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.

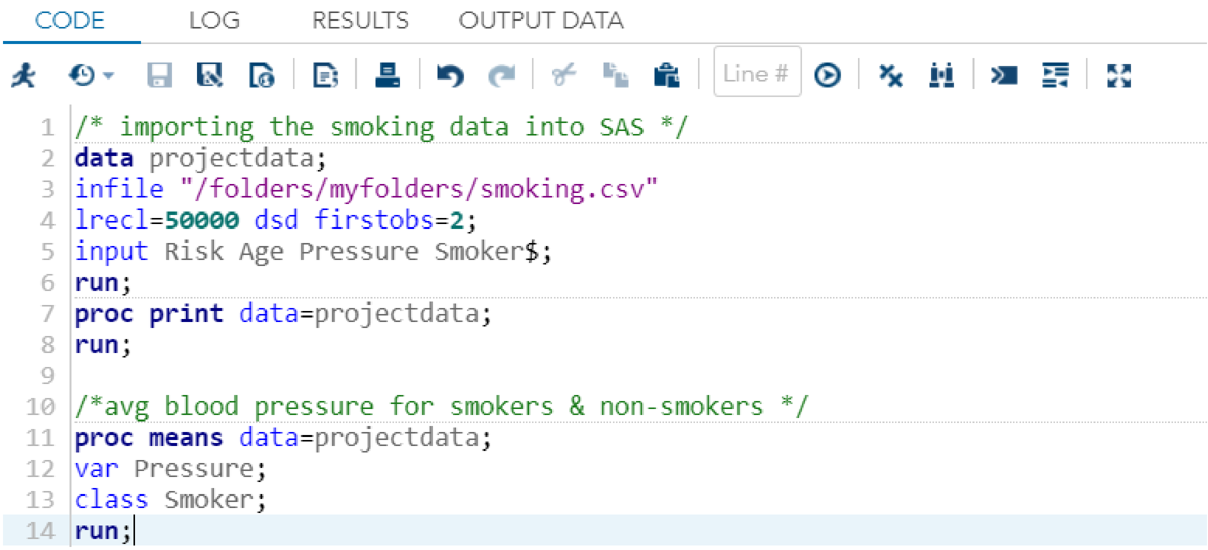
The question is whether the groups (Smokers & Non-smokers) differ in the means of blood pressure.

The required SAS codes are shown below.

1st Part

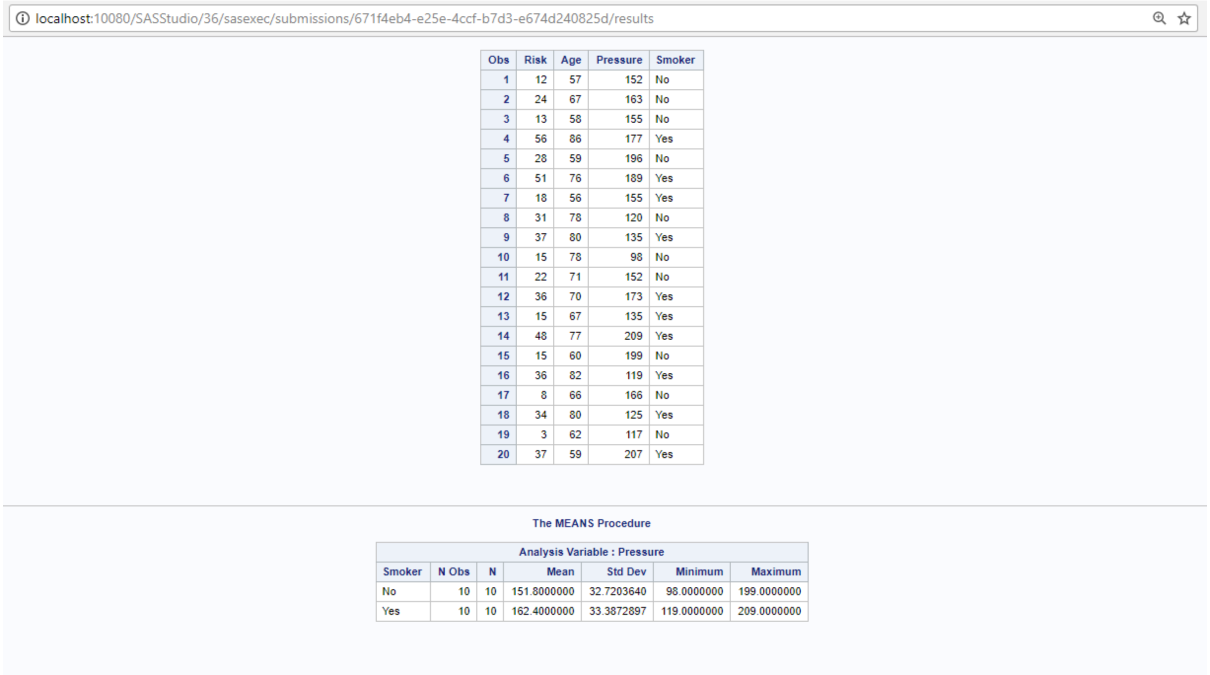
Code Window

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



```
/* importing the smoking data into SAS */
data projectdata;
infile "/folders/myfolders/smoking.csv"
lrecl=50000 dsd firstobs=2;
input Risk Age Pressure Smoker$;
run;
proc print data=projectdata;
run;

/*avg blood pressure for smokers & non-smokers */
proc means data=projectdata;
var Pressure;
class Smoker;
run;
```

Result Window



localhost:10080/SASStudio/36/sasexec/submissions/671f4eb4-e25e-4ccf-b7d3-e674d240825d/results

Obs	Risk	Age	Pressure	Smoker
1	12	57	152	No
2	24	67	163	No
3	13	58	155	No
4	56	86	177	Yes
5	28	59	196	No
6	51	76	189	Yes
7	18	56	155	Yes
8	31	78	120	No
9	37	80	135	Yes
10	15	78	98	No
11	22	71	152	No
12	36	70	173	Yes
13	15	67	135	Yes
14	48	77	209	Yes
15	15	60	199	No
16	36	82	119	Yes
17	8	66	166	No
18	34	80	125	Yes
19	3	62	117	No
20	37	59	207	Yes

The MEANS Procedure

Analysis Variable : Pressure						
Smoker	N Obs	N	Mean	Std Dev	Minimum	Maximum
No	10	10	151.8000000	32.7203640	98.0000000	199.0000000
Yes	10	10	162.4000000	33.3872897	119.0000000	209.0000000

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

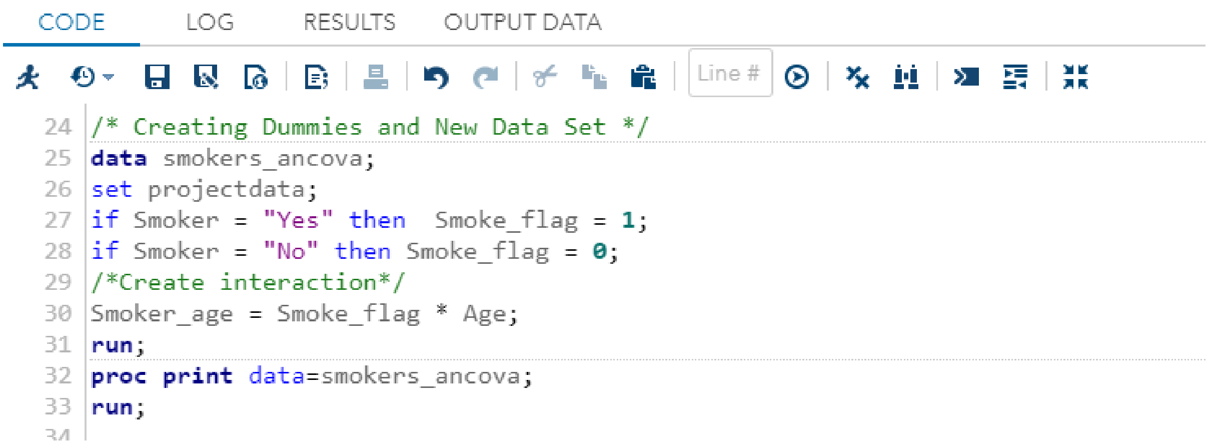
Here we have to predict the change in the dependent variable risk for change in age & blood pressure for two groups i.e., smokers & non-smokers. A general approach will be doing Regression Analysis. But here we have one binary independent variable (Smoker), two continuous variables (age & pressure) & 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.

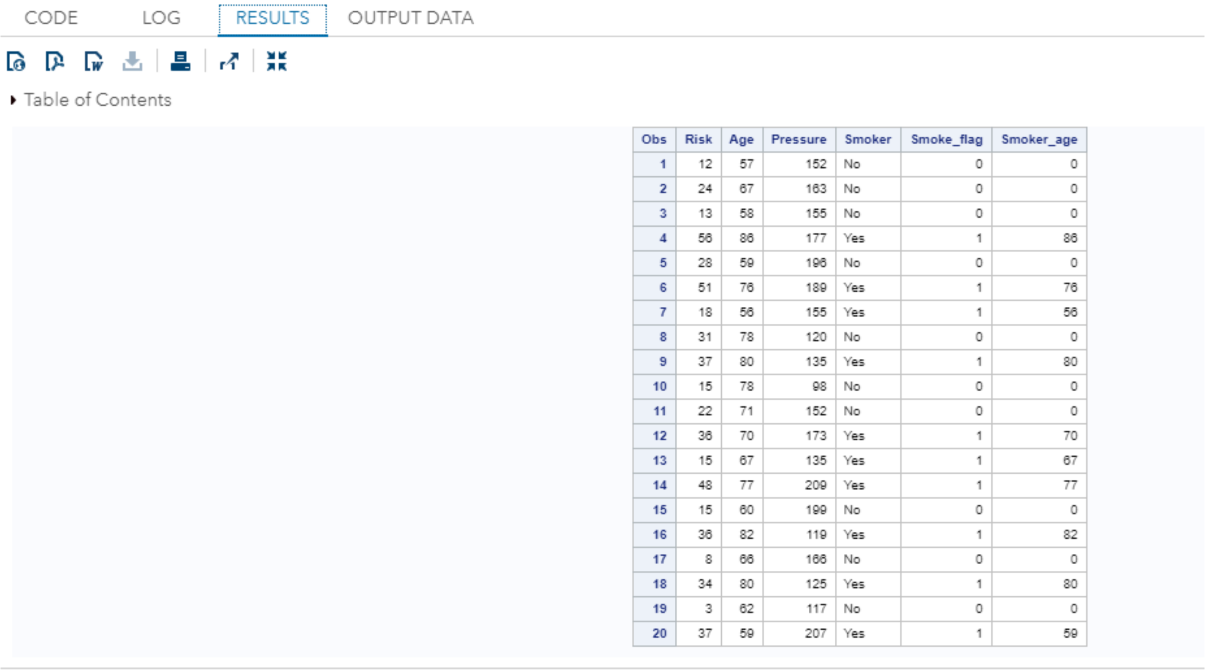
We create a dummy variable called “Smoke_flag” & another variable called “Smoker_age” which represents the interaction between Age & Smoker. These new variables will be used to fit the ANCOVA model.

Code Window



```
/* Creating Dummies and New Data Set */
data smokers_ancova;
set projectdata;
if Smoker = "Yes" then  Smoke_flag = 1;
if Smoker = "No" then Smoke_flag = 0;
/*Create interaction*/
Smoker_age = Smoke_flag * Age;
run;
proc print data=smokers_ancova;
run;
```

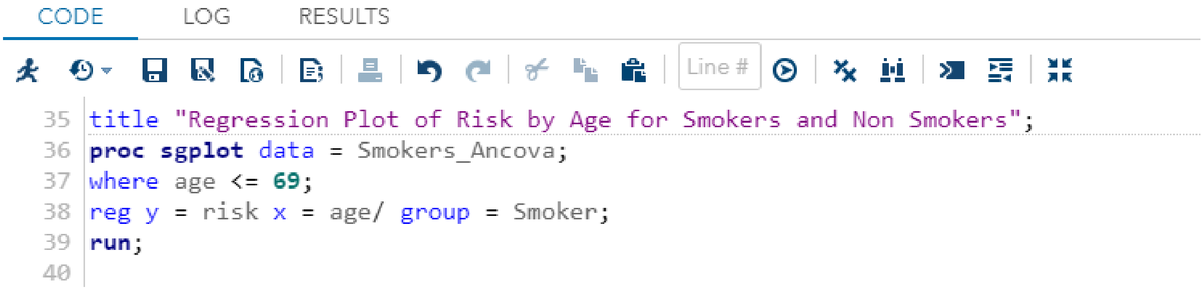
Result Window



Obs	Risk	Age	Pressure	Smoker	Smoke_flag	Smoker_age
1	12	57	152	No	0	0
2	24	67	163	No	0	0
3	13	58	155	No	0	0
4	56	86	177	Yes	1	86
5	28	59	196	No	0	0
6	51	76	189	Yes	1	76
7	18	56	155	Yes	1	56
8	31	78	120	No	0	0
9	37	80	135	Yes	1	80
10	15	78	98	No	0	0
11	22	71	152	No	0	0
12	36	70	173	Yes	1	70
13	15	67	135	Yes	1	67
14	48	77	209	Yes	1	77
15	15	60	199	No	0	0
16	36	82	119	Yes	1	82
17	8	66	166	No	0	0
18	34	80	125	Yes	1	80
19	3	62	117	No	0	0
20	37	59	207	Yes	1	59

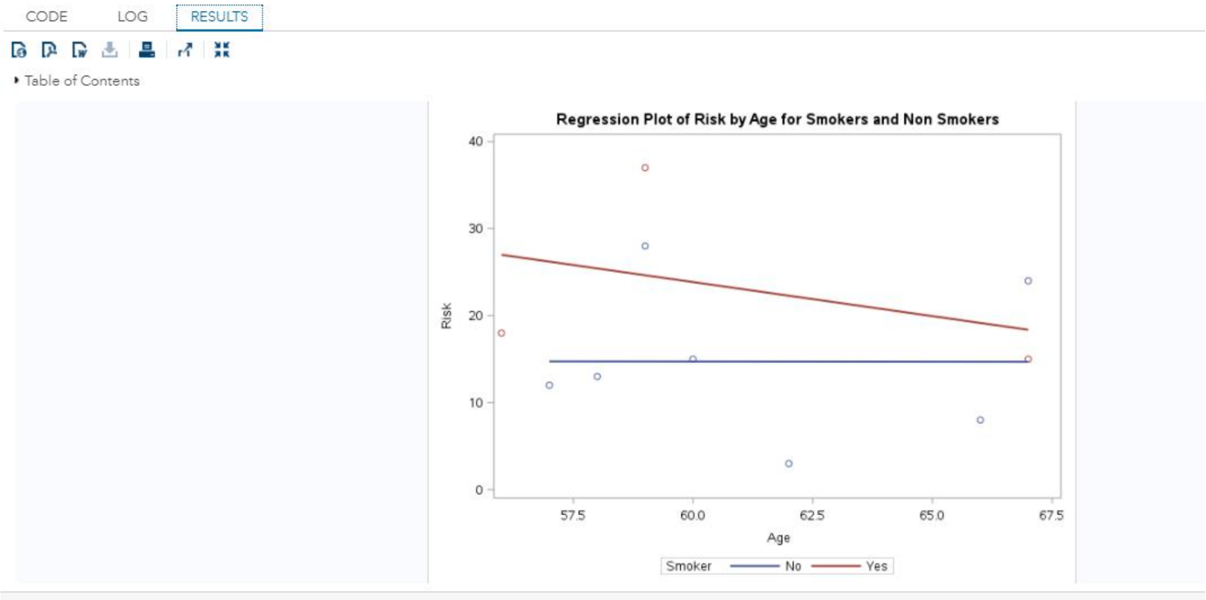
Now we generate a scatter plot with Risk as the Y & Age as the X, with separate regression line for Smokers & Non-smokers.

