Problem 1)

The data given in the project come from a study of the measurements of risk of stroke affected by age & blood pressure. The study is concerned on the basis of two groups of people, smokers & non-smokers, each with 10 observations.

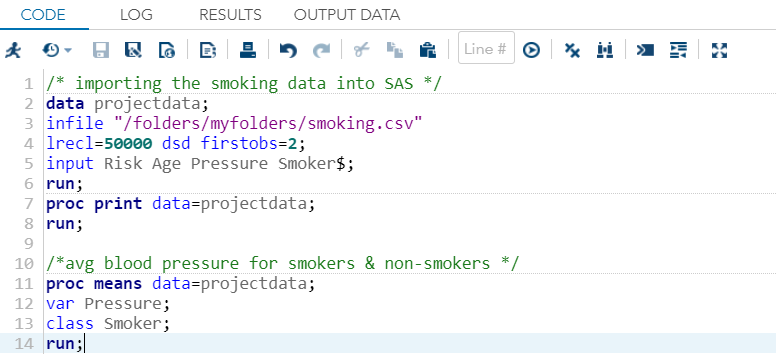
The question is whether the groups (Smokers & Non-smokers) differ in the means of blood pressure.

The required SAS codes are shown below.

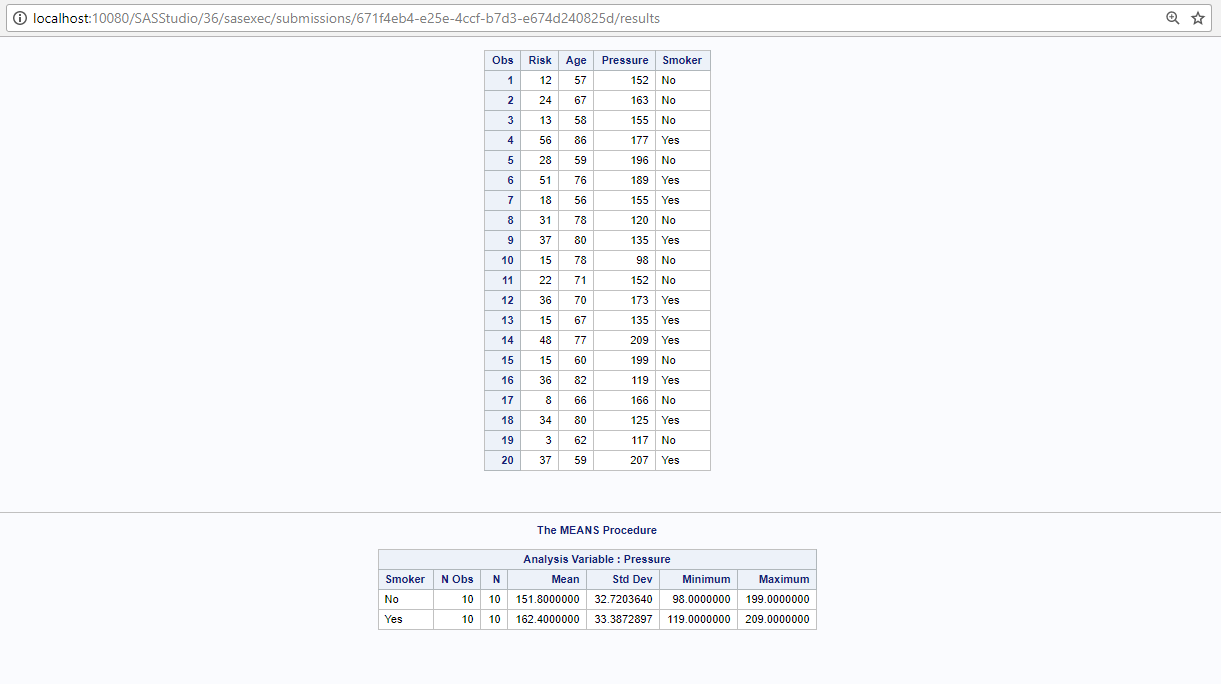
***1st Part***

***Code Window***

The following screenshot shows the codes for deciding whether the average blood pressure is higher for smokers or for non-smokers.



***Result Window***



So, average blood pressure for smokers = 162.4 & average blood pressure for non-smokers = 151.8.

Thus, it is true that the average blood pressure for smokers is **higher** than the average blood pressure for non-smokers.

***2nd Part***

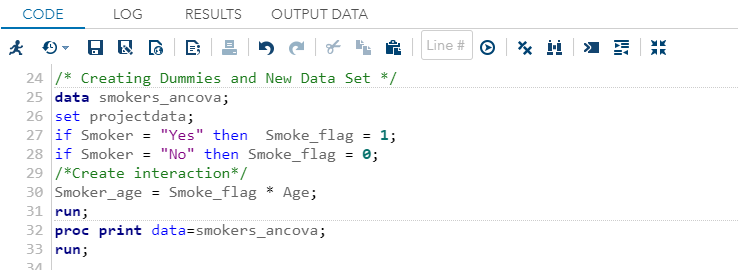
Here we have to predict the change in the dependent variable risk for change in age & blood pressure for two groups i.e., smokers & non-smokers. A general approach will be doing Regression Analysis. But here we have one binary independent variable (Smoker), two continuous variables (age & pressure) & one continuous dependent variable (risk). We want to evaluate whether the means of a dependent variable changes across the levels of categorical independent variable. So, age & pressure are considered as concomitant variables or covariates which we want to control in making comparison between the two groups.

Thus, here we can use ANCOVA or Analysis of Covariance. ANCOVA decomposes the variance in the dependent variable into variance explained by the covariates, variance explained by the categorical dependent variable & residual variance.

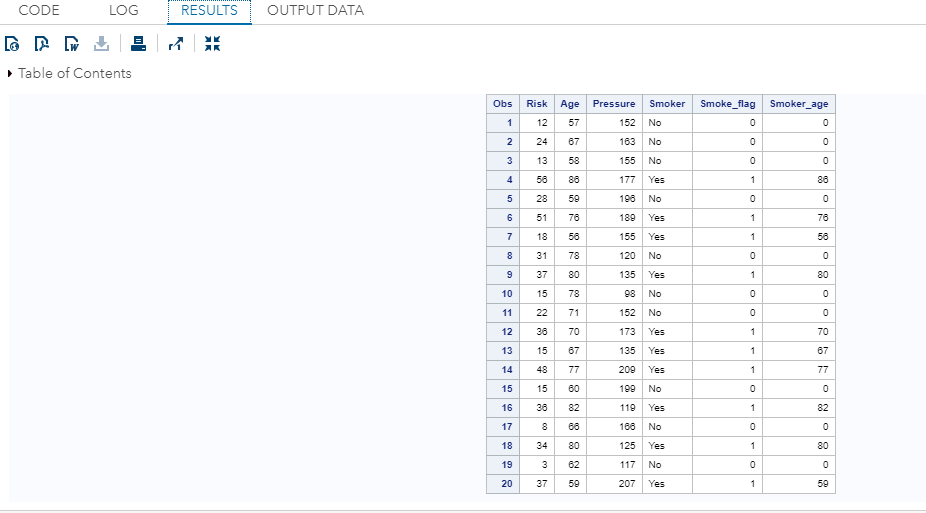
To fit the ANCOVA model, first we create involving interaction & dummy variables in the Data step.

We create a dummy variable called “Smoke\_flag” & another variable called “Smoker\_age” which represents the interaction between Age & Smoker. These new variables will be used to fit the ANCOVA model.

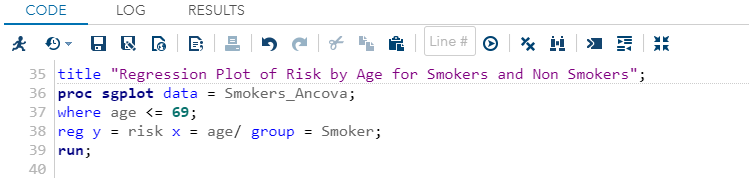
***Code Window***



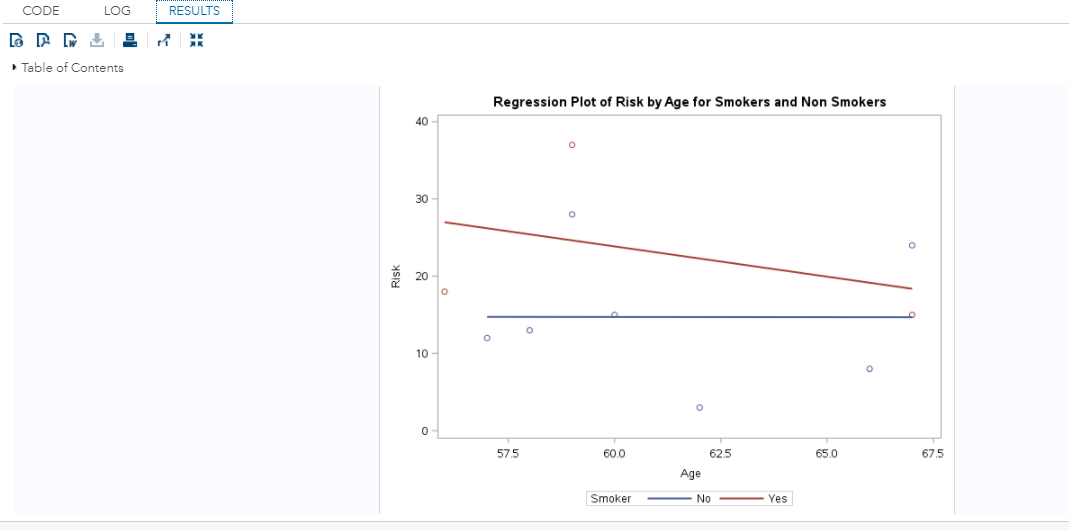
***Result Window***



Now we generate a scatter plot with Risk as the Y & Age as the X, with separate regression line for Smokers & Non-smokers.

