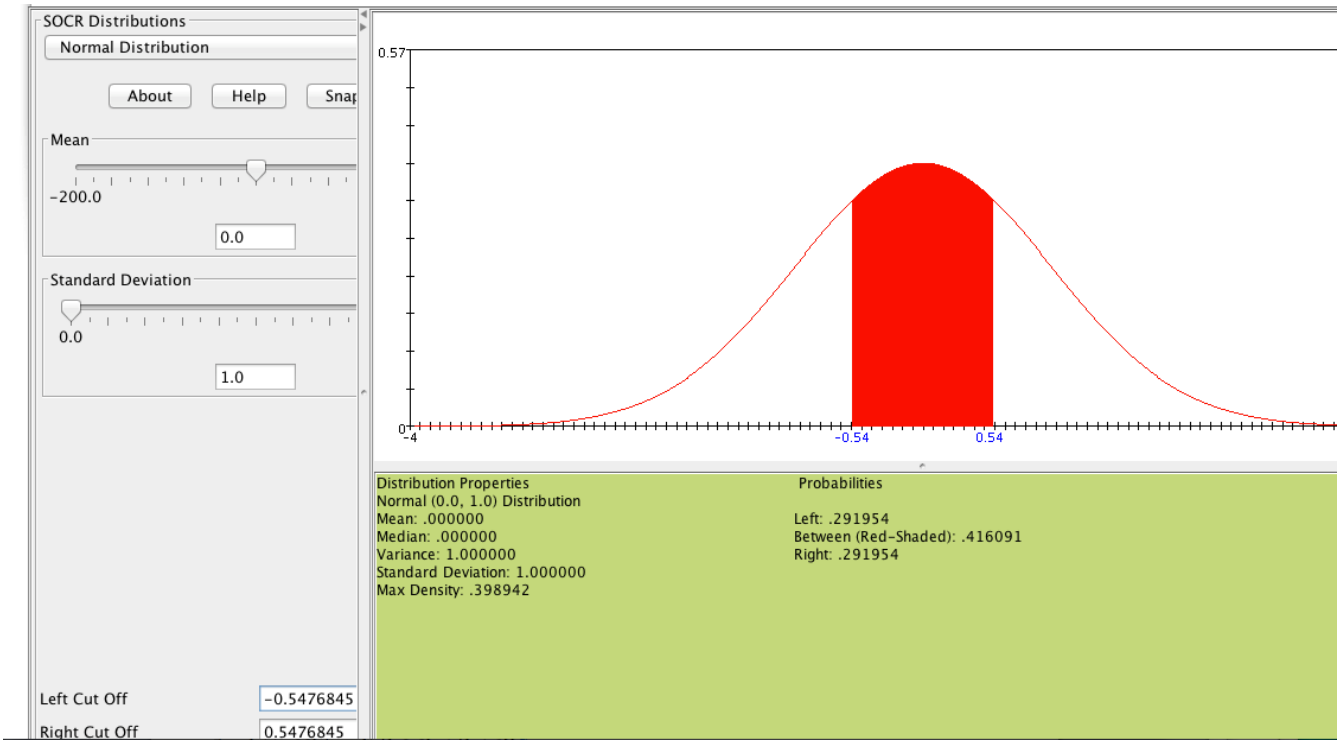
$$Z_0 = \frac{r_{11} - r_{22}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0.9441442/2 - 0.8598874/2}{\sqrt{\frac{1}{48-3} + \frac{1}{695-3}}} = -\frac{0.5476845}{2} = -0.2738423.$$

Since $|Z_0| < 1.96$, we don't have enough evidence to reject the null hypothesis of $r_1 = r_2$ at 5% level of significance. The conclusion is that the correlation between systolic and diastolic blood pressure didn't differ significantly, in fact they are very similar to each other.

To check this with the normal distribution: http://socr.ucla.edu/htmls/SOCR_Distributions.html



Note: We reject the null hypothesis if the test statistic falls in the red region of (-1.96, 1.96).
To calculate the p-value, we have



$$p.value = P(Z > Z_{0.025}) + P(Z < Z_{0.975}) = 0.291954 + 0.291954 \approx 0.5839$$

RCODE:

```
corr3 <- cor(g3$VSBPSYS,g3$VSBPDIA)  # r1=0.4052743
corr4 <- cor(g4$VSBPSYS,g4$VSBPDIA)  # r2=0.439872
r11 <- log((1+corr3)/(1-corr3),base=exp(1))
r22 <- log((1+corr4)/(1-corr4),base=exp(1))
z5 <- (r11-r22)/sqrt(1/(n3-3)+1/(n4-3))
p.value <- 2*(1-pnorm(z5,0,1,lower.tail=F))
```

Problem 6

Fit a simple linear regression of MMSCORE on VSTEMP and Weight_Kg and a brief summary of the model is given as below:

Call:
lm(formula = MMSCORE ~ VSTEMP + Weight_Kg)

Residuals:

Min	1Q	Median	3Q	Max
-8.7497	-1.7785	0.3089	2.1841	3.6242

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	22.465619	2.548285	8.816	<2e-16	***
VSTEMP	0.093785	0.067553	1.388	0.1655	
Weight_Kg	0.012355	0.006523	1.894	0.0586	

---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

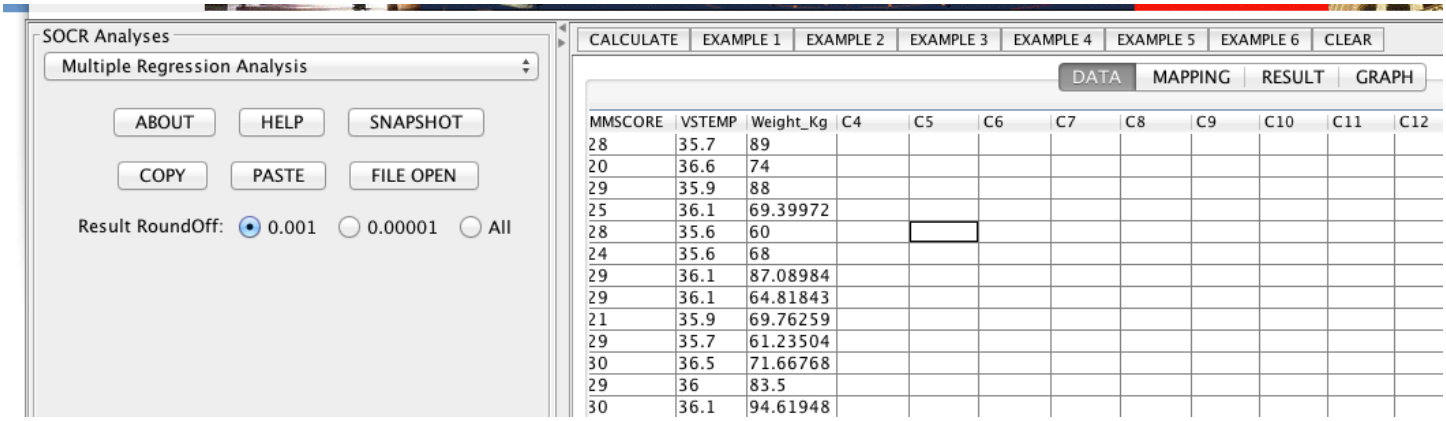
Residual standard error: 2.652 on 741 degrees of freedom
Multiple R-squared: 0.006825, Adjusted R-squared: 0.004144
F-statistic: 2.546 on 2 and 741 DF, p-value: 0.07909

From the regression model result, we can tell that the model didn't fit very well, p value of the coefficient of VSTEMP is 0.1655, which is not significant at all while the p value of the coefficient of Weight_Kg is 0.0586, which is right above 5% boundary and didn't seem to be significant either. Given the test on the coefficient is a test of trivial slope at the variable $H_0: \beta = 0$, and we fail to reject the null hypothesis for both cases. Hence, we can conclude there are trivial slope of the regression curve on VSTEMP and Weight_Kg at 5% level of significance.

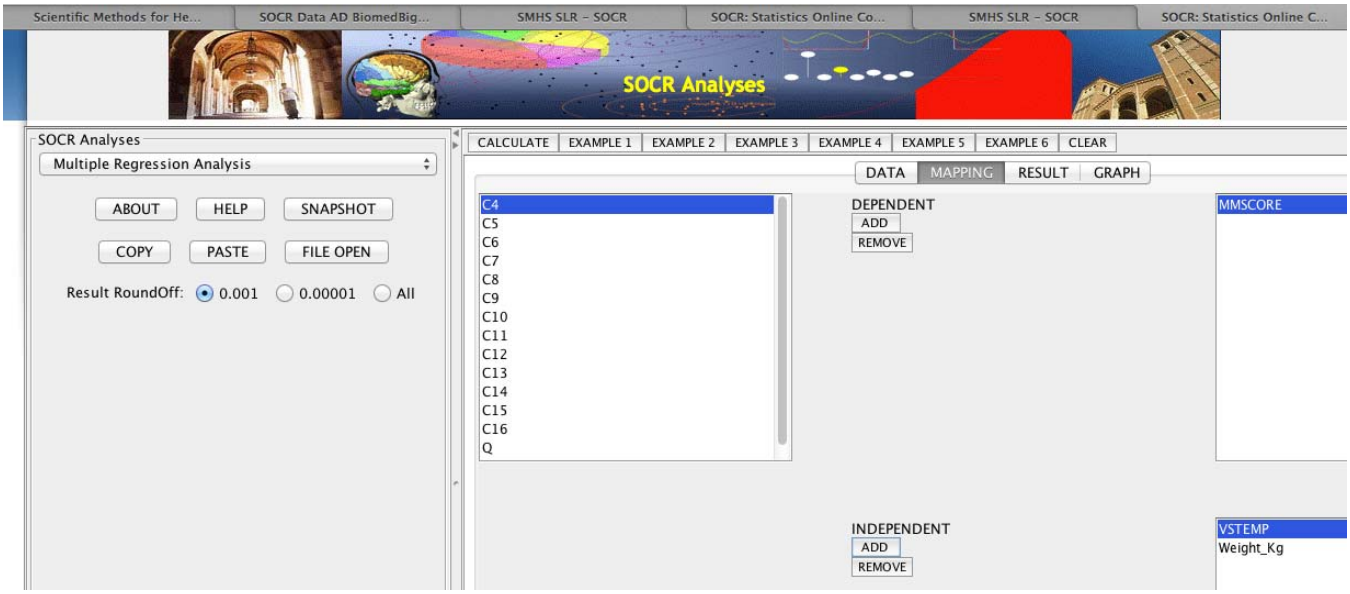
RCODE:

```
model <- lm(MMSCORE~ VSTEMP+Weight_Kg)
summary(model)
```