Code Window



```
title "Regression Plot of Risk by Age for Smokers and Non Smokers";
proc sgplot data = Smokers_Ancova;
where age <= 69;
reg y = risk x = age/ group = Smoker;
run;

```

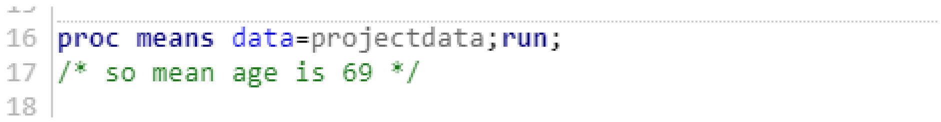
Result Window



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.

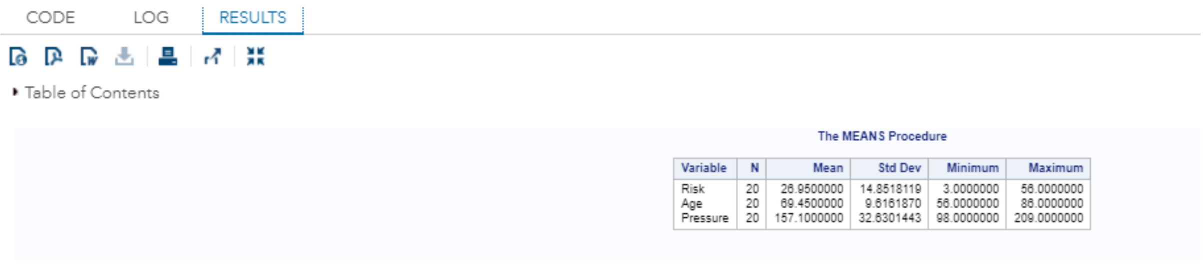
Now we fit an ANCOVA model, representing the relationship shown on the above graph. We include the main effects of Smoker (Smoker = "yes"), Age, their interaction & Pressure. Here we include the interaction created from the original AGE variable in this model.

Code Window



```
proc means data=projectdata;run;
/* so mean age is 69 */
```

Result Window



The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Risk	20	26.9500000	14.8518119	3.0000000	56.0000000
Age	20	69.4500000	9.6161870	56.0000000	86.0000000
Pressure	20	157.1000000	32.6301443	98.0000000	209.0000000

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.

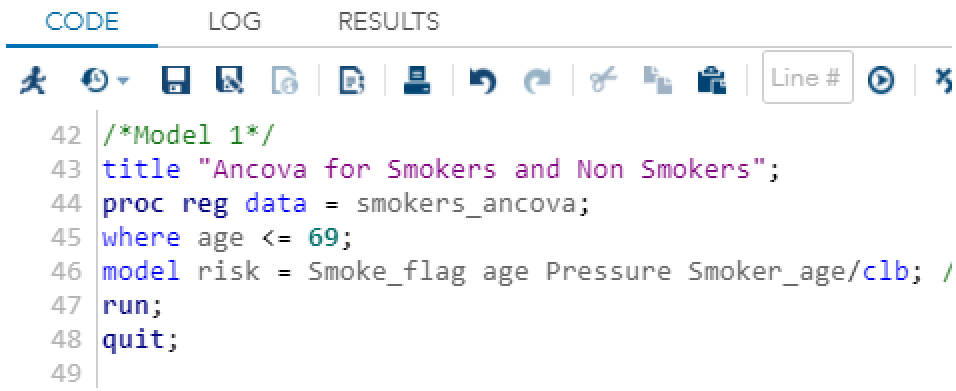
We use the clb option to get a 99.9% confidence interval for each of the parameters in the model. We have considered alpha as 0.1

The model that we are fitting is:

Risk = β_0+ β_1* Smoker$_i$ + β_2* Age$_i$ + β_3* Smoker_age$_i$ +

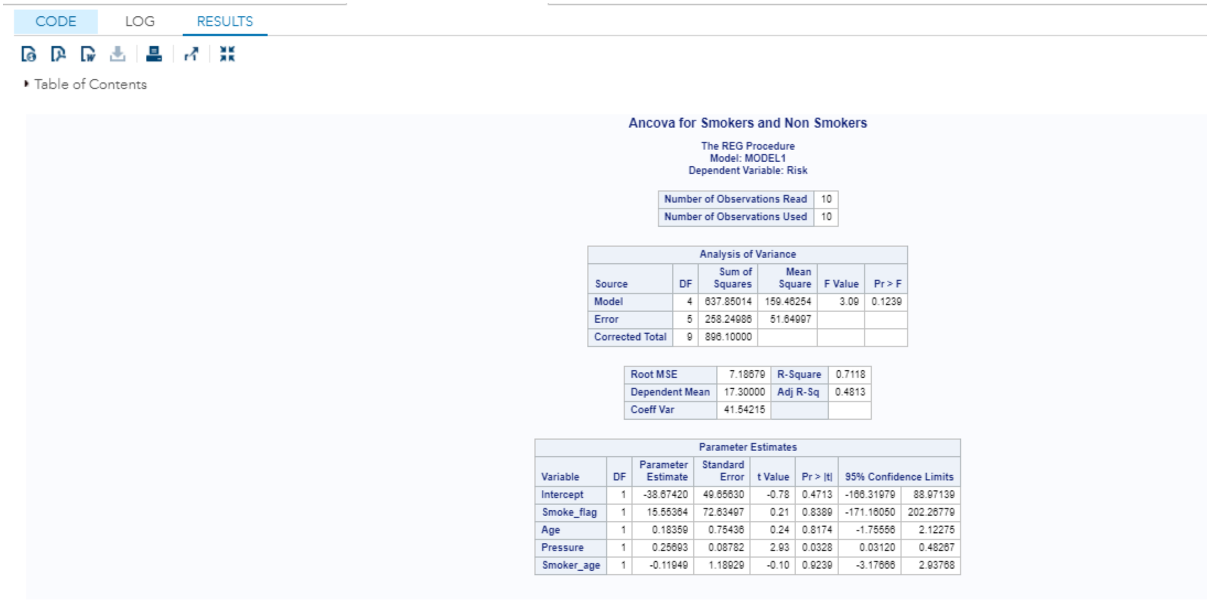
β_4* Pressure$_i$+ε_{ij}

Code Window



```
/*Model 1*/
title "Ancova for Smokers and Non Smokers";
proc reg data = smokers_ancova;
where age <= 69;
model risk = Smoke_flag age Pressure Smoker_age/clb; /
run;
quit;

```

Result Window



Ancova for Smokers and Non Smokers

The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	637.85014	159.46254	3.09	0.1239
Error	5	258.24986	51.64997		
Corrected Total	9	896.10000			

Root MSE	7.18679	R-Square	0.7118
Dependent Mean	17.30000	Adj R-Sq	0.4813
Coeff Var	41.54215		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-38.67420	49.65630	-0.78	0.4713	-166.31979	88.97139
Smoke_flag	1	15.55364	72.63497	0.21	0.8389	-171.16050	202.26779
Age	1	0.18359	0.75436	0.24	0.8174	-1.75556	2.12275
Pressure	1	0.25693	0.08782	2.93	0.0328	0.03120	0.48267
Smoker_age	1	-0.11949	1.18929	-0.10	0.9239	-3.17666	2.93768

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. R^2= 0.48.

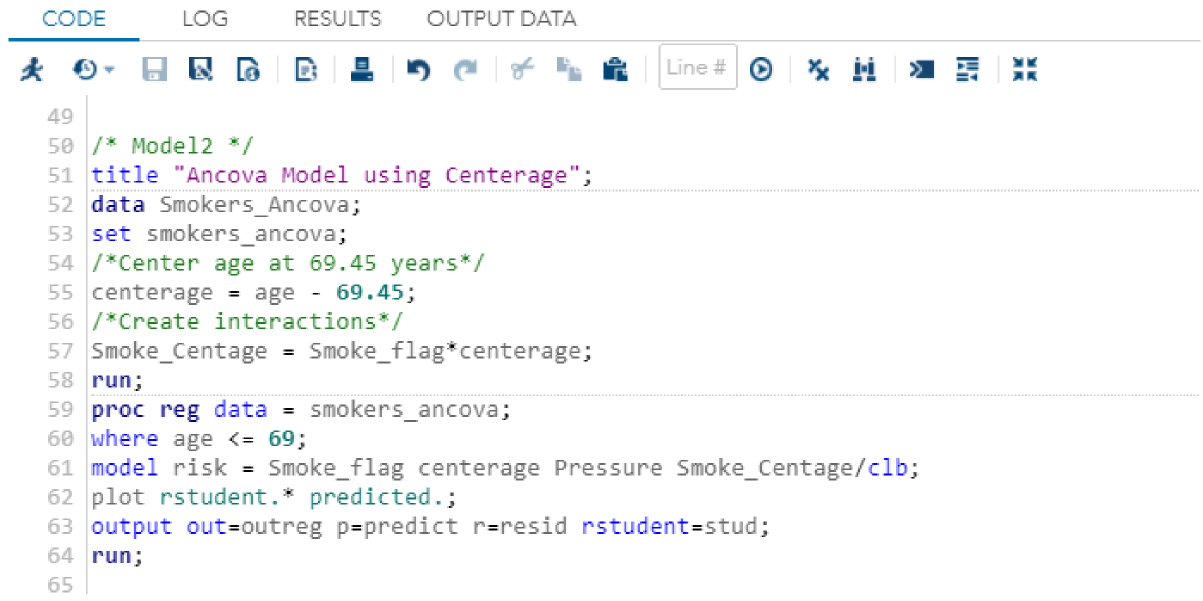
p>0.1, the model is near about significant. However, since the R^2 value is very close to 0. So this model is not predicting well.

We first look at the parameter estimate for the interaction term. The interaction term, β_4 (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:

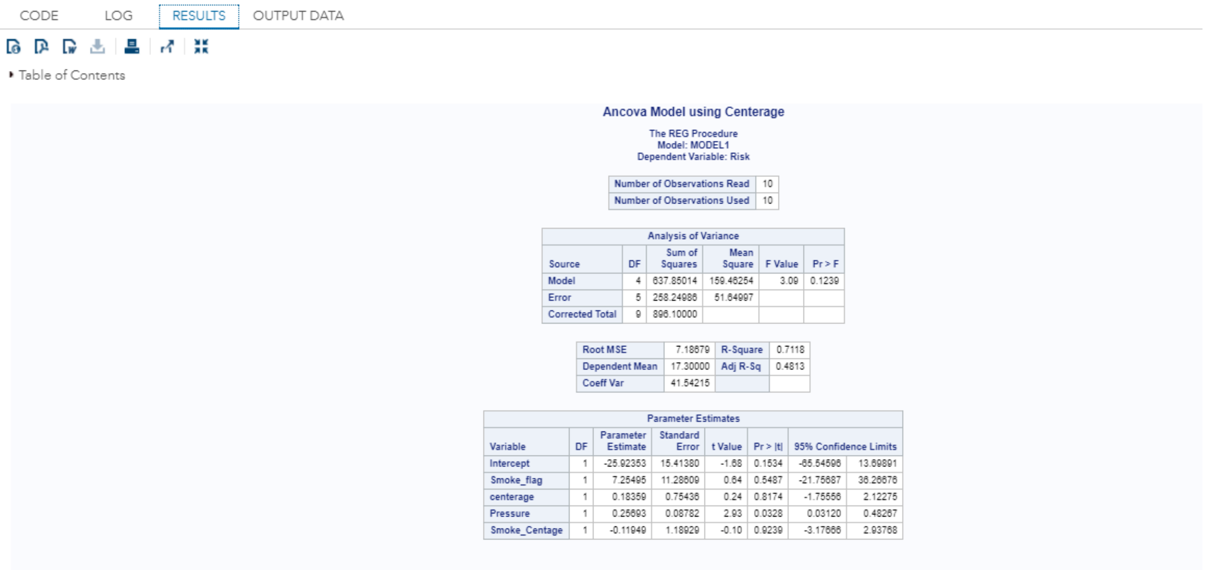
One way to help in the interpretation of the coefficients in a model like this is to center the continuous and then create an interaction term between centered age (centerage) & smokers. The new interaction will be called Smoke_Centage. Here we center age at 69.45 years, which is the approximate mean 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



```
/* Model2 */
title "Ancova Model using Centerage";
data Smokers_Ancova;
set smokers_ancova;
/*Center age at 69.45 years*/
centerage = age - 69.45;
/*Create interactions*/
Smoke_Centage = Smoke_flag*centerage;
run;
proc reg data = smokers_ancova;
where age <= 69;
model risk = Smoke_flag centerage Pressure Smoke_Centage/clb;
plot rstudent.* predicted.;
output out=outreg p=predict r=resid rstudent=stud;
run;
```

Result Window



Ancova Model using Centerage

The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	637.85014	159.46254	3.09	0.1239
Error	5	258.24986	51.64997		
Corrected Total	9	896.10000			

Root MSE	7.18679	R-Square	0.7118
Dependent Mean	17.30000	Adj R-Sq	0.4813
Coeff Var	41.54215		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-25.92353	15.41380	-1.68	0.1534	-65.54596	13.69891
Smoke_flag	1	7.25495	11.28609	0.64	0.5487	-21.75687	36.26676
centerage	1	0.18359	0.75436	0.24	0.8174	-1.75556	2.12275
Pressure	1	0.25693	0.08782	2.93	0.0328	0.03120	0.48267
Smoke_Centage	1	-0.11949	1.18929	-0.10	0.9239	-3.17666	2.93768

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.