***Code Window***

***Result Window***



From the above plots it can be seen that for Smokers, risk decreases with the increase of age. People, who smoke, have higher risk of stroke at the age of 57.5 & the risk decreases when the age is 67.5. However, people, who don’t smoke have no change in risk with the increase in age.

Now we fit an ANCOVA model, representing the relationship shown on the above graph. We include the main effects of Smoker (Smoker = “yes”), Age, their interaction & Pressure. Here we include the interaction created from the original AGE variable in this model.

***Code Window***



***Result Window***



Thus, the mean age is 69. So, we use a where statement to restrict the analysis to those who are less than or equal to 69 years old.

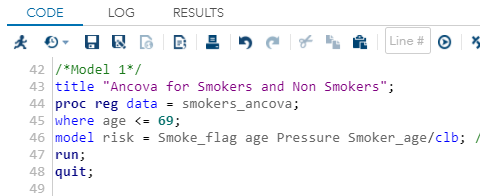
We use the clb option to get a 99.9% confidence interval for each of the parameters in the model. We have considered alpha as 0.1

The model that we are fitting is:

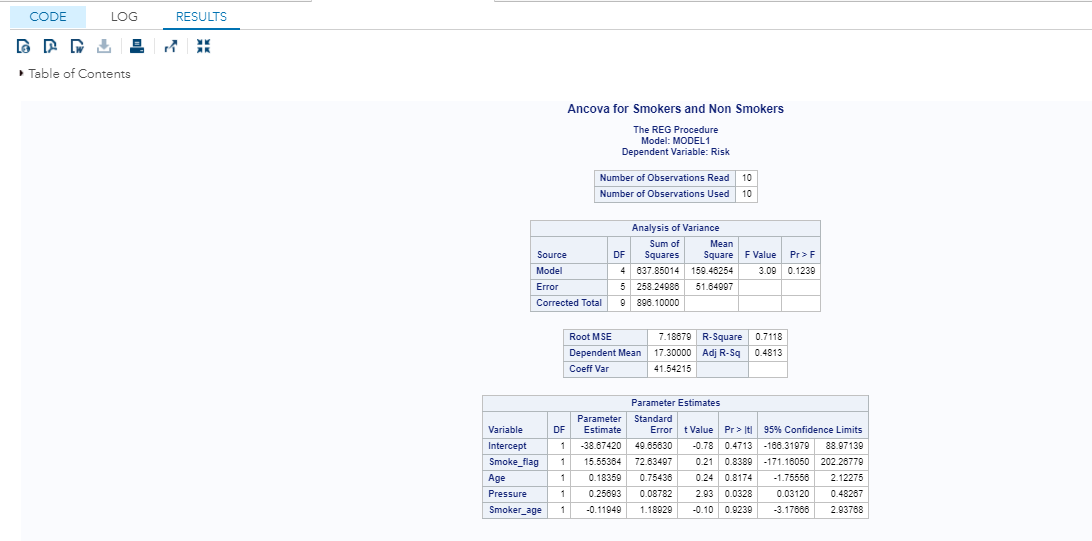
Risk = β0+ β1\* Smokeri + β2\* Agei + β3\* Smoker\_agei +

β4\* Pressurei+εij

***Code Window***



***Result Window***



We can see the model has 4 degrees of freedom, corresponding to the 4 predictors included in the model. We can interpret the overall significance by looking at the ANOVA table.

Here, F(4,5) = 3.09, p= 0.12 & Adj. R2= 0.48.

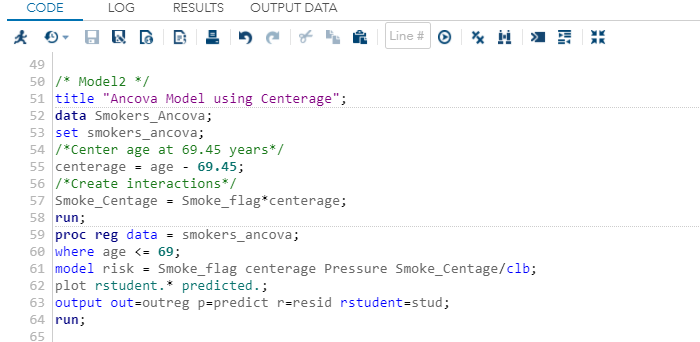
p>0.1, the model is near about significant. However, since the R2 value is very close to 0. So this model is not predicting well.

We first look at the parameter estimate for the interaction term. The interaction term, β4 (estimated to be -0.11949) represents the difference in the slope of the regression line for smokers vs. the reference category non-smokers.

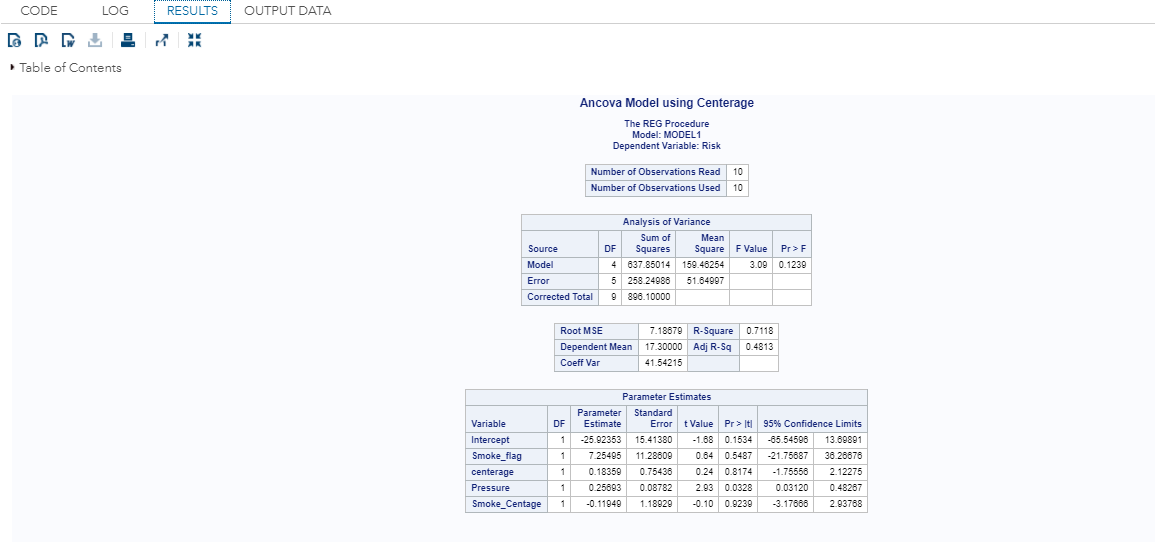
**ANCOVA Model with Centered Age:**

One way to help in the interpretation of the coefficients in a model like this is to center the continuous and then create an interaction term between centered age (centerage) & smokers. The new interaction will be called Smoke\_Centage. Here we center age at 69.45 years, which is the approximate mean of age variable. This is like shifting the X-Axis in our model, so that the value of 0 for centerage represents 69.45 years of actual age.

***Code Window***



***Result Window***



Note that the Analysis of Variance table and the model R-Square in the output below are the same as for the previous model. However, the parameter estimates are different.

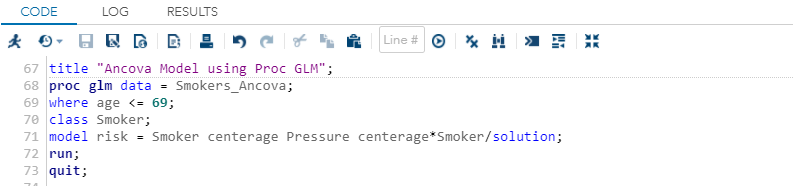
The interaction term which represents the difference in slope for smokers & non-smokers (estimated to be -0.11949) is the same as in the previous model. The coefficient for “centerage” is 0.18359 which represents the slope for the reference category (non-smokers), is the same as the coefficient for AGE in the previous model. However, the estimated values for the variables Smoke\_flag & the intercept are different than the previous model. We see that the estimated effect of Smoke\_flag is 7.25495. We can interpret this as the estimated difference in the average risk of smokers vs. non-smokers when they are 69.45 years of old (i.e. when centerage is zero). In other words, people who smoke, have 7.25 units lesser chance risks than people who don’t smoke at age 69.45 years.