Using SOCR multiple regression analysis to fit simple linear regression of VSTEMP and Weight_Kg w.r.t. MMSCORE http://www.socr.ucla.edu/htmls/ana/SimpleRegression_Analysis.html
Step 1: input data

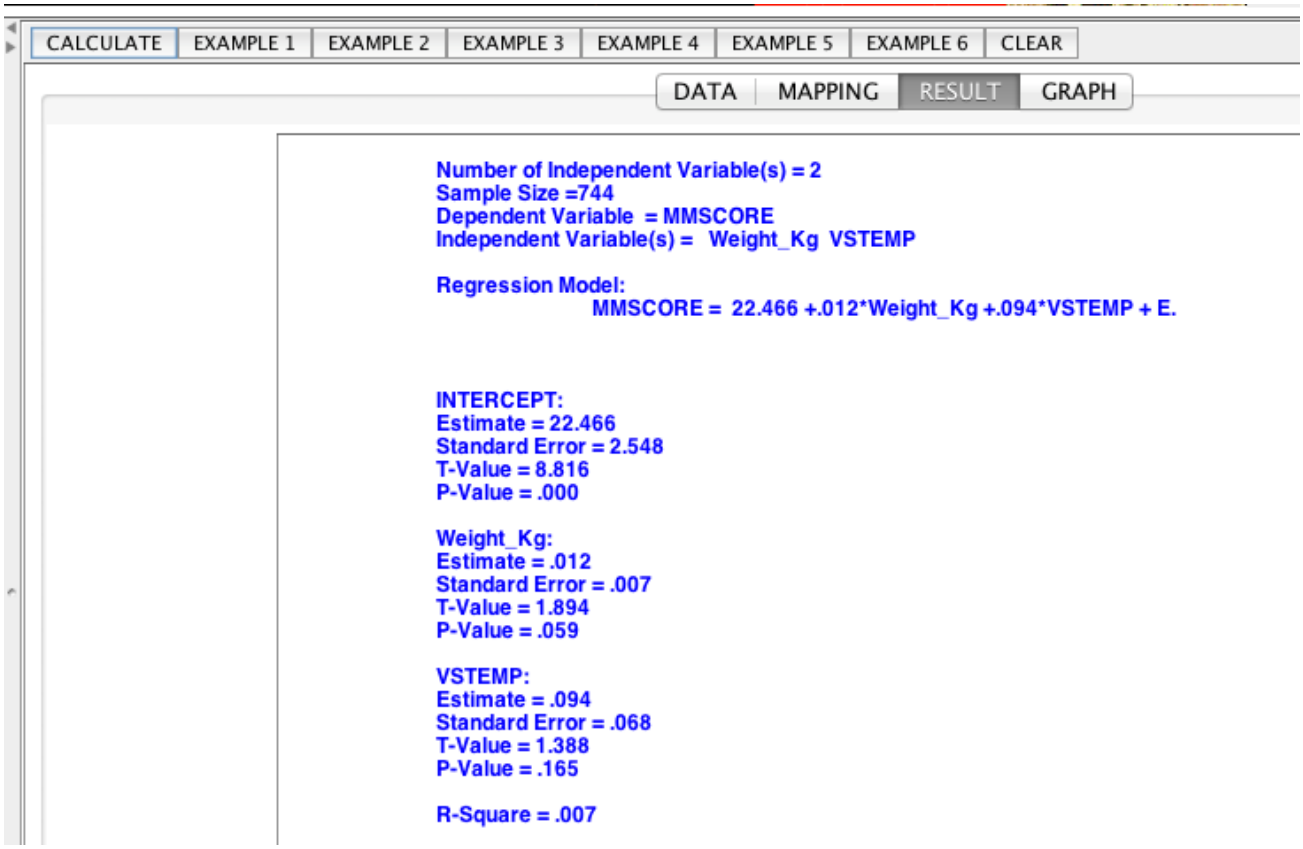


Step 2: Mapping



Step 3: Calculate

Step 4: Check result:



From the result, we can see that the p-value for estimate of coefficients of Weight_Kg and VSTEMP are 0.059 and 0.165 respectively, which suggest that we don't have enough evidence to reject the null hypothesis of coefficient equals zero for both cases. Hence, the conclusion is also consistent with the result from R that there are trivial slope of the regression curve on VSTEMP and Weight_Kg at 5% level of significance.