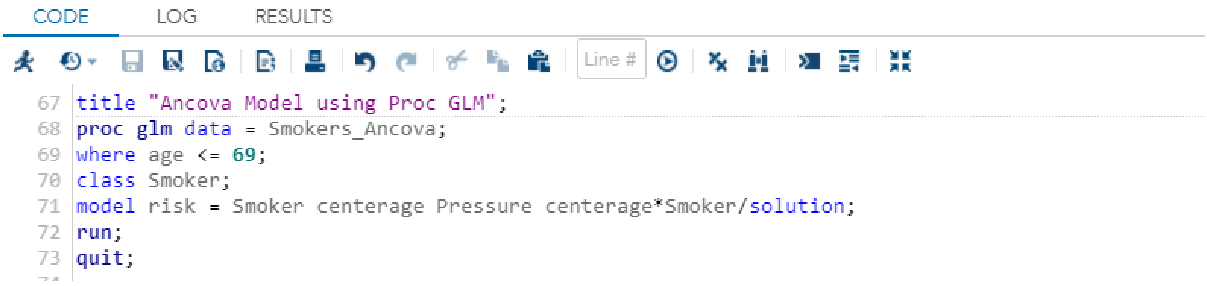
The interaction term which represents the difference in slope for smokers & non-smokers (estimated to be -0.11949) is the same as in the previous model. The coefficient for “centerage” is 0.18359 which represents the slope for the reference category (non-smokers), is the same as the coefficient for AGE 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:

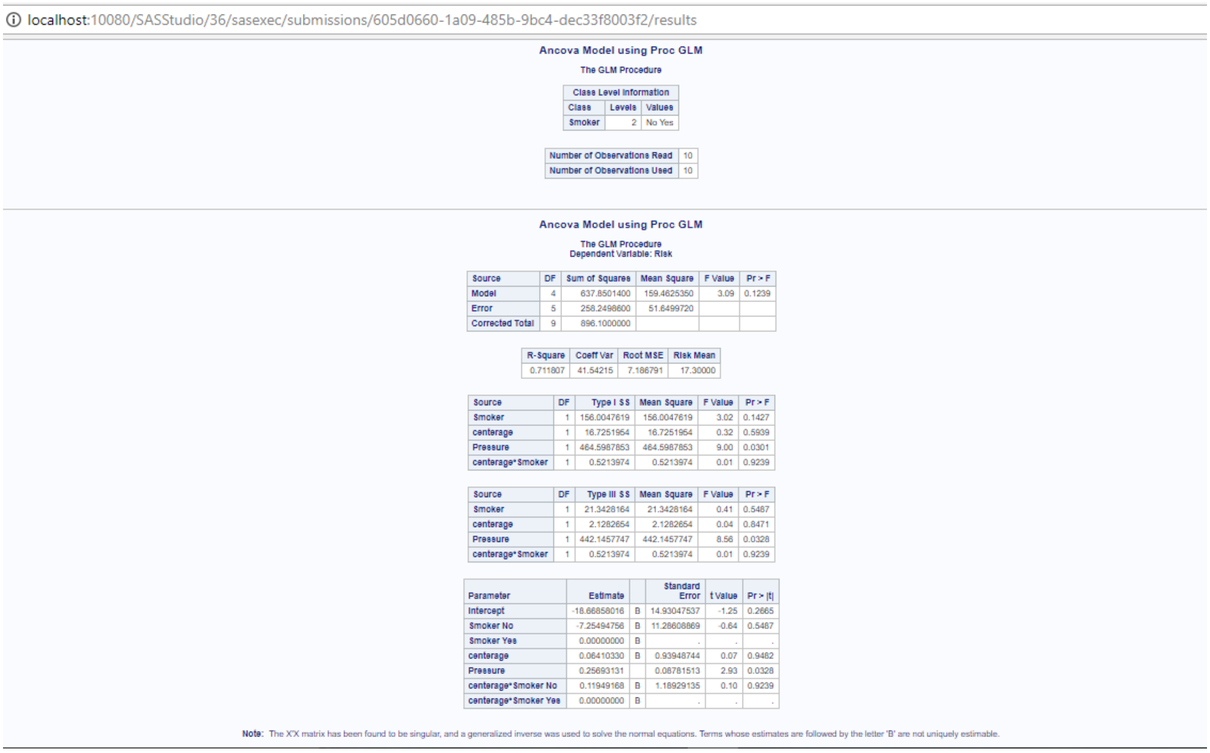
We now refit the model using centerage & smokers as predictors, but using Proc GLM. The advantage of using this procedure is that you don’t need to create dummy variables for our categorical predictors, and the interaction terms do not need to be created in advance. The categorical variable “Smoker” 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



```
title "Ancova Model using Proc GLM";
proc glm data = Smokers_Ancova;
where age <= 69;
class Smoker;
model risk = Smoker centerage Pressure centerage*Smoker/solution;
run;
quit;
```

Result Window


localhost:10080/SASStudio/36/sasexec/submissions/605d0660-1a09-485b-9bc4-dec33f8003f2/results

Ancova Model using Proc GLM

The GLM Procedure

Class Level Information		
Class	Levels	Values
Smoker	2	No Yes

Number of Observations Read	10
Number of Observations Used	10

Ancova Model using Proc GLM

The GLM Procedure
Dependent Variable: Risk

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	637.8501400	159.4625350	3.09	0.1239
Error	5	258.2498600	51.6499720		
Corrected Total	9	896.1000000			

R-Square	Coeff Var	Root MSE	Risk Mean
0.711807	41.54215	7.186791	17.30000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Smoker	1	156.0047619	156.0047619	3.02	0.1427
centerage	1	16.7251954	16.7251954	0.32	0.5939
Pressure	1	464.5987853	464.5987853	9.00	0.0301
centerage*Smoker	1	0.5213974	0.5213974	0.01	0.9239

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Smoker	1	21.3428164	21.3428164	0.41	0.5487
centerage	1	2.1282654	2.1282654	0.04	0.8471
Pressure	1	442.1457747	442.1457747	8.56	0.0328
centerage*Smoker	1	0.5213974	0.5213974	0.01	0.9239

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-18.66858016	B	14.93047537	-1.25	0.2665
Smoker No	-7.25494756	B	11.28608869	-0.64	0.5487
Smoker Yes	0.00000000	B	.	.	.
centerage	0.06410330	B	0.93948744	0.07	0.9482
Pressure	0.25693131		0.08781513	2.93	0.0328
centerage*Smoker No	0.11949168	B	1.18929135	0.10	0.9239
centerage*Smoker Yes	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

From the above result we can see that the coefficient of determination i.e., R^2 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.

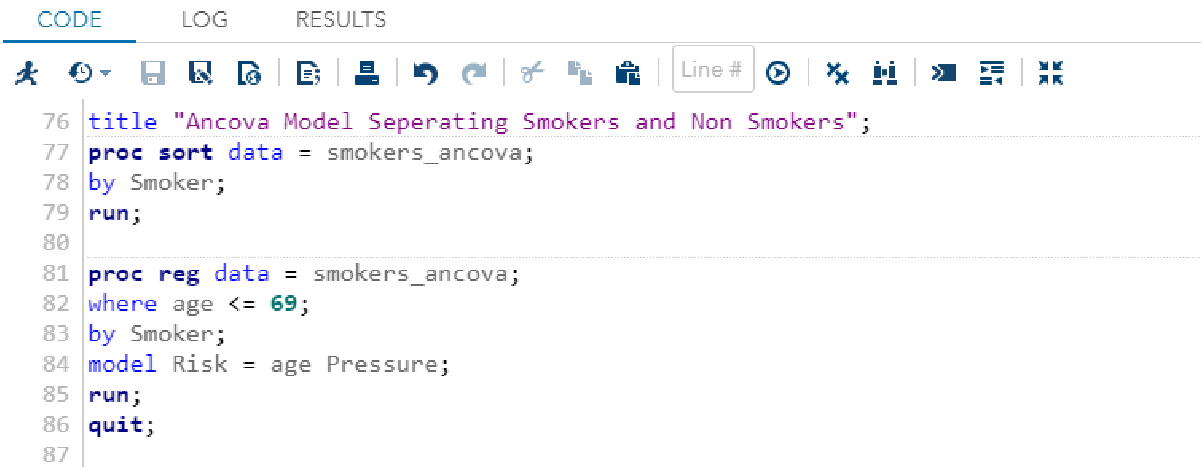
Notice that we get the same parameter estimates using Proc GLM as we did in Proc Reg. By default, Proc GLM over parameterizes the model, including a parameter for each level of Smoker. The parameter estimate for the highest level of Smoking is set to zero, which has the effect in this case of making 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.

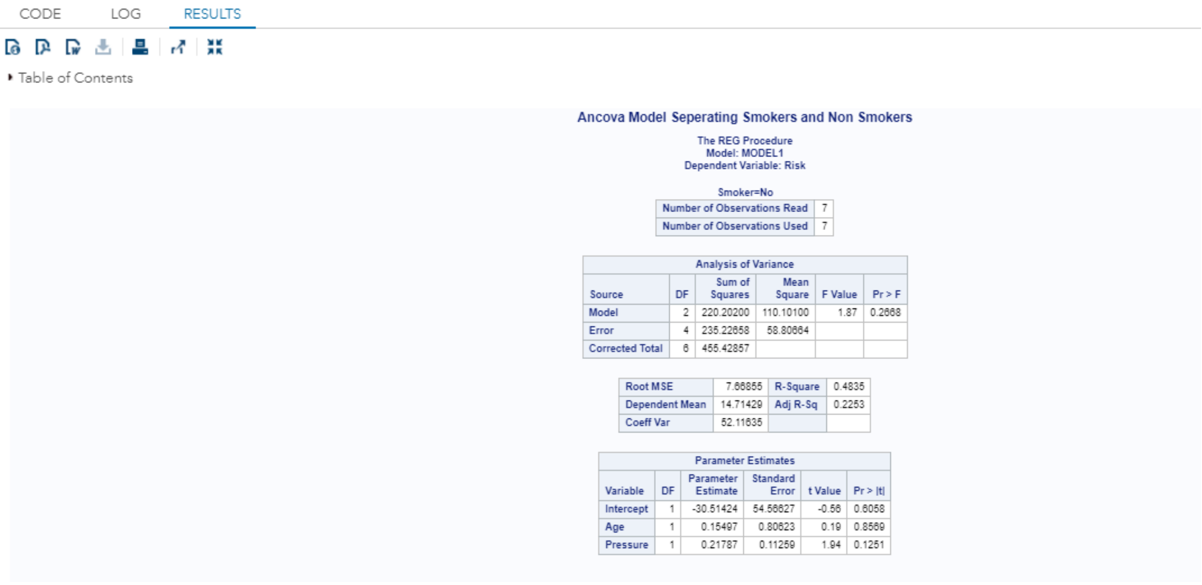
Code Window



```
title "Ancova Model Seperating Smokers and Non Smokers";
proc sort data = smokers_ancova;
by Smoker;
run;

proc reg data = smokers_ancova;
where age <= 69;
by Smoker;
model Risk = age Pressure;
run;
quit;
```

Result Window

Output for “Smoker”=no



Ancova Model Seperating Smokers and Non Smokers

The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=No

Number of Observations Read	7
Number of Observations Used	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	220.20200	110.10100	1.87	0.2668
Error	4	235.22658	58.80664		
Corrected Total	6	455.42857			

Root MSE	7.66855	R-Square	0.4835
Dependent Mean	14.71429	Adj R-Sq	0.2253
Coeff Var	52.11635		

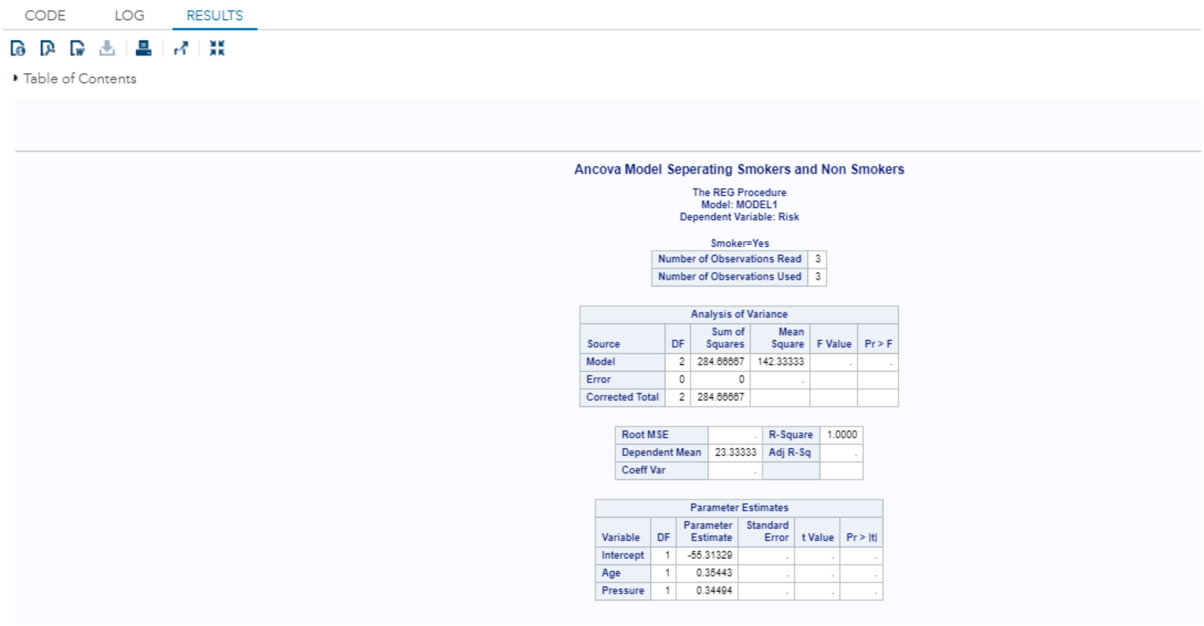
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-30.51424	54.56627	-0.56	0.6058
Age	1	0.15497	0.80623	0.19	0.8569
Pressure	1	0.21787	0.11259	1.94	0.1251

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



Ancova Model Seperating Smokers and Non Smokers

The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=Yes

Number of Observations Read	3
Number of Observations Used	3

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	284.66667	142.33333	.	.
Error	0	0	.		
Corrected Total	2	284.66667			

Root MSE	.	R-Square	1.0000
Dependent Mean	23.33333	Adj R-Sq	.
Coeff Var	.		