The INTERCEPT in this model tells us the estimated average risk of smokers when Centerage is zero. It is often helpful to center continuous variables in a regression model. It helps in interpreting the intercept in the model, and can also help in interpreting the main effects of variables that are included in interactions. When one centers the continuous variable, the interaction term is computed by multiplying the dummy variable for Smoke\_flag times the Centered version of the continuous variable.

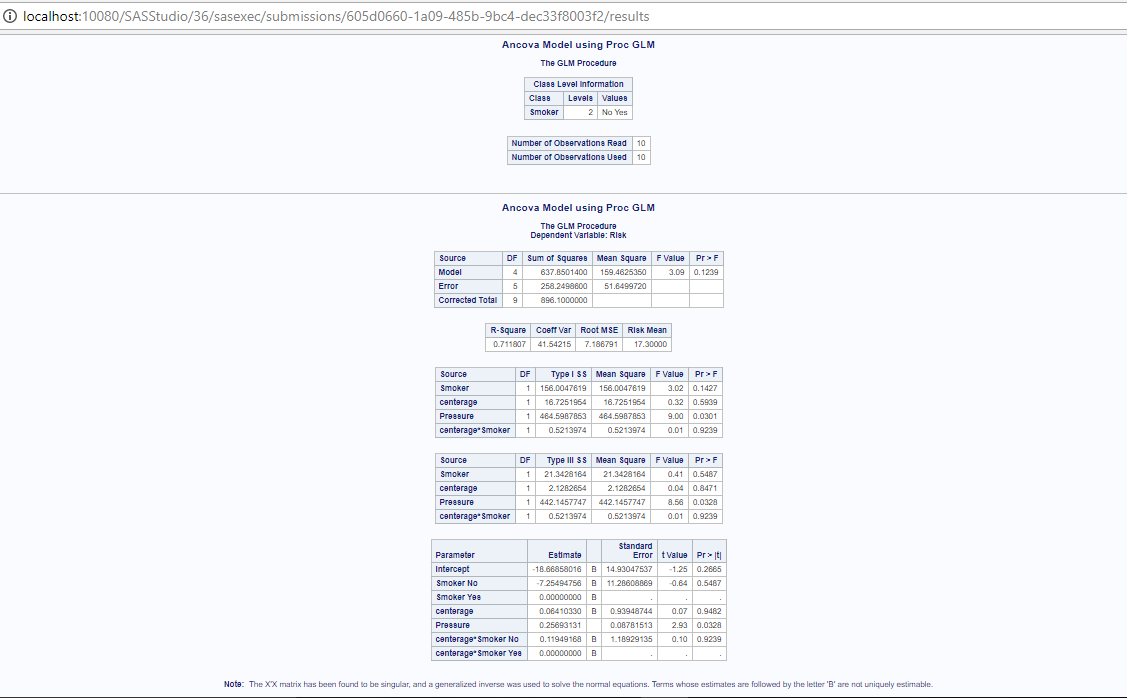
**ANCOVA Model Using Proc GLM:**

We now refit the model using centerage & smokers as predictors, but using Proc GLM. The advantage of using this procedure is that you don’t need to create dummy variables for our categorical predictors, and the interaction terms do not need to be created in advance. The categorical variable “Smoker” is listed in the Class statement in SAS. The solution option is used to request that SAS print out the parameter estimates from the model. This option is not necessary, but is used for comparison with the parameter estimates from Proc Reg.

***Code Window***



***Result Window***



From the above result we can see that the coefficient of determination i.e., R2 is 0.711807 which means that the model is predicting the dependent variable risk quite well.

In the output, the Type I SS shows the effect of each predictor in the model, sequentially. That is, the effect of Smoking is evaluated without controlling for the other predictors. The effect of AGE is evaluated with only Smoking in the model, and the effect of the CENTERAGE by Smoking interaction is evaluated, after adjusting the main effects. The total of the Type I SS is equal to the total model SS.

The Type III SS below shows the effect of each predictor in the model, controlling for all other effects. The Type III SS is sometimes called the regression sum of squares or partial sum of squares. In this case, the total of the Type III SS does not equal the total model SS.

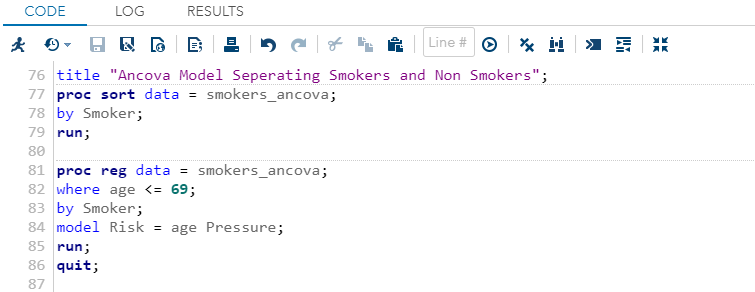
Notice that we get the same parameter estimates using Proc GLM as we did in Proc Reg. By default, Proc GLM over parameterizes the model, including a parameter for each level of Smoker. The parameter estimate for the highest level of Smoking is set to zero, which has the effect in this case of making non-smokers the reference category, as we had when we fit the model using Proc Reg. Although the parameters are not uniquely estimable in this over parameterized model, we can interpret the parameter estimates, knowing the convention that SAS uses for the parameters in the model.

**Separate Regression Models for Males and Females:**

We now fit separate regression models for smokers & non-smokers. To do this, we first sort the data by Smoker & then fit the regression model by Smoker, using a by statement. We select only cases with AGE<=69.

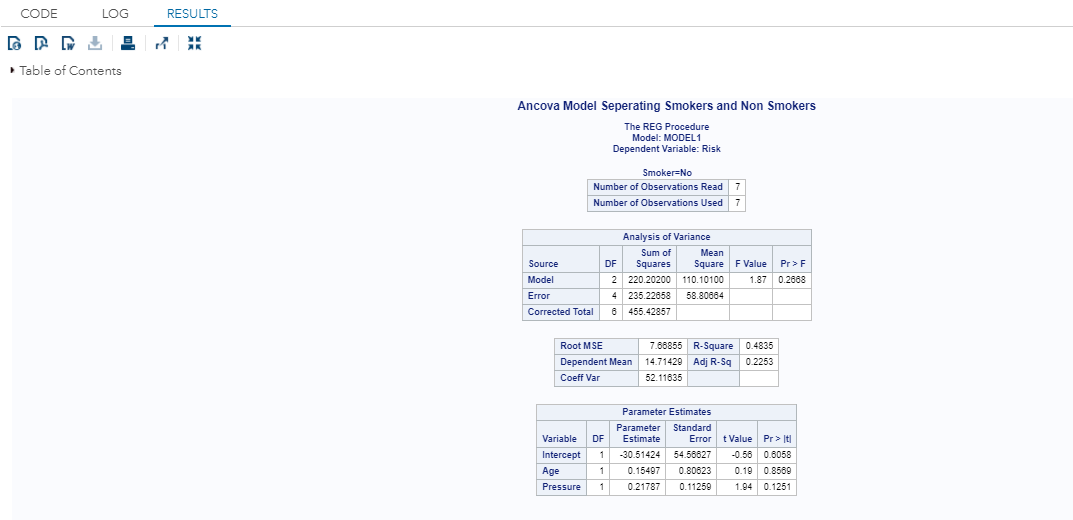
The advantage of the ANCOVA model is that we get a direct test of whether the slope for AGE is the same for smokers and non-smokers, whereas in the individual regression models, we do not.

***Code Window***



***Result Window***

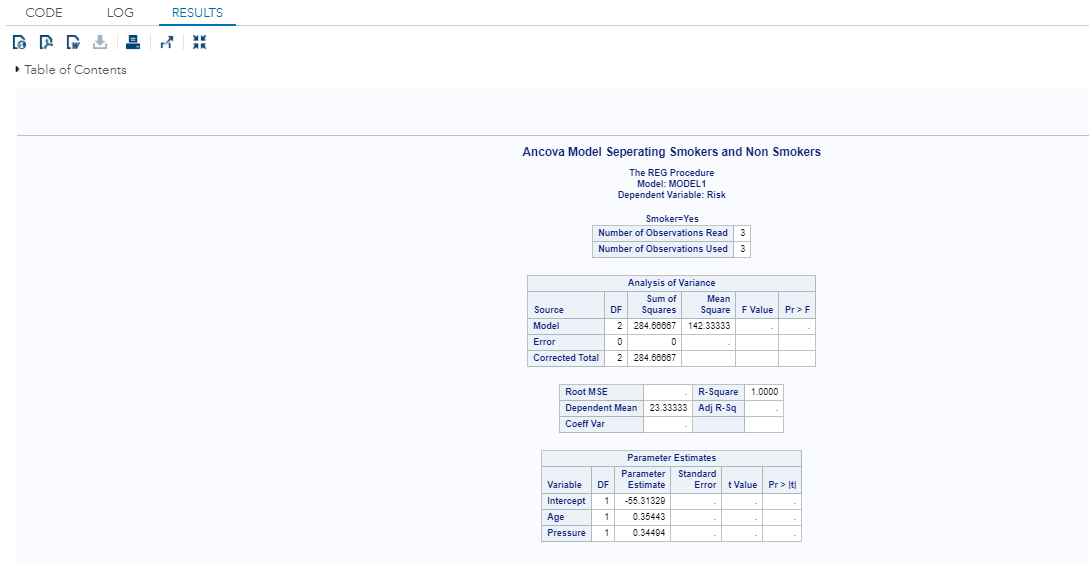
Output for “Smoker”=no



From the above result it is seen that for one unit increase in Age, risk increases 0.15 units & altering the pressure by one unit, the risk increase by 0.217 units.