Parameter Estimates

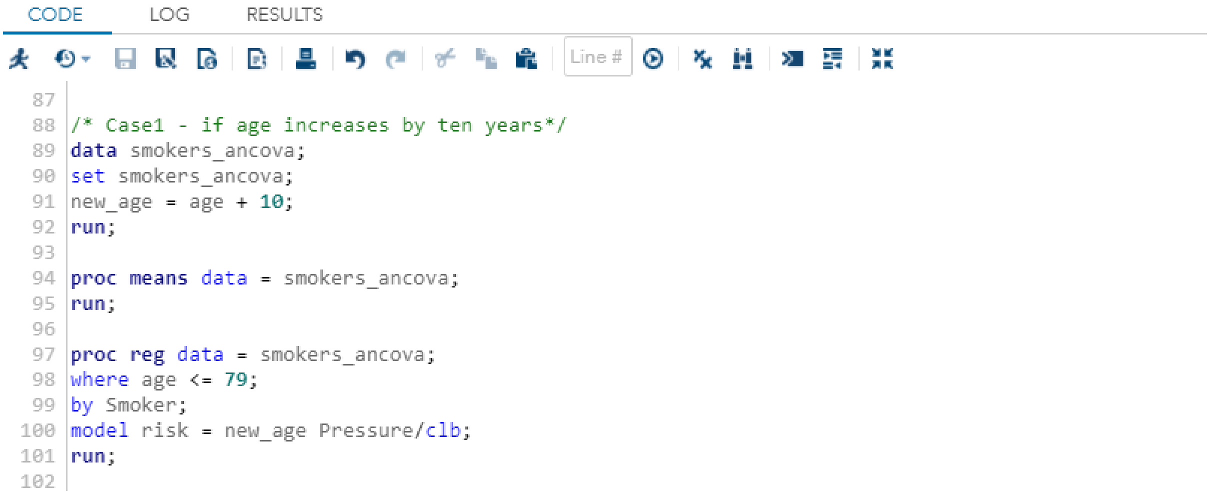
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-55.31329	.	.	.
Age	1	0.35443	.	.	.
Pressure	1	0.34494	.	.	.

In the above output the value of R^2 is 1.0. An R^2 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.

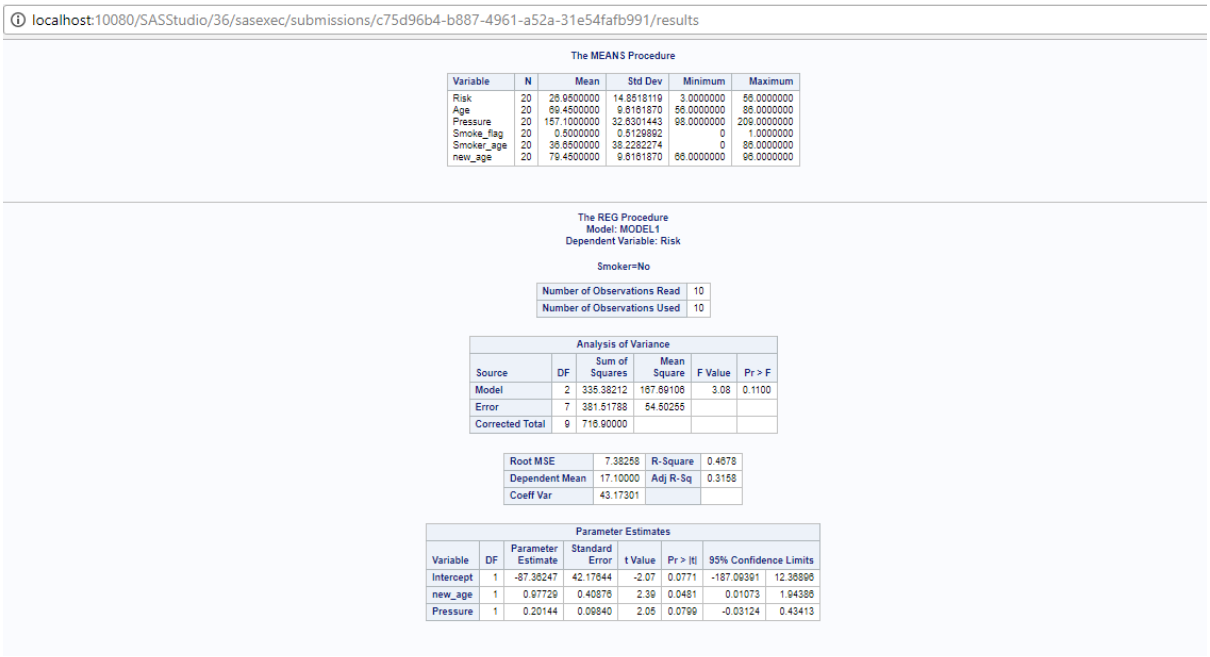
Code Window



```

/* Case1 - if age increases by ten years*/
data smokers_ancova;
set smokers_ancova;
new_age = age + 10;
run;

proc means data = smokers_ancova;
run;

proc reg data = smokers_ancova;
where age <= 79;
by Smoker;
model risk = new_age Pressure/clb;
run;

```

Result Window

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



localhost:10080/SASStudio/36/sasexec/submissions/c75d96b4-b887-4961-a52a-31e54fafb991/results

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Risk	20	26.9500000	14.8518119	3.0000000	56.0000000
Age	20	69.4500000	9.6161870	56.0000000	86.0000000
Pressure	20	157.1000000	32.6301443	98.0000000	209.0000000
Smoke_flag	20	0.5000000	0.5129892	0	1.0000000
Smoker_age	20	36.6500000	38.2282274	0	86.0000000
new_age	20	79.4500000	9.6161870	66.0000000	96.0000000

The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=No

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	335.38212	167.69106	3.08	0.1100
Error	7	381.51788	54.50255		
Corrected Total	9	716.90000			

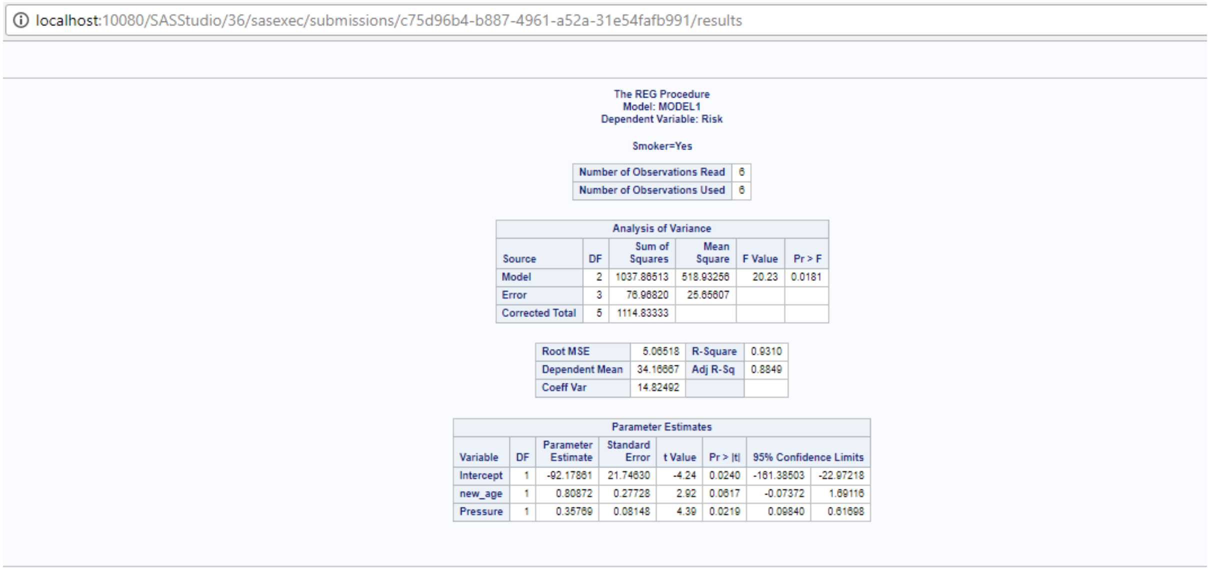
Root MSE	7.38258	R-Square	0.4678
Dependent Mean	17.10000	Adj R-Sq	0.3158
Coeff Var	43.17301		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-87.36247	42.17644	-2.07	0.0771	-187.09391	12.36896
new_age	1	0.97729	0.40876	2.39	0.0481	0.01073	1.94386
Pressure	1	0.20144	0.09840	2.05	0.0799	-0.03124	0.43413

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



localhost:10080/SASStudio/36/sasexec/submissions/c75d96b4-b887-4961-a52a-31e54fafb991/results

The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=Yes

Number of Observations Read	6
Number of Observations Used	6

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1037.86513	518.93256	20.23	0.0181
Error	3	76.96820	25.65607		
Corrected Total	5	1114.83333			

Root MSE	5.06518	R-Square	0.9310
Dependent Mean	34.16667	Adj R-Sq	0.8849
Coeff Var	14.82492		

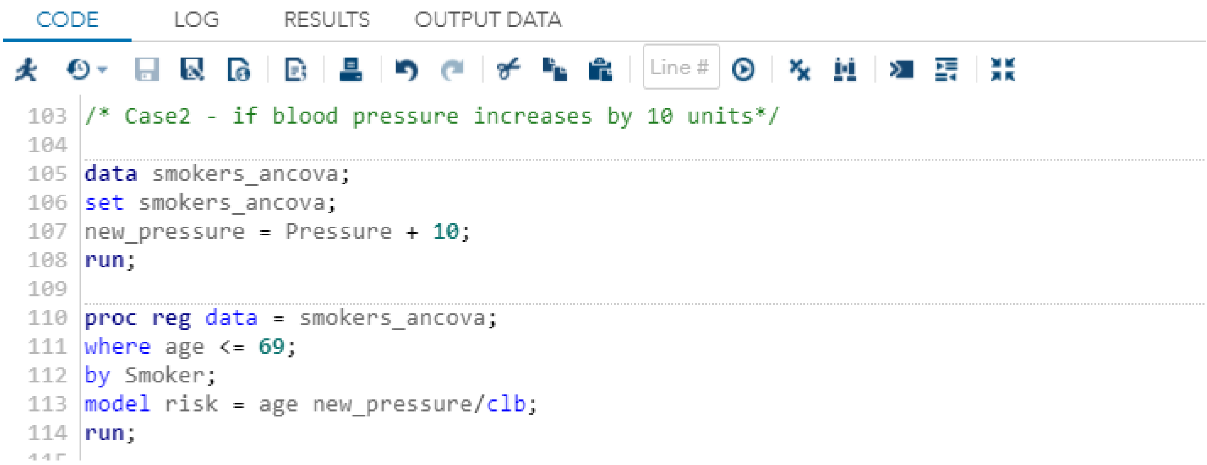
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-92.17861	21.74630	-4.24	0.0240	-161.38503	-22.97218
new_age	1	0.80872	0.27728	2.92	0.0617	-0.07372	1.69116
Pressure	1	0.35769	0.08148	4.39	0.0219	0.09840	0.61698

The model is statistically insignificant as the p values are more than alpha =0.1. But the R^2 value = 0.93 i.e., high R^2 value indicates predicted values are very close to the regression line.

Blood pressure is increased by 10 units:-

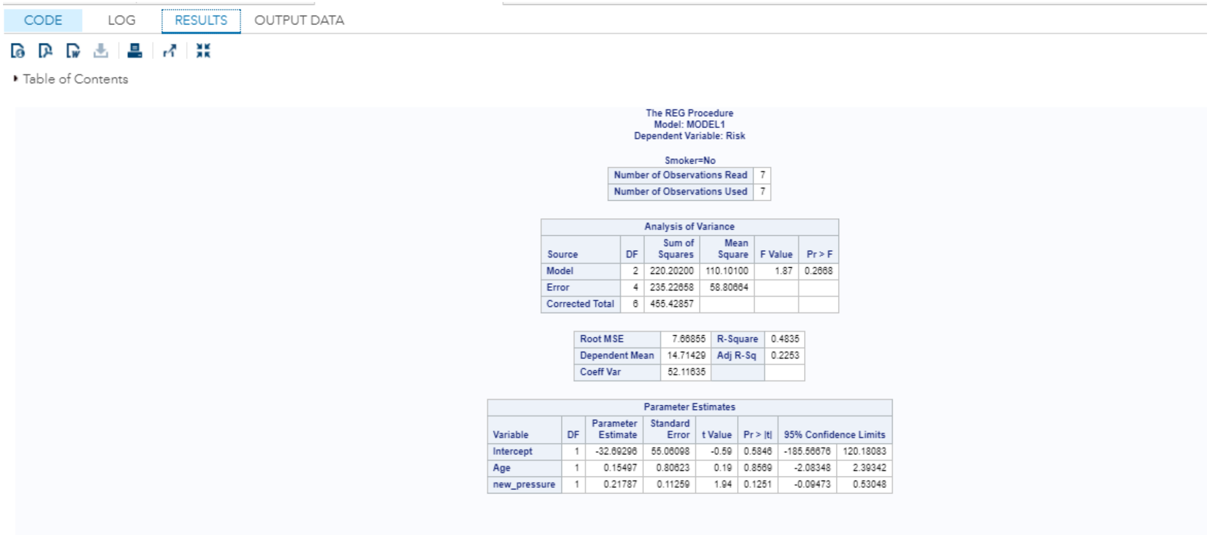
Code Window



```
/* Case2 - if blood pressure increases by 10 units*/

data smokers_ancova;
set smokers_ancova;
new_pressure = Pressure + 10;
run;

proc reg data = smokers_ancova;
where age <= 69;
by Smoker;
model risk = age new_pressure/clb;
run;
```

Result Window

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



The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=No

Number of Observations Read	7
Number of Observations Used	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	220.20200	110.10100	1.87	0.2668
Error	4	235.22658	58.80664		
Corrected Total	6	455.42857			

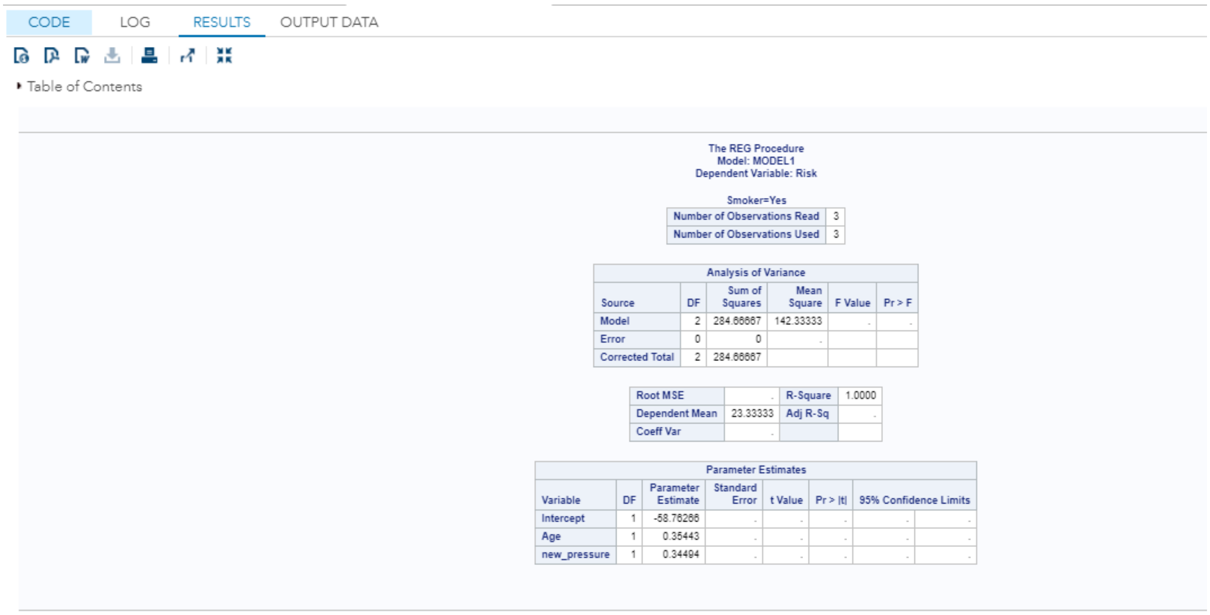
Root MSE	7.66855	R-Square	0.4835
Dependent Mean	14.71429	Adj R-Sq	0.2253
Coeff Var	52.11635		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-32.69296	55.06098	-0.59	0.5846	-185.56676	120.18083
Age	1	0.15497	0.80623	0.19	0.8569	-2.08348	2.39342
new_pressure	1	0.21787	0.11259	1.94	0.1251	-0.09473	0.53048

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



The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=Yes

Number of Observations Read	3
Number of Observations Used	3

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	284.66667	142.33333	.	.
Error	0	0	.		
Corrected Total	2	284.66667			

Root MSE	.	R-Square	1.0000
Dependent Mean	23.33333	Adj R-Sq	.
Coeff Var	.		

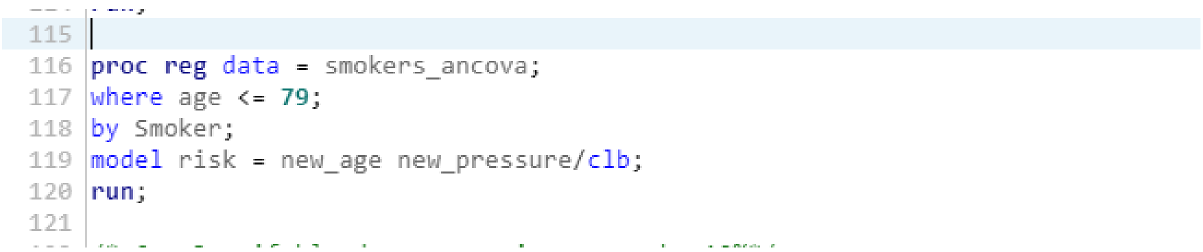
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-58.76266	.	.	.	.	.
Age	1	0.35443	.	.	.	.	.
new_pressure	1	0.34494	.	.	.	.	.

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.

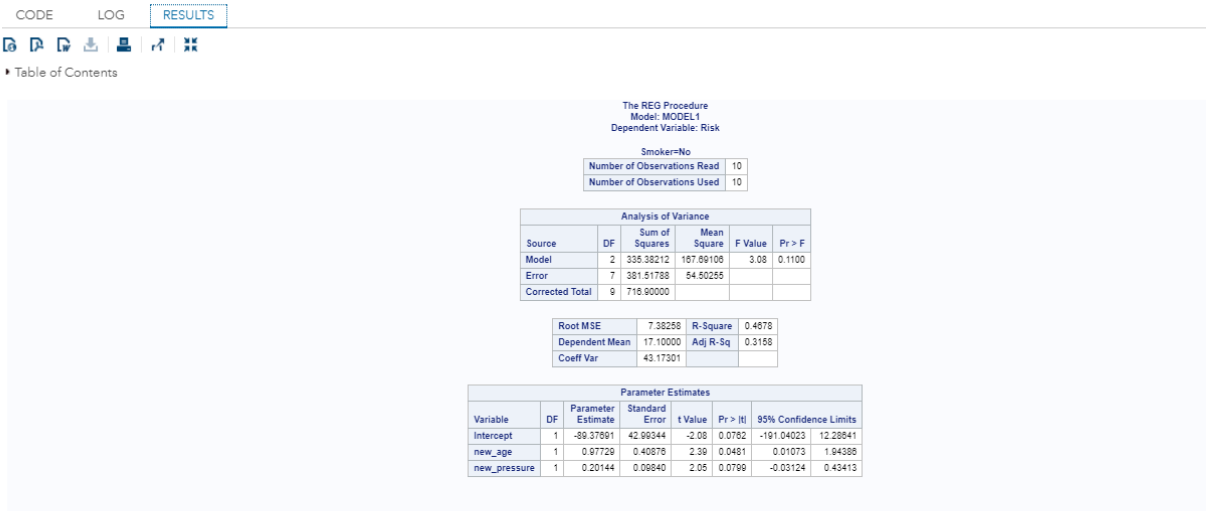
Code Window



```

proc reg data = smokers_ancova;
where age <= 79;
by Smoker;
model risk = new_age new_pressure/clb;
run;

```

Result Window

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



The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=No

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	335.38212	167.69106	3.08	0.1100
Error	7	381.51788	54.50255		
Corrected Total	9	716.90000			

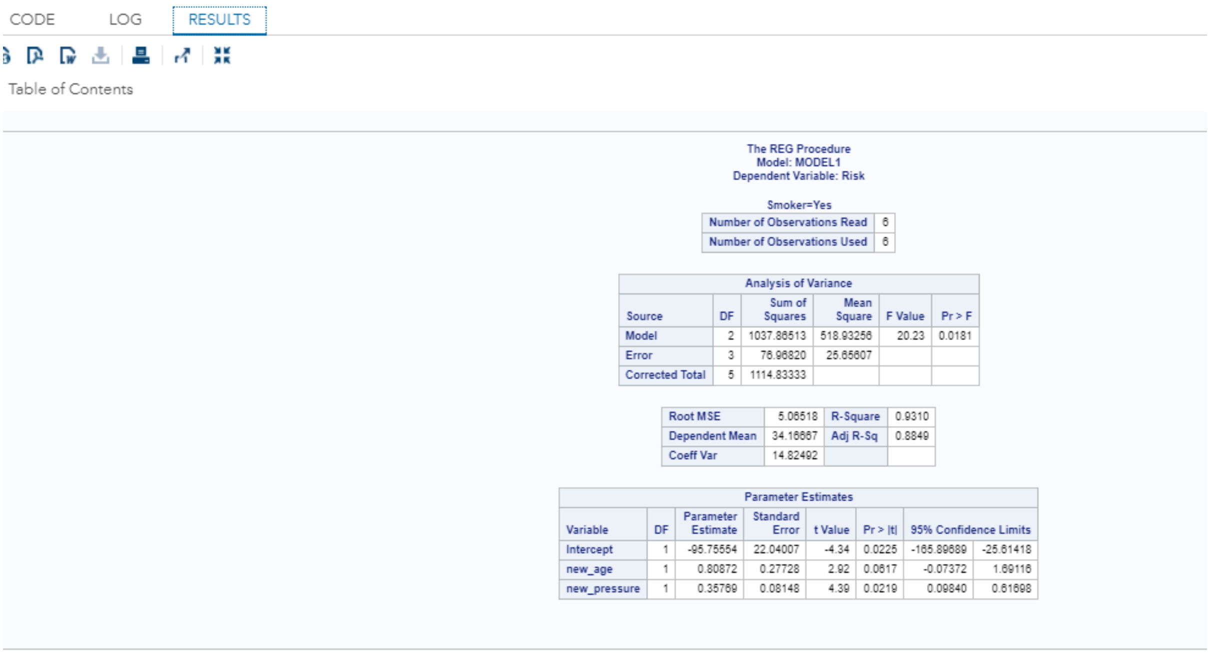
Root MSE	7.38258	R-Square	0.4678
Dependent Mean	17.10000	Adj R-Sq	0.3158
Coeff Var	43.17301		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-89.37691	42.99344	-2.08	0.0762	-191.04023	12.28641
new_age	1	0.97729	0.40876	2.39	0.0481	0.01073	1.94386
new_pressure	1	0.20144	0.09840	2.05	0.0799	-0.03124	0.43413

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)



The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=Yes

Number of Observations Read	6
Number of Observations Used	6

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1037.86513	518.93256	20.23	0.0181
Error	3	76.96820	25.65607		
Corrected Total	5	1114.83333			

Root MSE	5.06518	R-Square	0.9310
Dependent Mean	34.16667	Adj R-Sq	0.8849
Coeff Var	14.82492		

Parameter Estimates

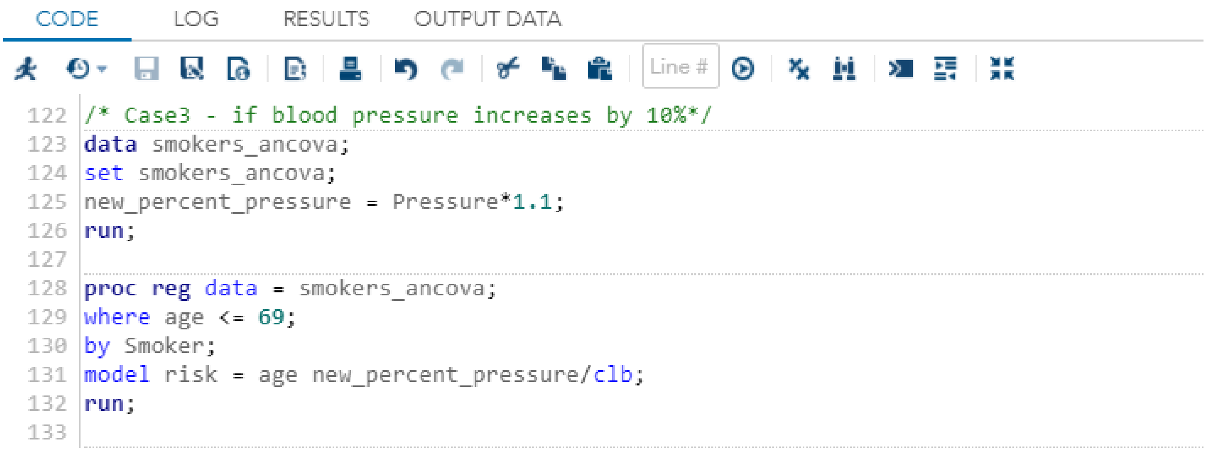
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-95.75554	22.04007	-4.34	0.0225	-165.89689	-25.61418
new_age	1	0.80872	0.27728	2.92	0.0617	-0.07372	1.69116
new_pressure	1	0.35769	0.08148	4.39	0.0219	0.09840	0.61698

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%.

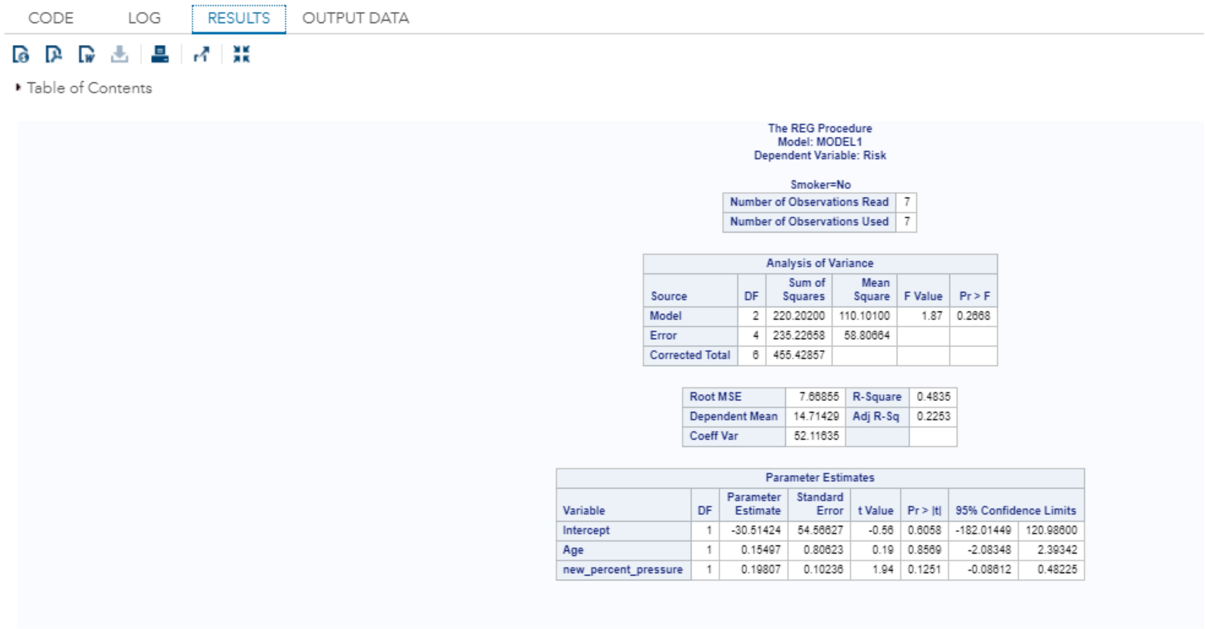
Code Window



```
/* Case3 - if blood pressure increases by 10%*/
data smokers_ancova;
set smokers_ancova;
new_percent_pressure = Pressure*1.1;
run;

proc reg data = smokers_ancova;
where age <= 69;
by Smoker;
model risk = age new_percent_pressure/clb;
run;

```

Result Window

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



The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=No

Number of Observations Read	7
Number of Observations Used	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	220.20200	110.10100	1.87	0.2668
Error	4	235.22658	58.80664		
Corrected Total	6	455.42857			