The above model is not at all significant in predicting the dependent variable “risk”.

Output for “Smoker”=yes



In the above output the value of R2 is 1.0. An R2 of 1 indicates that the regression line perfectly fits the data. So, the smokers having age<=69, the result shows that Risk is positively & highly correlated with the independent variables age & pressure i.e., the risk increases with the increase in pressure & age for people who smoke.

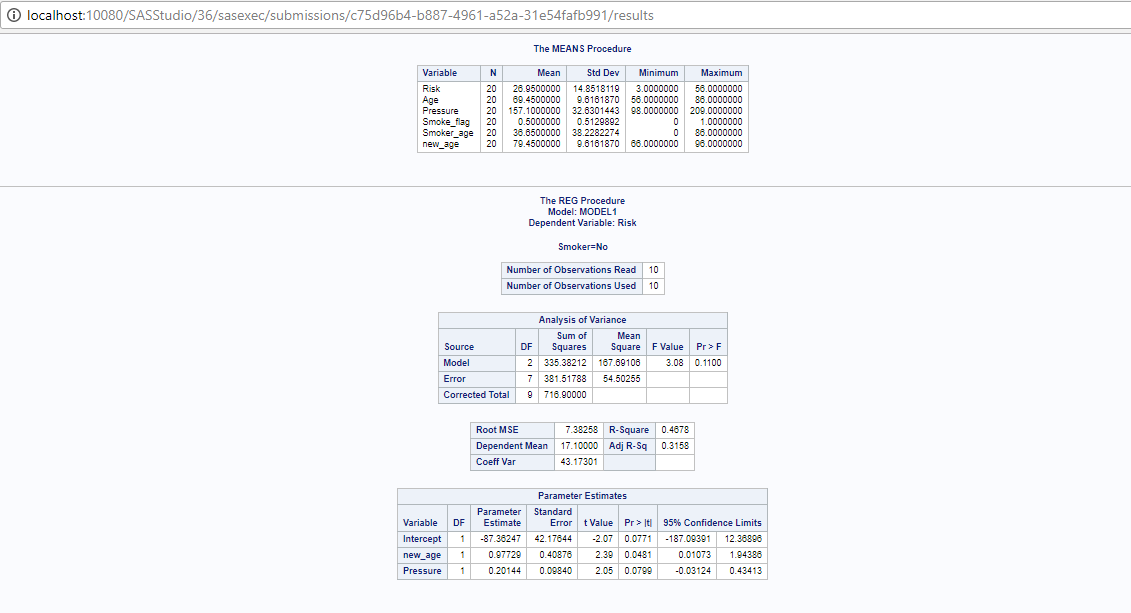
***Age is increased by 10 years:-***

The following codes show the change in risk if age is increased by 10 years for smokers & non-smokers.

***Code Window***

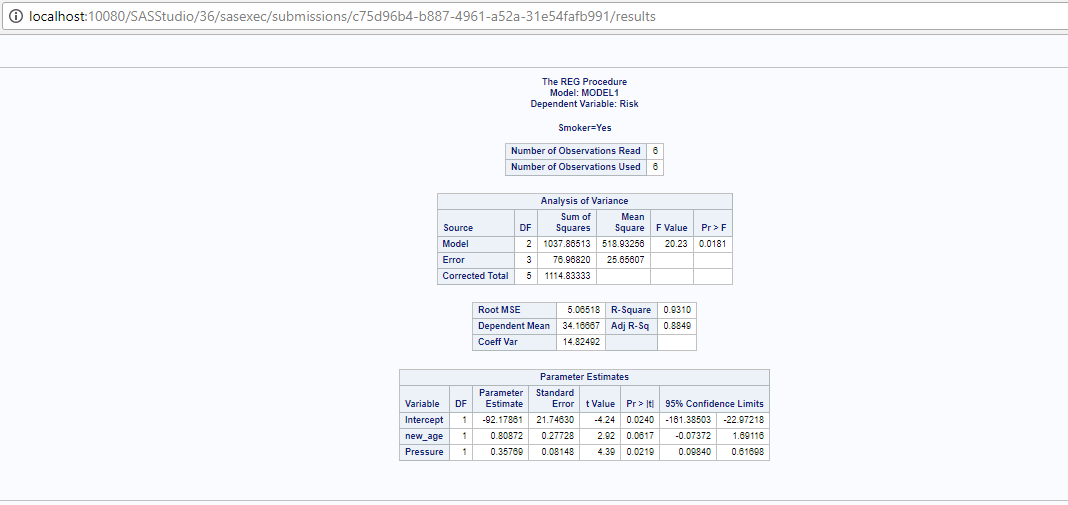
***Result Window***

Question #D~The output for “Smoker”=no



The above output shows the model is statistically insignificant as p=0.40876>0.1. If the age is increased by 10 years, then altering the new\_age by one unit increases the risk by 0.97 units. Also, one unit for one unit change is pressure the risk changes by 0.201 units.

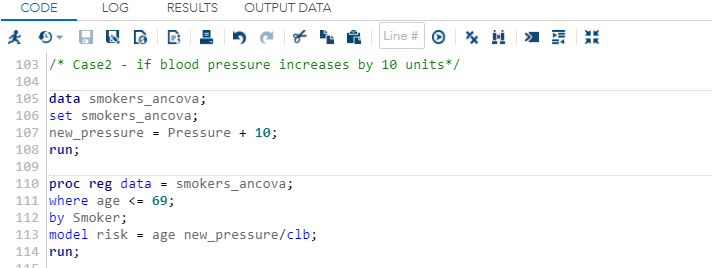
Question #A~The output for “Smoker”=yes



The model is statistically insignificant as the p values are more than alpha =0.1. But the R2 value = 0.93 i.e., high R2 value indicates predicted values are very close to the regression line.

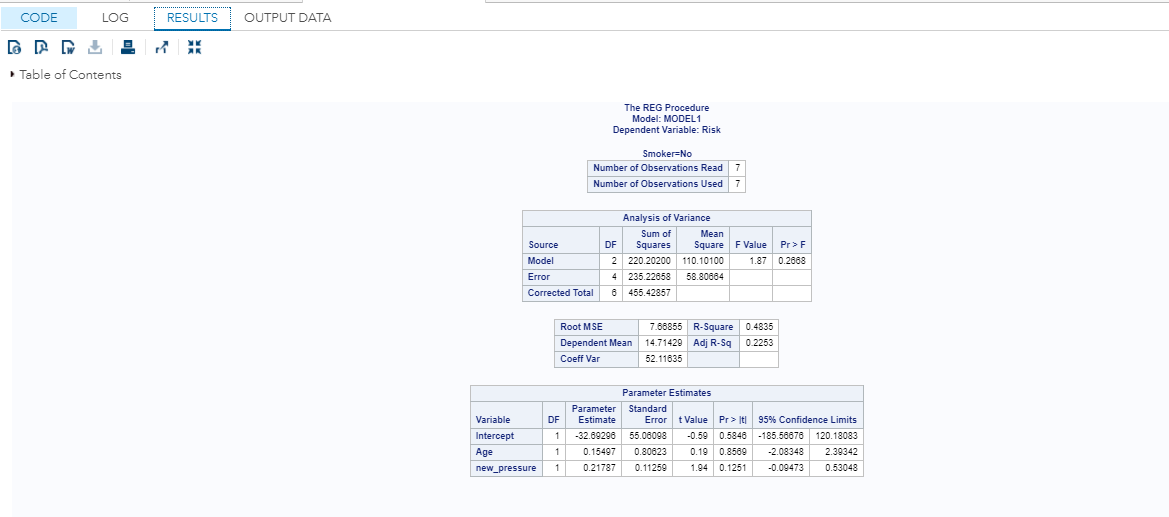
***Blood pressure is increased by 10 units:-***