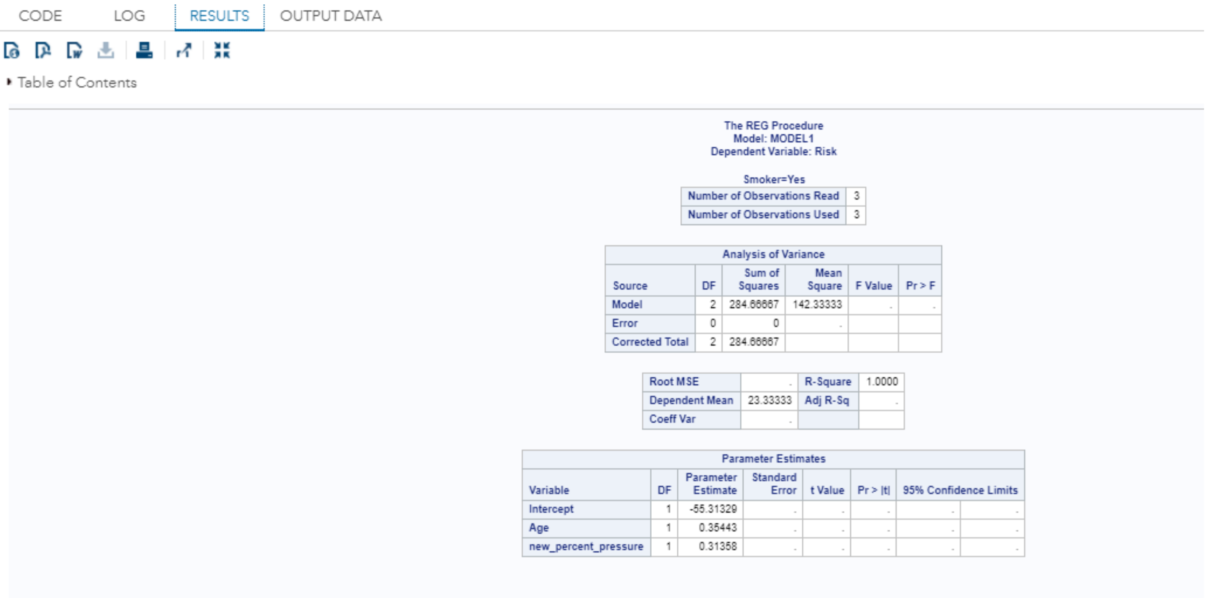
Root MSE	7.66855	R-Square	0.4835
Dependent Mean	14.71429	Adj R-Sq	0.2253
Coeff Var	52.11635		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-30.51424	54.56627	-0.56	0.6058	-182.01449	120.98600
Age	1	0.15497	0.80623	0.19	0.8569	-2.08348	2.39342
new_percent_pressure	1	0.19807	0.10236	1.94	0.1251	-0.08612	0.48225

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



The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=Yes

Number of Observations Read	3
Number of Observations Used	3

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	284.66667	142.33333	.	.
Error	0	0	.		
Corrected Total	2	284.66667			

Root MSE	.	R-Square	1.0000
Dependent Mean	23.33333	Adj R-Sq	.
Coeff Var	.		

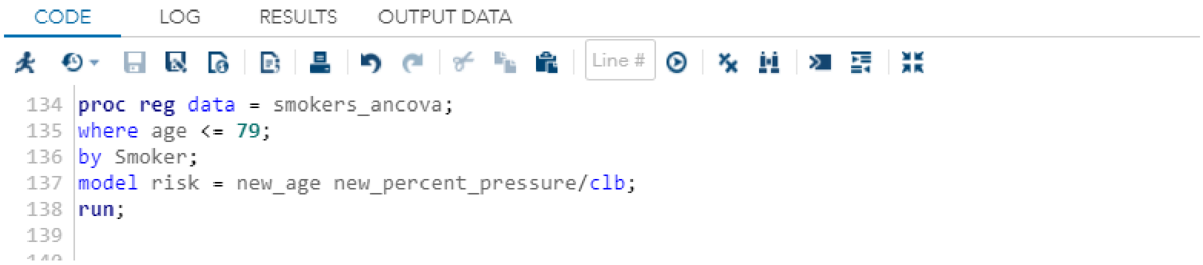
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-55.31329	.	.	.	.	.
Age	1	0.35443	.	.	.	.	.
new_percent_pressure	1	0.31358	.	.	.	.	.

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.

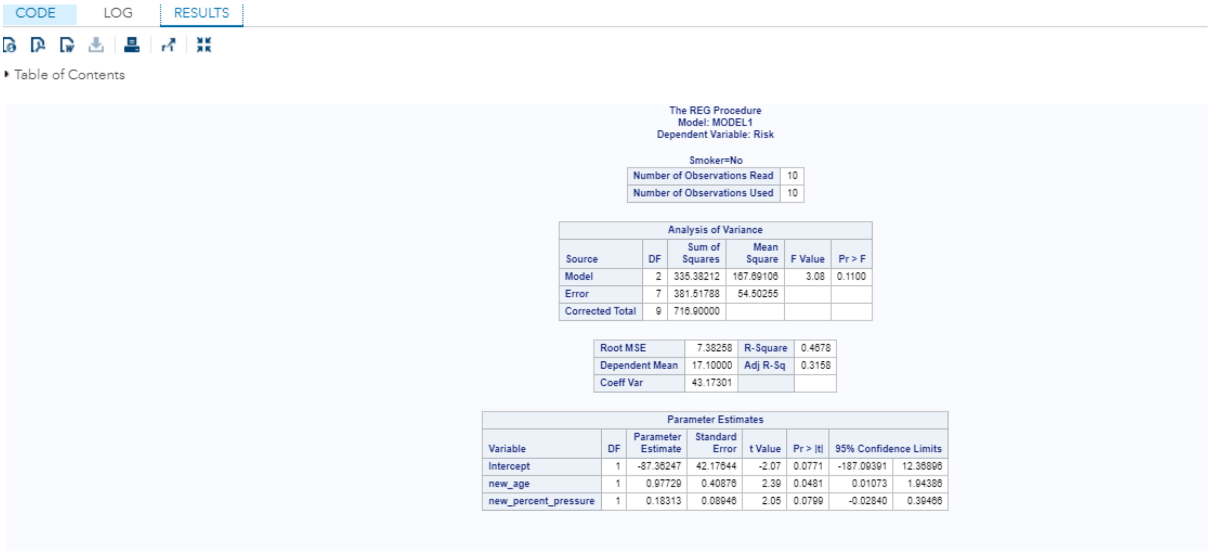
Code Window



```
proc reg data = smokers_ancova;
where age <= 79;
by Smoker;
model risk = new_age new_percent_pressure/clb;
run;
```

Result Window

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



The REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=No

Number of Observations Read	10
Number of Observations Used	10

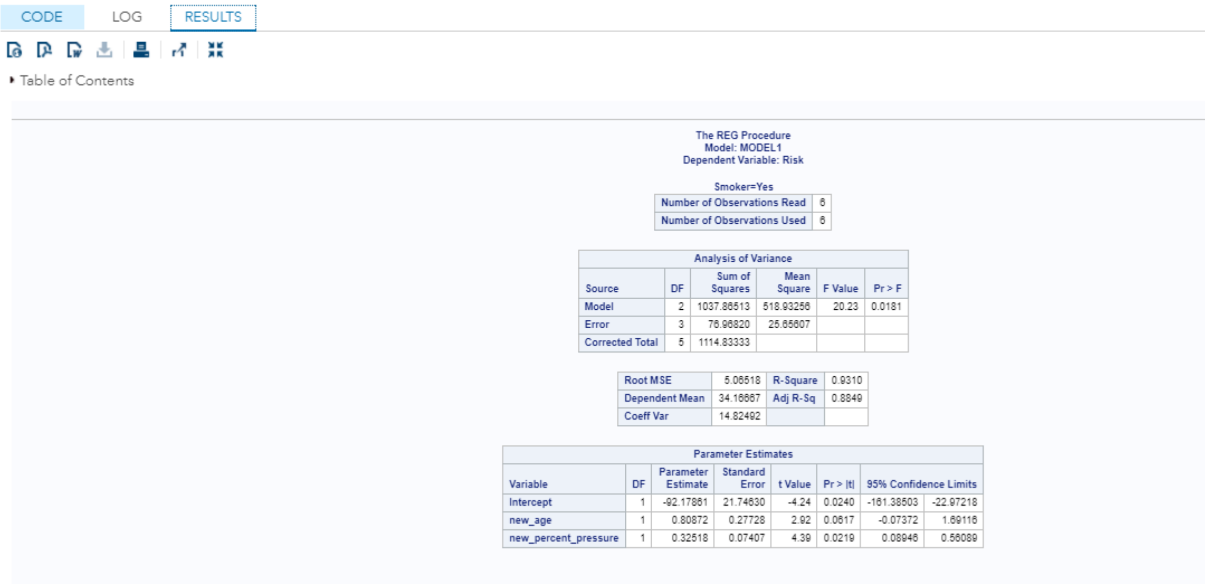
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	335.38212	167.69106	3.08	0.1100
Error	7	381.51788	54.50255		
Corrected Total	9	716.90000			

Root MSE	7.38258	R-Square	0.4678
Dependent Mean	17.10000	Adj R-Sq	0.3158
Coeff Var	43.17301		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-87.36247	42.17644	-2.07	0.0771	-187.09391	12.36896
new_age	1	0.97729	0.40876	2.39	0.0481	0.01073	1.94386
new_percent_pressure	1	0.18313	0.08946	2.05	0.0799	-0.02840	0.39466

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 REG Procedure
Model: MODEL1
Dependent Variable: Risk

Smoker=Yes

Number of Observations Read	6
Number of Observations Used	6

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1037.86513	518.93256	20.23	0.0181
Error	3	76.96820	25.65607		
Corrected Total	5	1114.83333			

Root MSE	5.06518	R-Square	0.9310
Dependent Mean	34.16667	Adj R-Sq	0.8849
Coeff Var	14.82492		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-92.17861	21.74630	-4.24	0.0240	-161.38503	-22.97218
new_age	1	0.80872	0.27728	2.92	0.0617	-0.07372	1.69116
new_percent_pressure	1	0.32518	0.07407	4.39	0.0219	0.08946	0.56089

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 R^2 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.

$$\text{Logit}(p) = \text{Log}(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots.. + \beta_m x_m$$

Where β_0 = slope & $\beta_1, \ldots., \beta_m$ are coefficients of response variables $x_1, \ldots.\ x_m$.

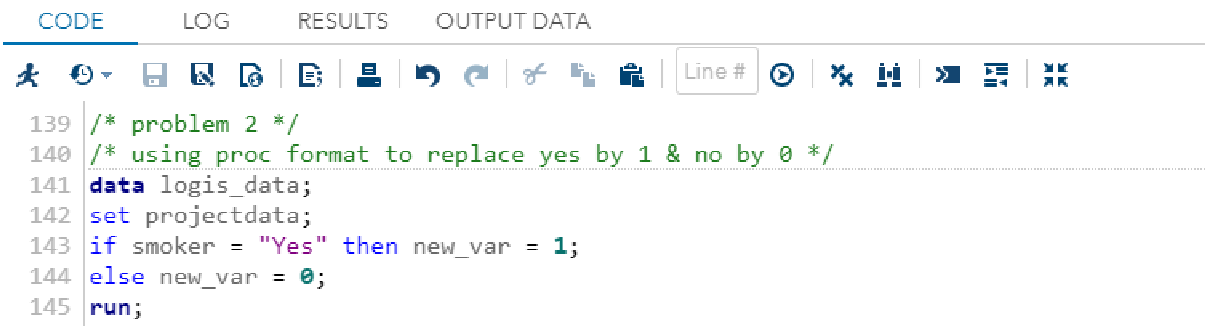
p = probability of success (probability that the person is a smoker).

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



```
/* problem 2 */
/* using proc format to replace yes by 1 & no by 0 */
data logis_data;
set projectdata;
if smoker = "Yes" then new_var = 1;
else new_var = 0;
run;
```

Output Window

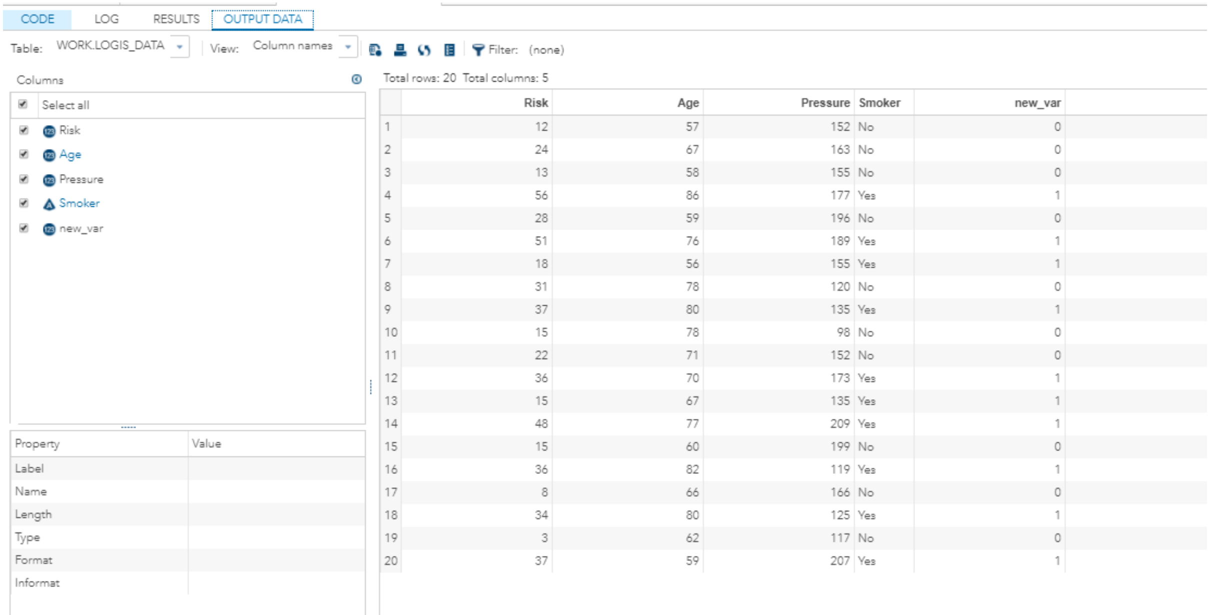


Table: WORK.LOGIS_DATA

Columns

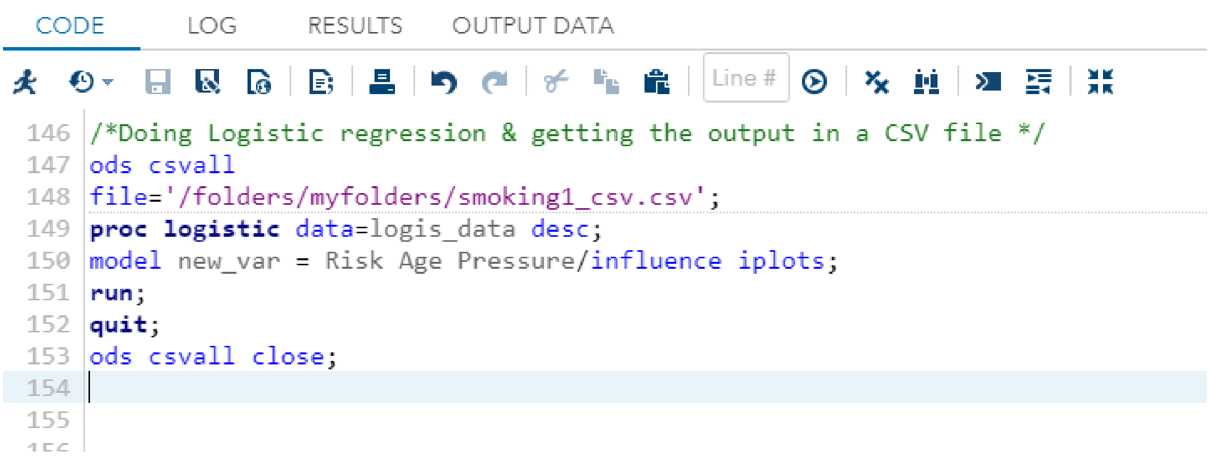
- ☑ Select all
- ☑ Risk
- ☑ Age
- ☑ Pressure
- ☑ Smoker
- ☑ new_var