***Code Window***



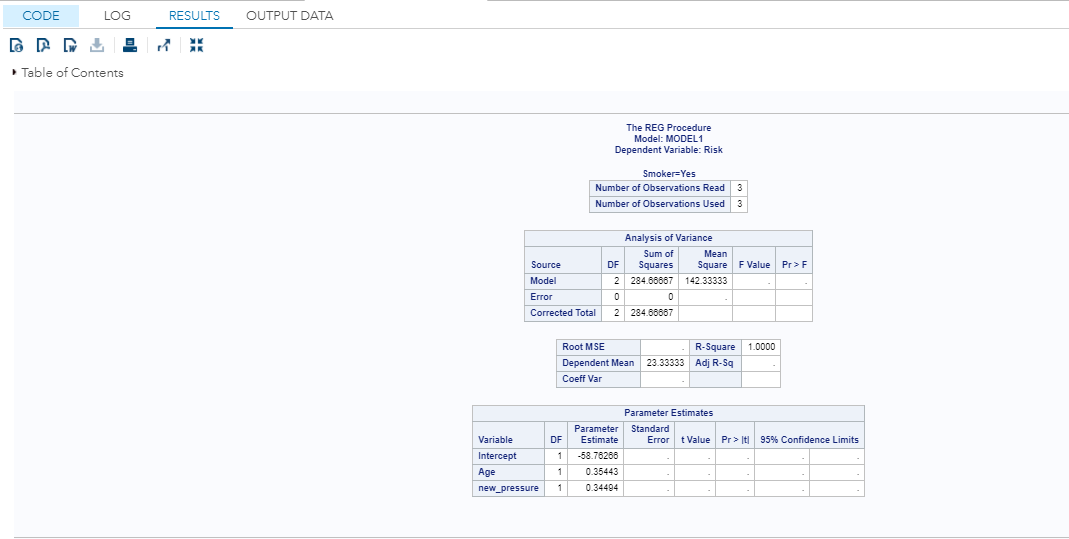
***Result Window***

Question #E~The output for “Smoker”=no



If the pressure is increased by 10 units then the change in one unit pressure will increase the risk 0.21787 units. But the coefficient of determination is 0.4835. Thus the model is not predicting well i.e. the predicted values are not very closer to the regression line.

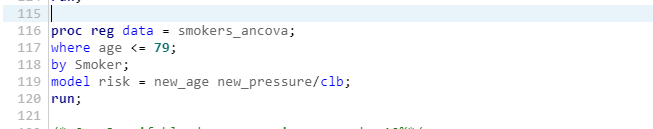
Question #B~The output for “Smoker”=yes



For smokers, the coefficient of determination is 1. So, the independent variables are perfectly positively correlated to the dependent variable risk. Therefore 100% variability is explained by the dependent variables.

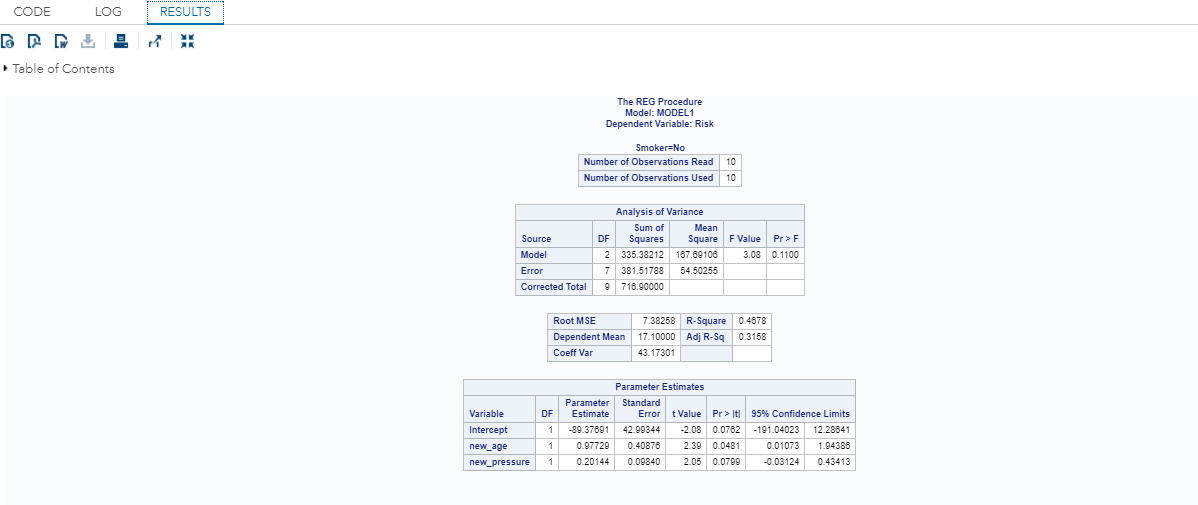
Now we check the outputs of Question B & Question E by taking the variable “new\_age” into consideration i.e., for case we are studying the change in risk variable for 10 years increase in age & 10 units increase in pressure for smokers & non-smokers respectively.

***Code Window***



***Result Window***

Question #E~The output for “Smoker”=no (considering the new\_age variable)



If new\_age is taken into consideration, then altering a unit of new\_pressure will increase the risk by 0.20144 units.

Question #B~The output for “Smoker”=yes (considering the new\_age variable)

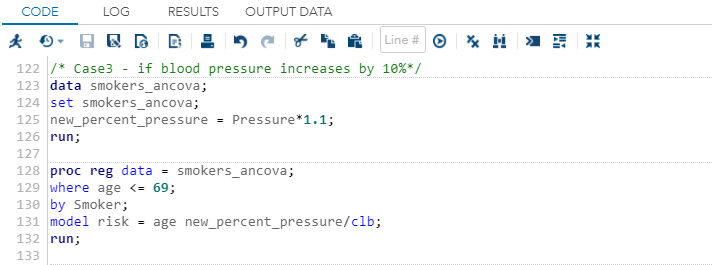


When smoker=yes then the coefficient of determination is very high i.e., 0.93. It indicates that the values are very close to the regression line. The change in one unit new\_pressure will increase the risk by 0.35769 units.

***Blood pressure is increased by 10%:-***

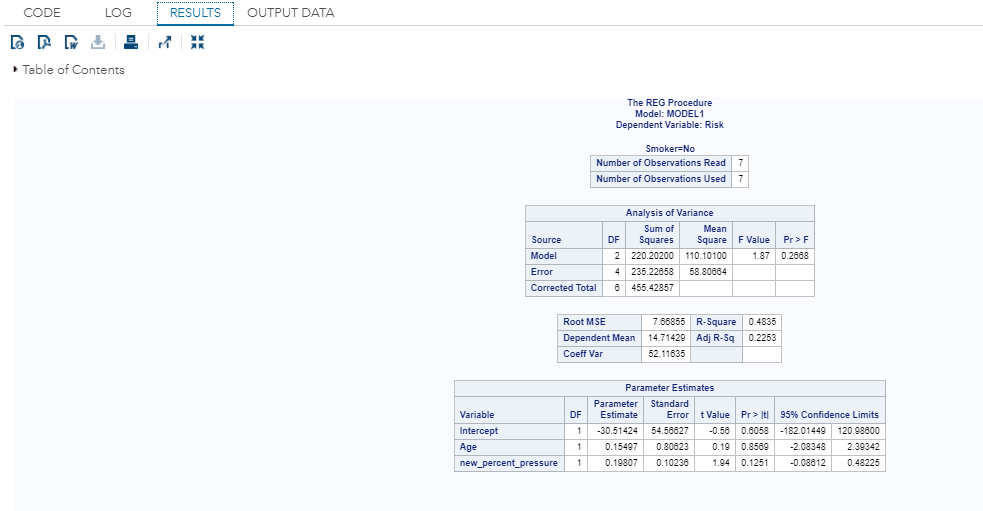
The following codes show the change in risk if pressure is increased by 10%.

***Code Window***



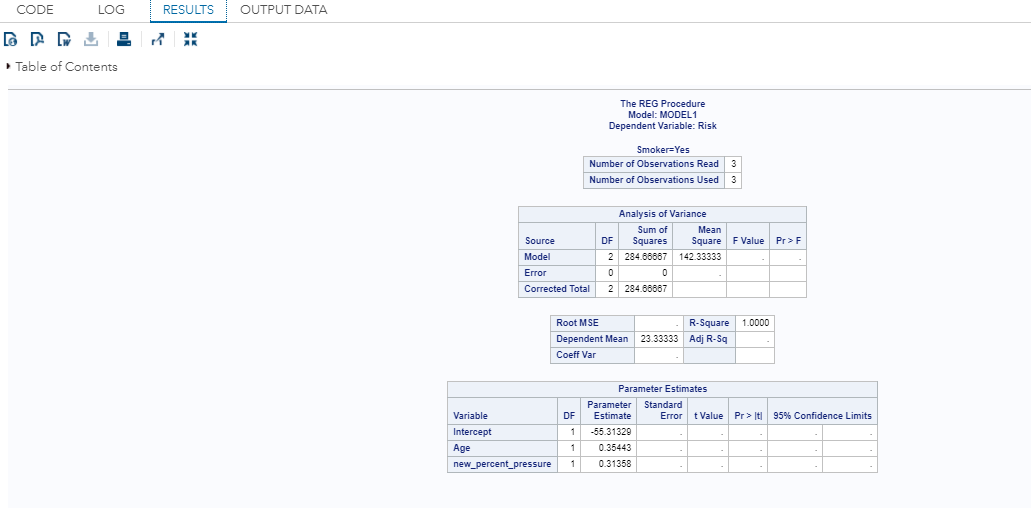
***Result Window***

Question #F~The output for “Smoker”=no



The new\_percent\_pressure is the new variable that is constructed to get the value of pressure when pressure is increased by 10%. When smoker =no, then altering one unit of “new\_percent\_pressure” changes the risk value by 0.19807 units.

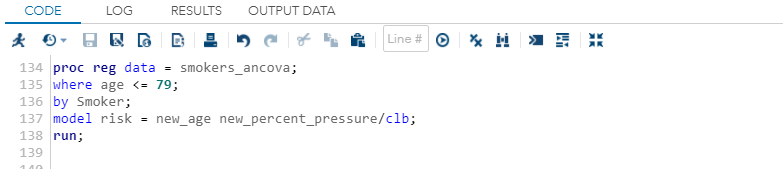
Question #C~The output for “Smoker”=yes



As in the previous model, when smoker = yes, then the coefficient of determination is 1. If one unit of new\_percent\_pressure is changed then the risk is changed by 0.31358 units.

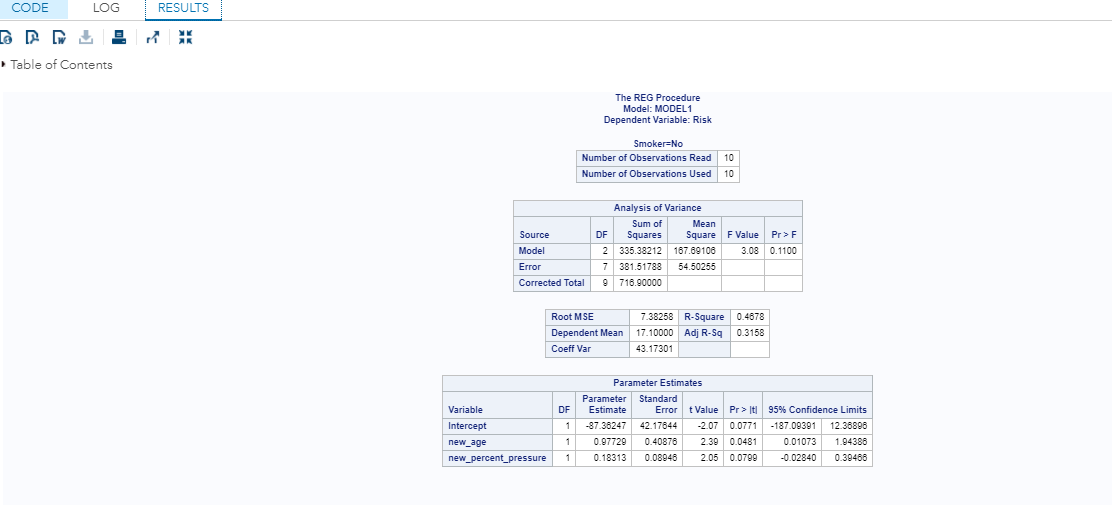
Now we check the outputs of Question C & Question F by taking the variable “new\_age” into consideration i.e., we are studying the change in risk variable for 10 years increase in age & 10% increase in pressure for smokers & non-smokers respectively.

***Code Window***



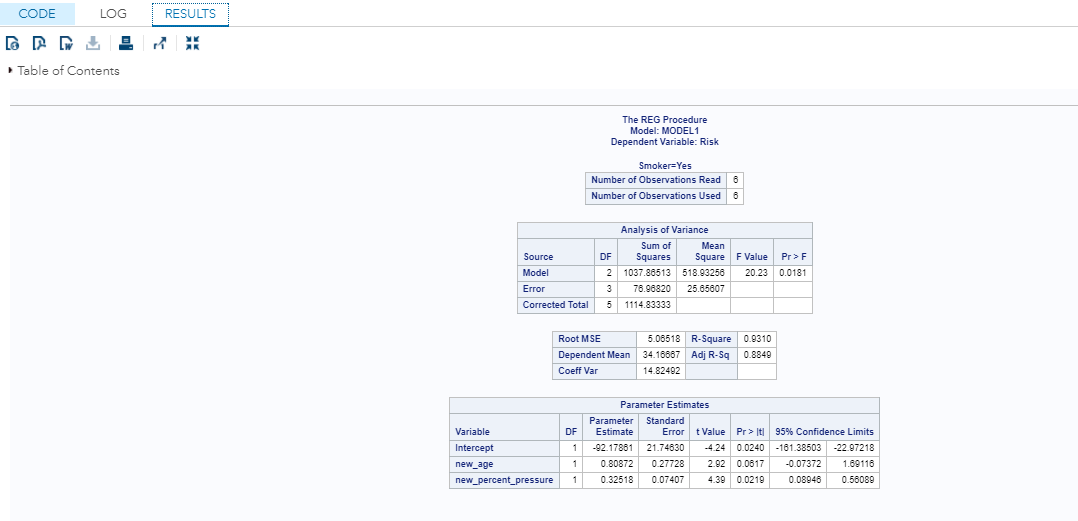
***Result Window***

Question #F~The output for “Smoker”=no (considering the new\_age variable)



The change in the new\_percent\_pressure is influenced by new\_age variable. Altering a unit of new\_percent\_pressure will change the risk by 0.18313 units. The model is significant for new\_percent\_pressure as p=0.0799<0.1.

Question #C~The output for “Smoker”=yes (considering the new\_age variable)



The dataset for smoker =yes, the change of one unit in new\_percent\_pressure will increase the risk value by 0.32518 units. The coefficient of determination is 0.93 which indicates that the predicted values are very close to the regression line.

**Conclusion:**

So from the previous study of different models, it can be concluded that the output for smoker=”yes” for each case explains 100% variability of the dependent variable by the independent variables.

On the other hand, when smoker = “no” the value of R2 is very low & near about 48% of variability is explained for each case by the independent variables age & pressure.

Problem 2)

To estimate the odds that the person is a smoker we have to perform Logistic regression where the dichotomous dependent variable is “Smoker” & independent variables are “Risk”, “Age” & “Pressure”. Logistic regression models a relationship between predictor variables and a categorical response variable. The Logistic regression equation is expressed as the inverse of the logit of p.

Logit(p) = Log( ) = β0+ β1x1+ β2x2 + β3x3 +…..+ βmxm

Where β0 = slope & β1,…., βm are coefficients of response variables x1,…. xm.

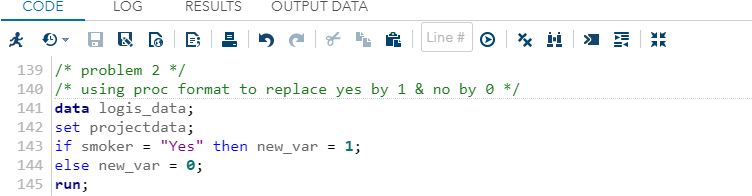
p = probability of success (probability that the person is a smoker).

The above equation states the (natural) logarithm of the odds is a linear function of the X variables (and is often called the log odds).

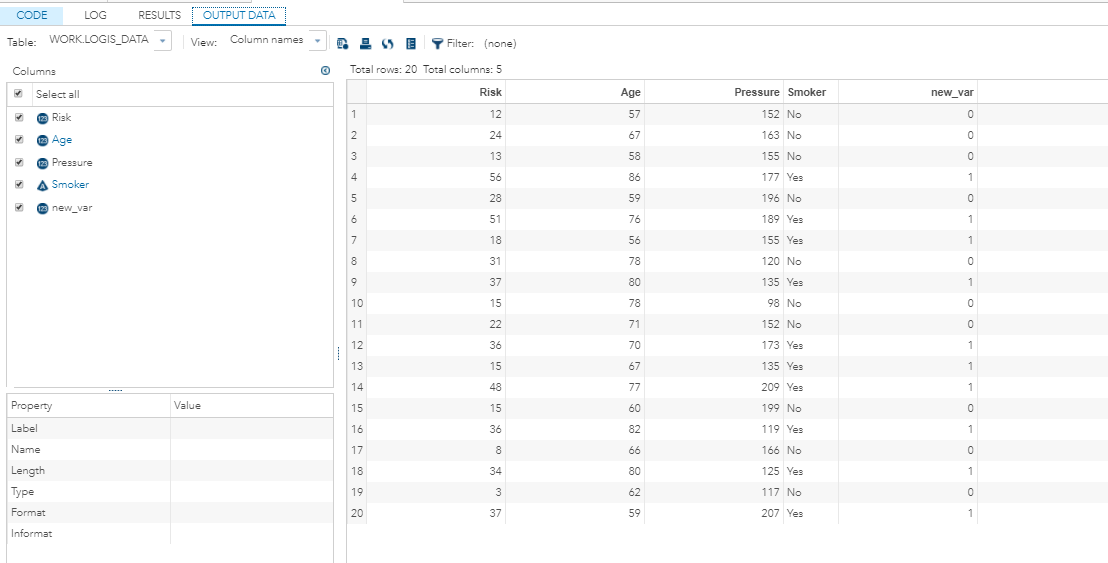
To yield the output in a CSV file, we have to SAS Output Delivery System or SAS ODS.

First we define a new variable “new\_var” which holds the value 1 for Smoker =”yes” & 0 for Smoker = “no”. This new\_var will be used as the dependent variable for the further calculation.