Property	Value
Label	
Name	
Length	
Type	
Format	
Informat	

Total rows: 20 Total columns: 5

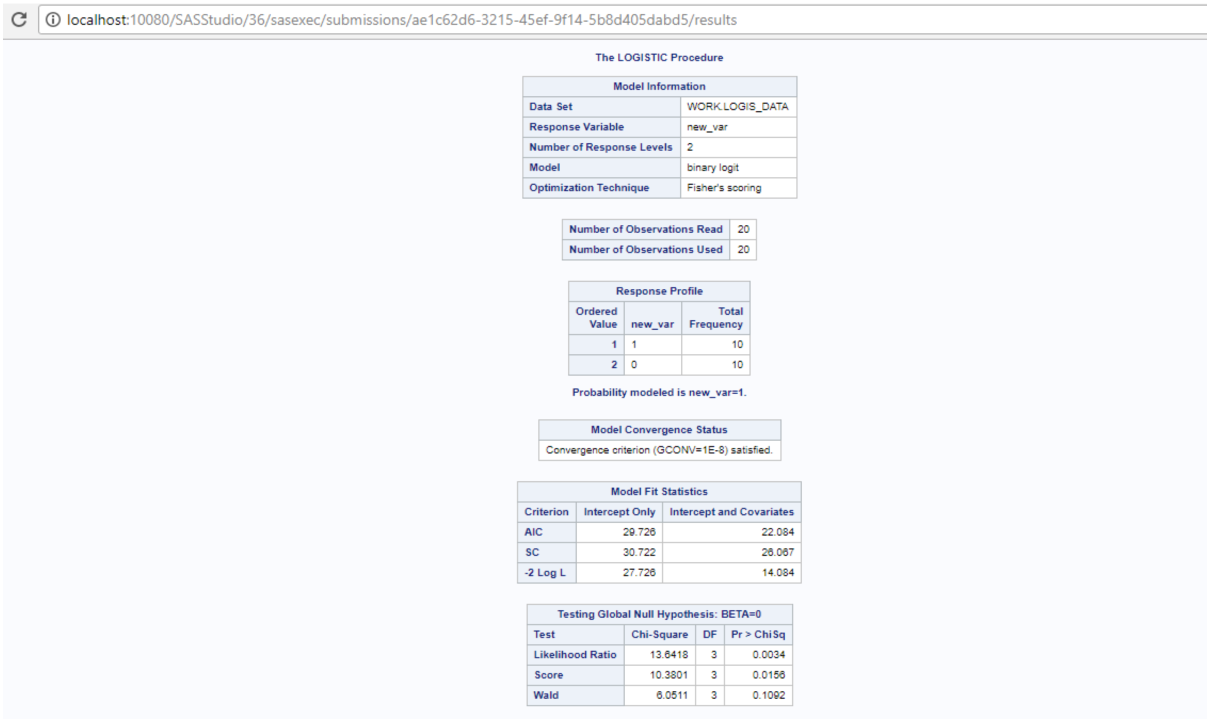
	Risk	Age	Pressure	Smoker	new_var
1	12	57	152	No	0
2	24	67	163	No	0
3	13	58	155	No	0
4	56	86	177	Yes	1
5	28	59	196	No	0
6	51	76	189	Yes	1
7	18	56	155	Yes	1
8	31	78	120	No	0
9	37	80	135	Yes	1
10	15	78	98	No	0
11	22	71	152	No	0
12	36	70	173	Yes	1
13	15	67	135	Yes	1
14	48	77	209	Yes	1
15	15	60	199	No	0
16	36	82	119	Yes	1
17	8	66	166	No	0
18	34	80	125	Yes	1
19	3	62	117	No	0
20	37	59	207	Yes	1

Code Window



```
/*Doing Logistic regression & getting the output in a CSV file */
ods csvall
file='/folders/myfolders/smoking1_csv.csv';
proc logistic data=logis_data desc;
model new_var = Risk Age Pressure/influence iplots;
run;
quit;
ods csvall close;
```

Result Window



localhost:10080/SASStudio/36/sasexec/submissions/ae1c62d6-3215-45ef-9f14-5b8d405dabd5/results

The LOGISTIC Procedure

Model Information	
Data Set	WORK.LOGIS_DATA
Response Variable	new_var
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	20
Number of Observations Used	20

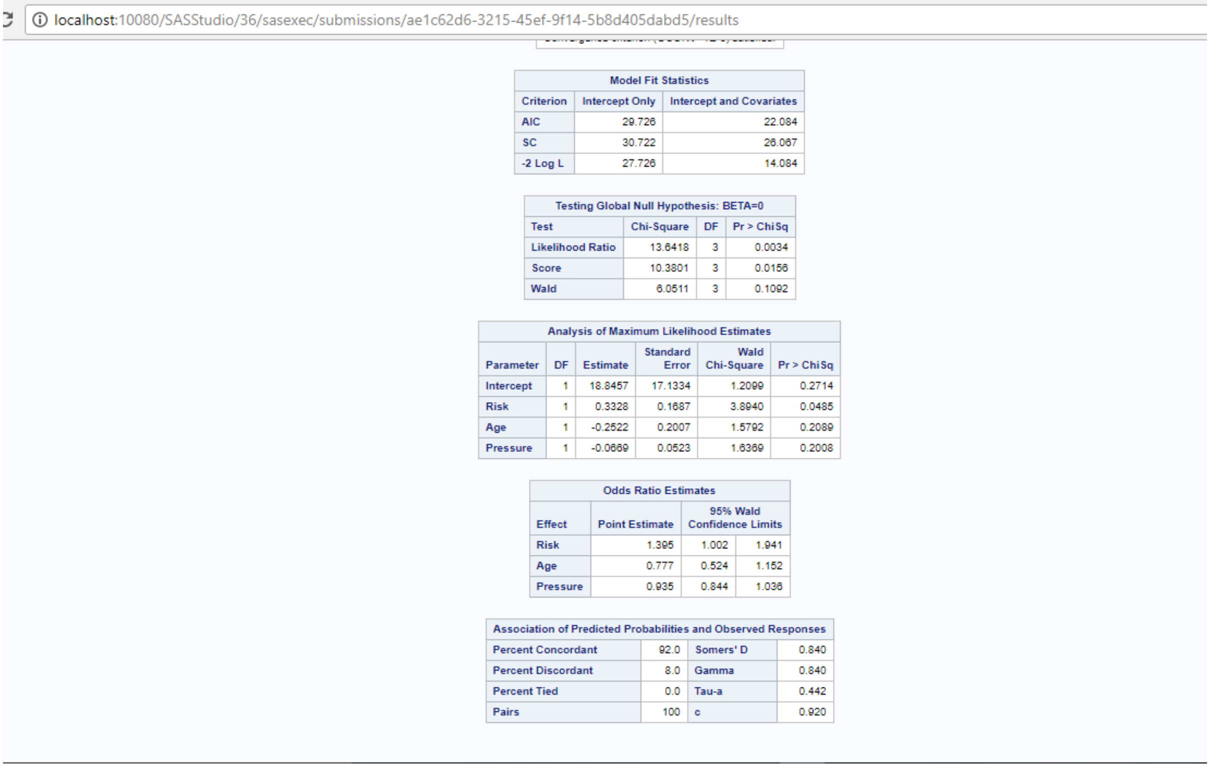
Response Profile		
Ordered Value	new_var	Total Frequency
1	1	10
2	0	10

Probability modeled is new_var=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	29.726	22.084
SC	30.722	26.067
-2 Log L	27.726	14.084

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	13.6418	3	0.0034
Score	10.3801	3	0.0156
Wald	6.0511	3	0.1092



localhost:10080/SASStudio/36/sasexec/submissions/ae1c62d6-3215-45ef-9f14-5b8d405dabd5/results

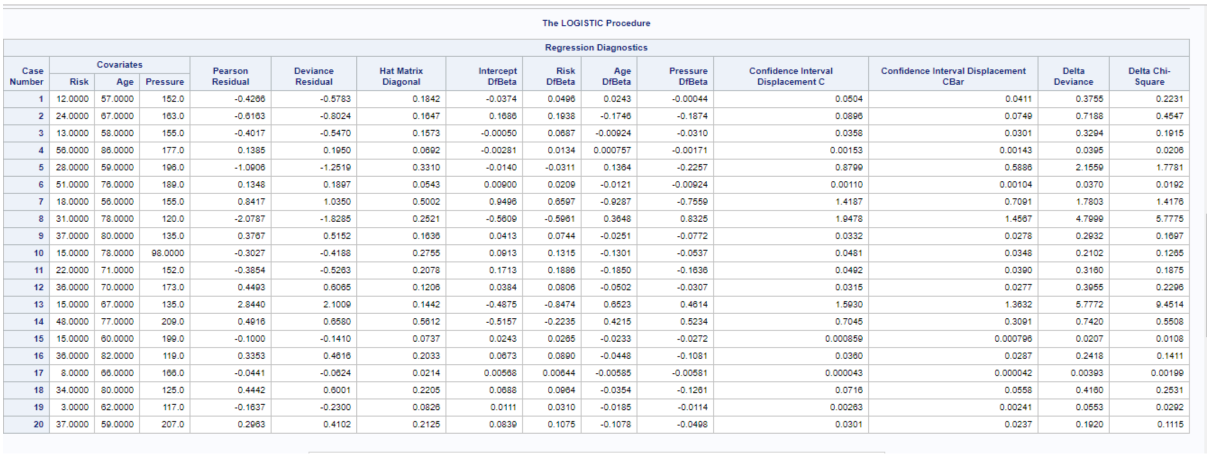
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	29.726	22.084
SC	30.722	26.067
-2 Log L	27.726	14.084

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	13.6418	3	0.0034
Score	10.3801	3	0.0156
Wald	6.0511	3	0.1092

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	18.8457	17.1334	1.2099	0.2714
Risk	1	0.3328	0.1687	3.8940	0.0485
Age	1	-0.2522	0.2007	1.5792	0.2089
Pressure	1	-0.0669	0.0523	1.6369	0.2008

Odds Ratio Estimates			
		95% Wald Confidence Limits	
Effect	Point Estimate		
Risk	1.395	1.002	1.941
Age	0.777	0.524	1.152
Pressure	0.935	0.844	1.036

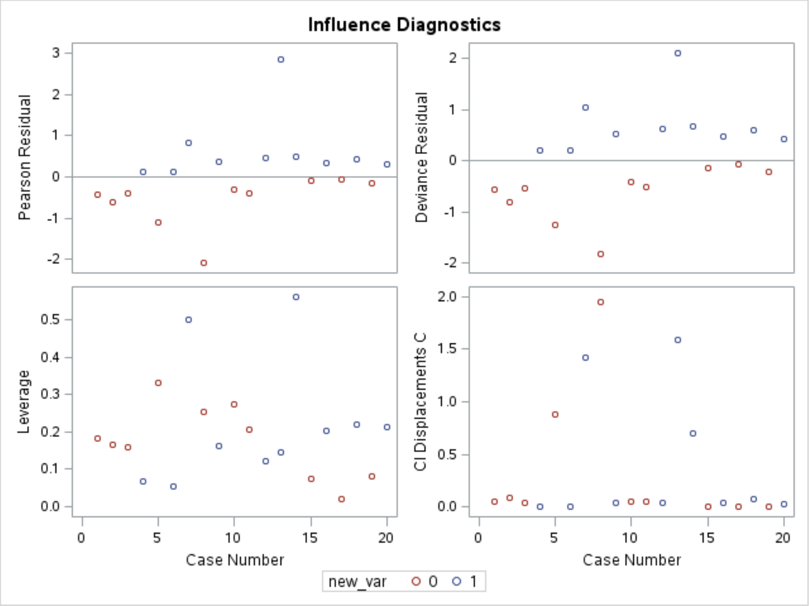
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	92.0	Somers' D	0.840
Percent Discordant	8.0	Gamma	0.840
Percent Tied	0.0	Tau-a	0.442
Pairs	100	c	0.920