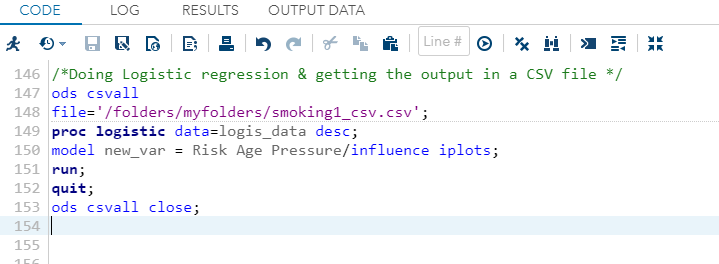
***Code Window***



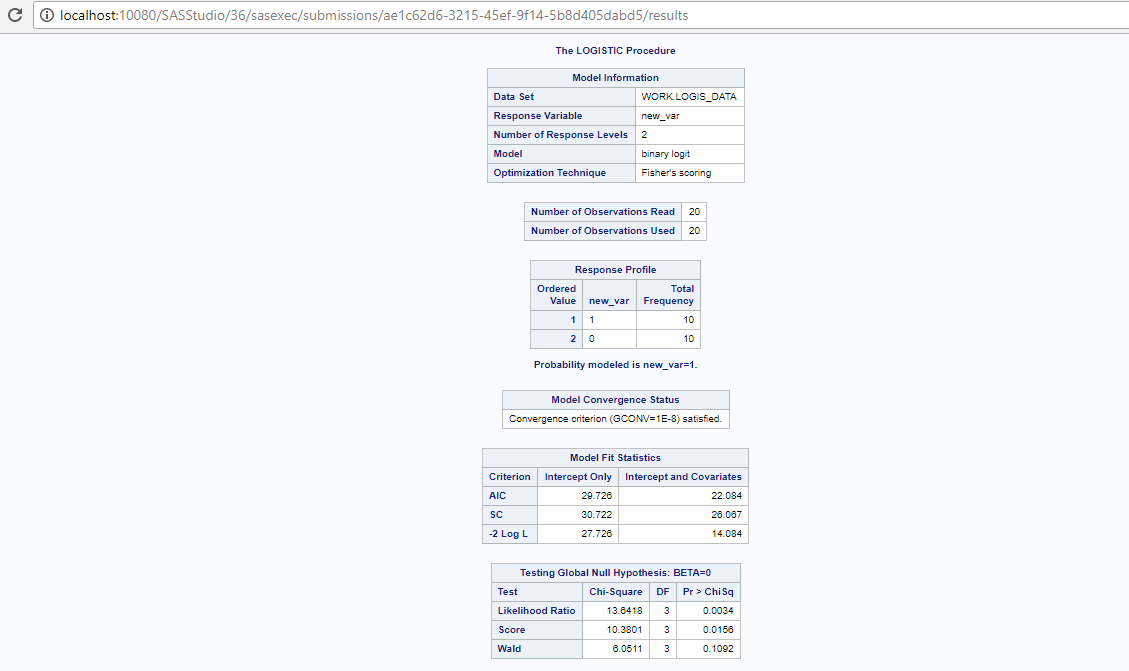
***Output Window***

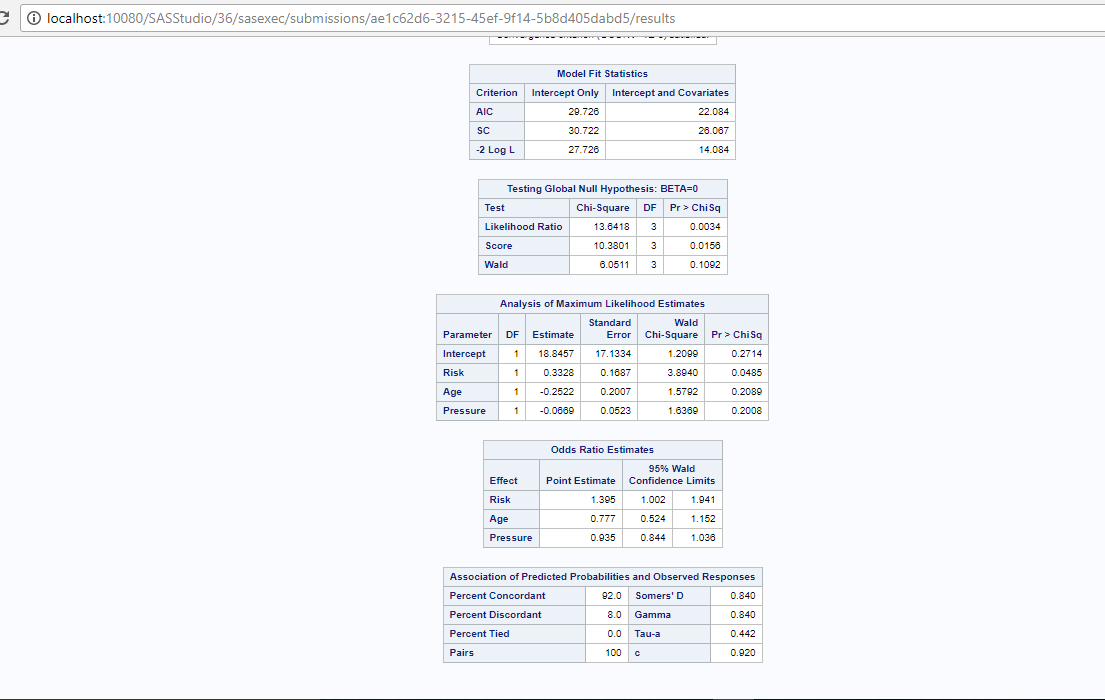


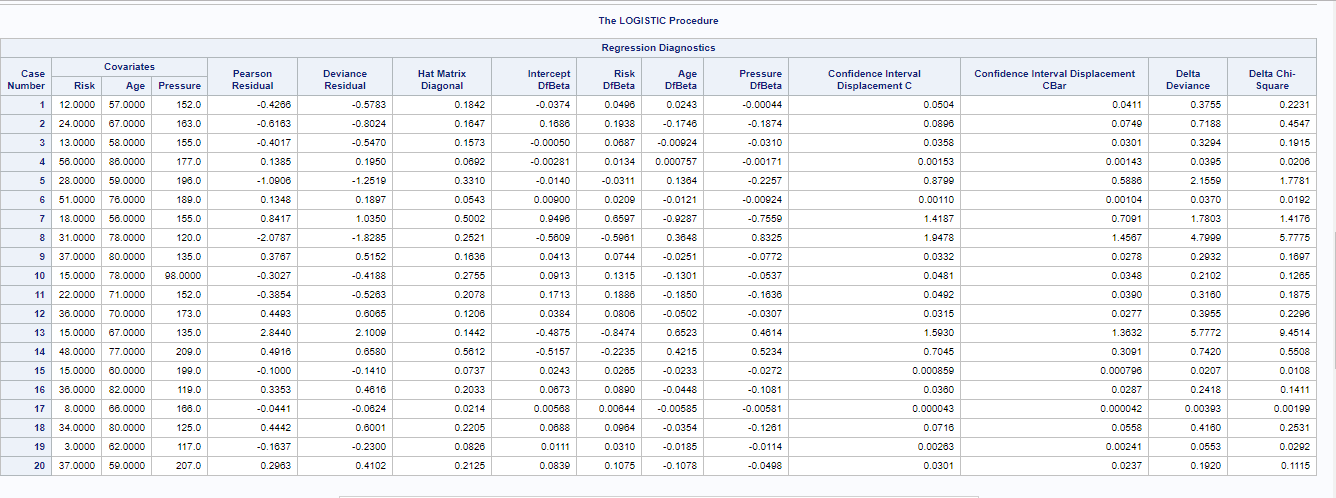
***Code Window***

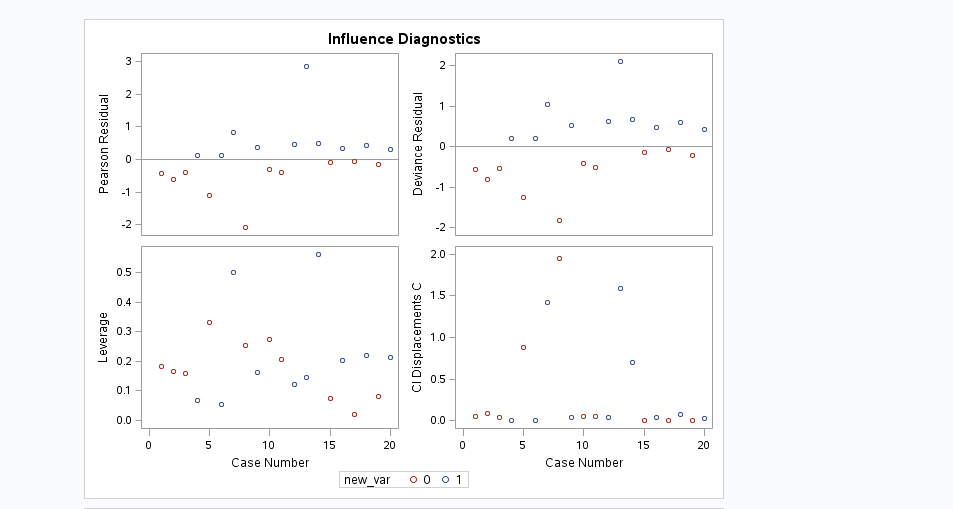


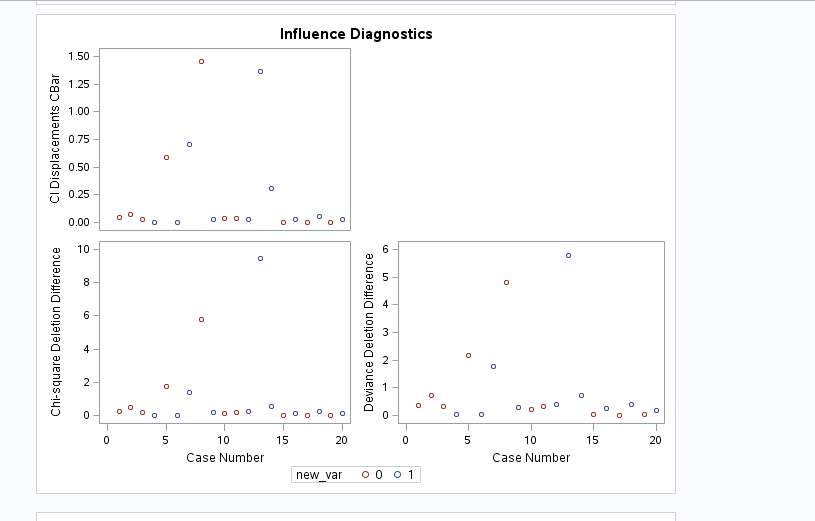
***Result Window***

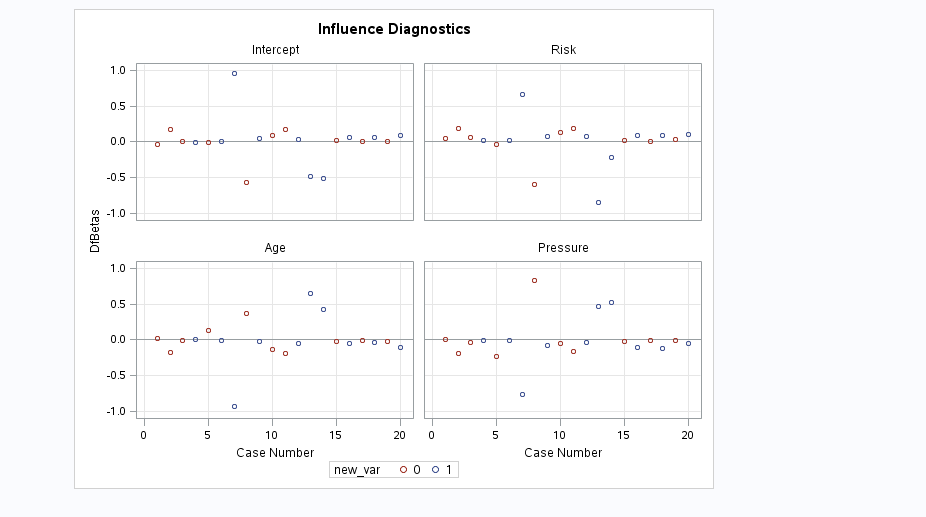




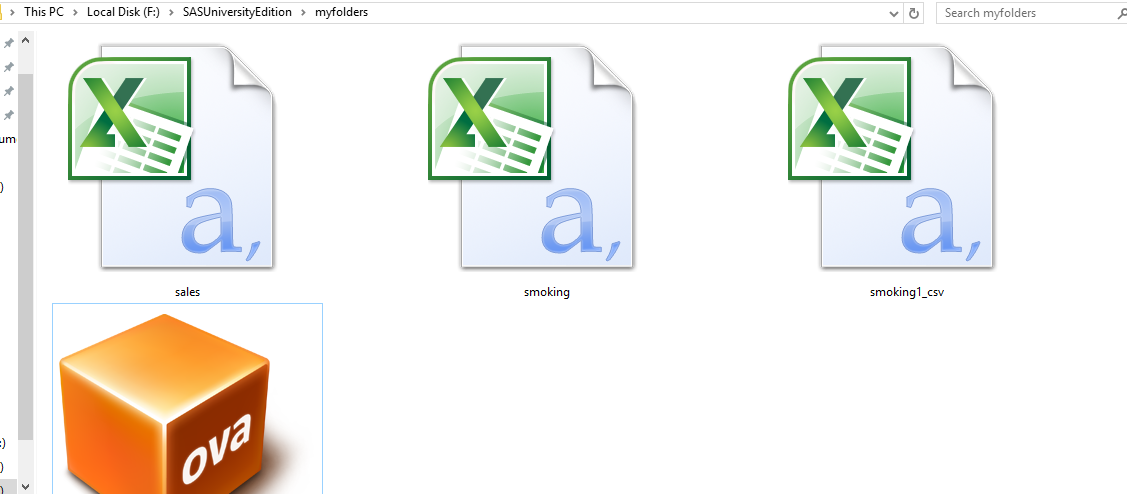




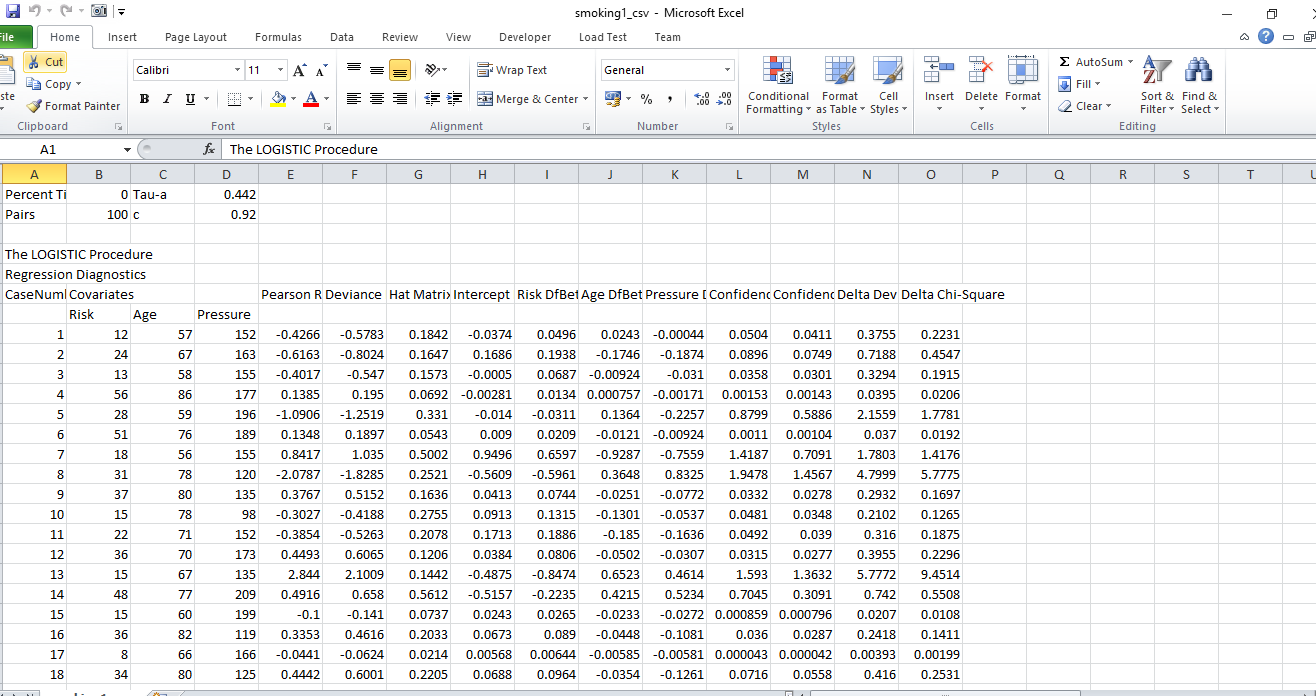




The output CSV file smoking1\_csv.csv is saved in the local drive.



The CSV output for Risk =12 is shown below



The result above shows a table named “Regression Diagnostics” shows the change in the output for the risk value =12. On the first row of the table i.e. for “case number 1”, we can see that deviance residual = -0.5783 & the risk value =12. Here we take into consideration the deviance residual since is the easiest residual to understand. The logistic regression can be understood in terms of fitting the function p = logit-1(Xβ) for known X in such a way as to minimise the total deviance residuals of all the data points.

In absolutes terms, the squared deviance of each data point is equal to (-2 times) the logarithms of the difference between its predicted probability logit-1(Xβ) and the complement of its actual value (1 for control, 0 for a case). A perfect fit for a point (which never occurs) gives a deviance of zero (as log(1) = 0). A poorly fitting has a large residual deviance as -2 times of the log of a very small number is a large number.

Here, for risk value = 12 we can find the deviance residual = -0.5783 in the CSV output.

Since the value is quite small, so here our fitting is not so poor. It seems a standard fitting.