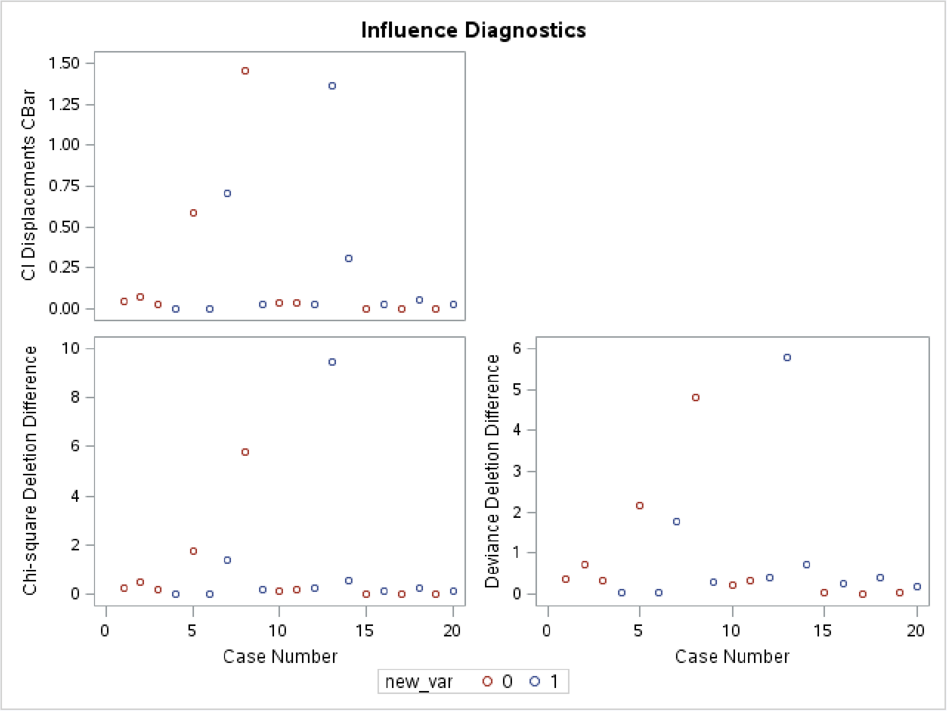


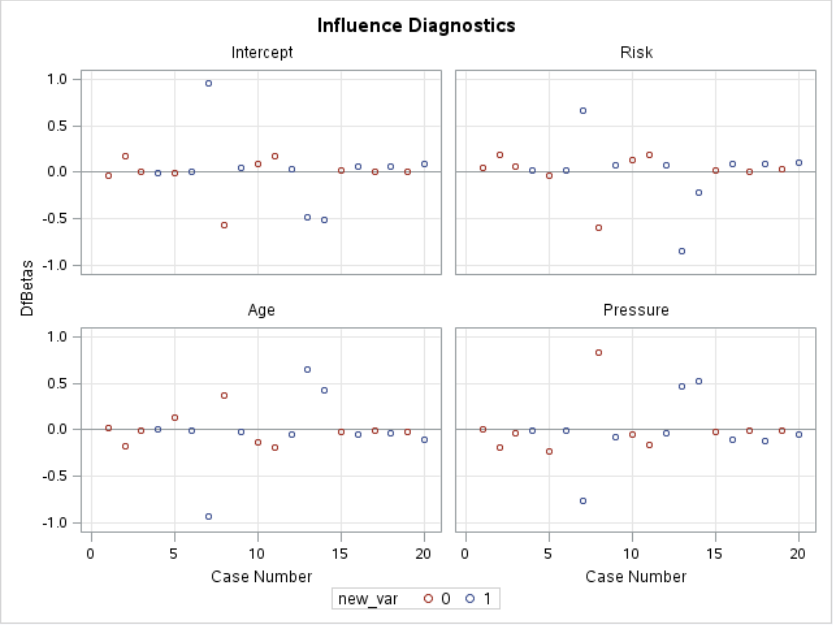
The LOGISTIC Procedure

Regression Diagnostics

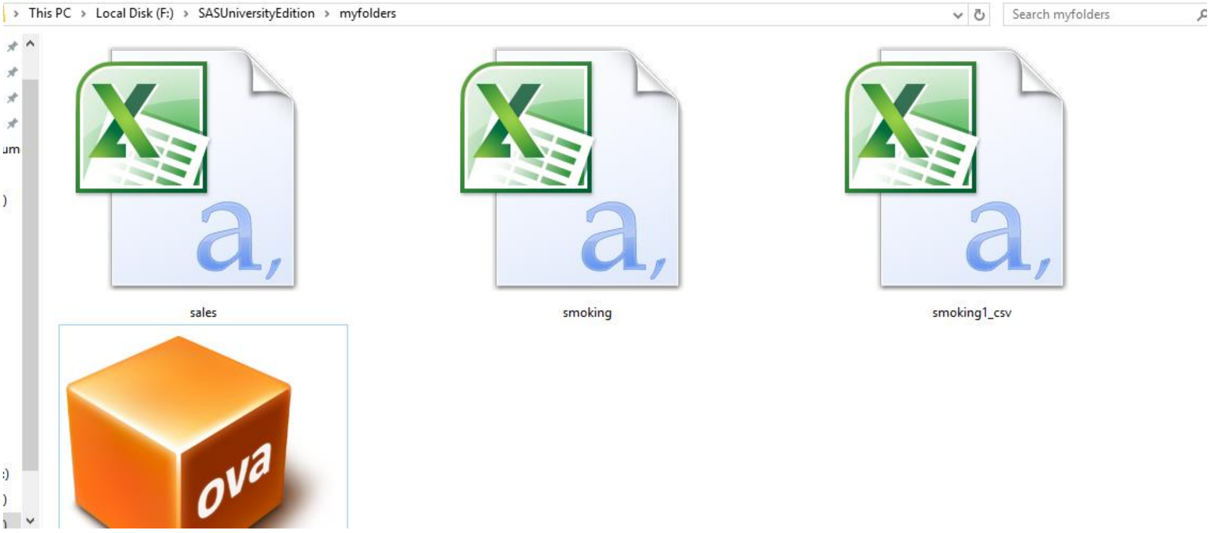
Case Number	Covariates			Pearson Residual	Deviance Residual	Hat Matrix Diagonal	Intercept DfBeta	Risk DfBeta	Age DfBeta	Pressure DfBeta	Confidence Interval Displacement C	Confidence Interval Displacement CBar	Delta Deviance	Delta Chi-Square
	Risk	Age	Pressure											
1	12.0000	57.0000	152.0	-0.4266	-0.5783	0.1842	-0.0374	0.0496	0.0243	-0.00044	0.0504	0.0411	0.3755	0.2231
2	24.0000	67.0000	163.0	-0.6163	-0.8024	0.1647	0.1686	0.1938	-0.1746	-0.1874	0.0896	0.0749	0.7188	0.4547
3	13.0000	58.0000	155.0	-0.4017	-0.5470	0.1573	-0.00050	0.0687	-0.00924	-0.0310	0.0358	0.0301	0.3294	0.1915
4	56.0000	86.0000	177.0	0.1385	0.1950	0.0692	-0.00281	0.0134	0.000757	-0.00171	0.00153	0.00143	0.0395	0.0206
5	28.0000	59.0000	196.0	-1.0906	-1.2519	0.3310	-0.0140	-0.0311	0.1364	-0.2257	0.8799	0.5886	2.1559	1.7781
6	51.0000	76.0000	189.0	0.1348	0.1897	0.0543	0.00900	0.0209	-0.0121	-0.00924	0.00110	0.00104	0.0370	0.0192
7	18.0000	56.0000	155.0	0.8417	1.0350	0.5002	0.9496	0.6597	-0.9287	-0.7559	1.4187	0.7091	1.7803	1.4176
8	31.0000	78.0000	120.0	-2.0787	-1.8285	0.2521	-0.5609	-0.5961	0.3648	0.8325	1.9478	1.4567	4.7999	5.7775
9	37.0000	80.0000	135.0	0.3767	0.5152	0.1636	0.0413	0.0744	-0.0251	-0.0772	0.0332	0.0278	0.2932	0.1697
10	15.0000	78.0000	98.0000	-0.3027	-0.4188	0.2755	0.0913	0.1315	-0.1301	-0.0537	0.0481	0.0348	0.2102	0.1265
11	22.0000	71.0000	152.0	-0.3854	-0.5263	0.2078	0.1713	0.1886	-0.1850	-0.1636	0.0492	0.0390	0.3160	0.1875
12	36.0000	70.0000	173.0	0.4493	0.6065	0.1206	0.0384	0.0806	-0.0502	-0.0307	0.0315	0.0277	0.3955	0.2296
13	15.0000	67.0000	135.0	2.8440	2.1009	0.1442	-0.4875	-0.8474	0.6523	0.4614	1.5930	1.3632	5.7772	9.4514
14	48.0000	77.0000	209.0	0.4916	0.6580	0.5612	-0.5157	-0.2235	0.4215	0.5234	0.7045	0.3091	0.7420	0.5508
15	15.0000	60.0000	199.0	-0.1000	-0.1410	0.0737	0.0243	0.0265	-0.0233	-0.0272	0.000859	0.000796	0.0207	0.0108
16	36.0000	82.0000	119.0	0.3353	0.4616	0.2033	0.0673	0.0890	-0.0448	-0.1081	0.0360	0.0287	0.2418	0.1411
17	8.0000	66.0000	166.0	-0.0441	-0.0624	0.0214	0.00568	0.00644	-0.00585	-0.00581	0.000043	0.000042	0.00393	0.00199
18	34.0000	80.0000	125.0	0.4442	0.6001	0.2205	0.0688	0.0964	-0.0354	-0.1261	0.0716	0.0558	0.4160	0.2531
19	3.0000	62.0000	117.0	-0.1637	-0.2300	0.0826	0.0111	0.0310	-0.0185	-0.0114	0.00263	0.00241	0.0553	0.0292
20	37.0000	59.0000	207.0	0.2963	0.4102	0.2125	0.0839	0.1075	-0.1078	-0.0498	0.0301	0.0237	0.1920	0.1115



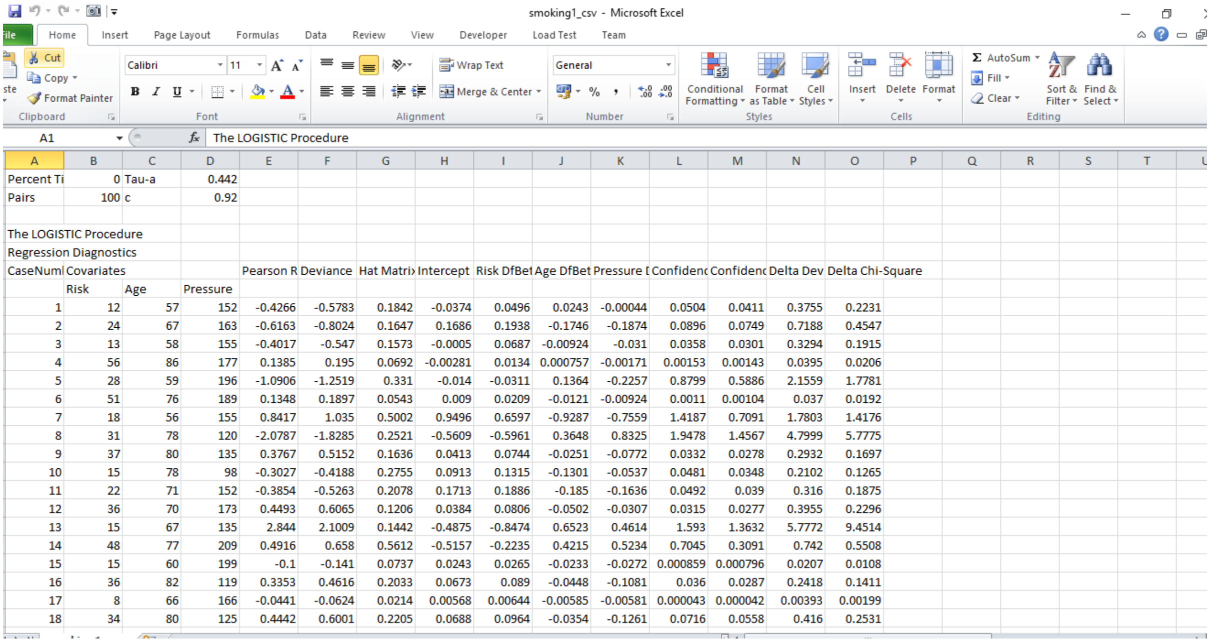




The output CSV file smoking1_csv.csv is saved in the local drive.


This PC › Local Disk (F:) › SASUniversityEdition › myfolders
Search myfolders
sales
smoking
smoking1_csv
ova

The CSV output for Risk =12 is shown below



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Percent Ti	0	Tau-a	0.442											
Pairs	100	c	0.92											
The LOGISTIC Procedure														
Regression Diagnostics														
CaseNuml	Covariates			Pearson R	Deviance	Hat Matrix	Intercept	Risk DfBet	Age DfBet	Pressure [	Confidenc	Confidenc	Delta Dev	Delta Chi-Square
	Risk	Age	Pressure											
1	12	57	152	-0.4266	-0.5783	0.1842	-0.0374	0.0496	0.0243	-0.00044	0.0504	0.0411	0.3755	0.2231
2	24	67	163	-0.6163	-0.8024	0.1647	0.1686	0.1938	-0.1746	-0.1874	0.0896	0.0749	0.7188	0.4547
3	13	58	155	-0.4017	-0.547	0.1573	-0.0005	0.0687	-0.00924	-0.031	0.0358	0.0301	0.3294	0.1915
4	56	86	177	0.1385	0.195	0.0692	-0.00281	0.0134	0.000757	-0.00171	0.00153	0.00143	0.0395	0.0206
5	28	59	196	-1.0906	-1.2519	0.331	-0.014	-0.0311	0.1364	-0.2257	0.8799	0.5886	2.1559	1.7781
6	51	76	189	0.1348	0.1897	0.0543	0.009	0.0209	-0.0121	-0.00924	0.0011	0.00104	0.037	0.0192
7	18	56	155	0.8417	1.035	0.5002	0.9496	0.6597	-0.9287	-0.7559	1.4187	0.7091	1.7803	1.4176
8	31	78	120	-2.0787	-1.8285	0.2521	-0.5609	-0.5961	0.3648	0.8325	1.9478	1.4567	4.7999	5.7775
9	37	80	135	0.3767	0.5152	0.1636	0.0413	0.0744	-0.0251	-0.0772	0.0332	0.0278	0.2932	0.1697
10	15	78	98	-0.3027	-0.4188	0.2755	0.0913	0.1315	-0.1301	-0.0537	0.0481	0.0348	0.2102	0.1265
11	22	71	152	-0.3854	-0.5263	0.2078	0.1713	0.1886	-0.185	-0.1636	0.0492	0.039	0.316	0.1875
12	36	70	173	0.4493	0.6065	0.1206	0.0384	0.0806	-0.0502	-0.0307	0.0315	0.0277	0.3955	0.2296
13	15	67	135	2.844	2.1009	0.1442	-0.4875	-0.8474	0.6523	0.4614	1.593	1.3632	5.7772	9.4514
14	48	77	209	0.4916	0.658	0.5612	-0.5157	-0.2235	0.4215	0.5234	0.7045	0.3091	0.742	0.5508
15	15	60	199	-0.1	-0.141	0.0737	0.0243	0.0265	-0.0233	-0.0272	0.000859	0.000796	0.0207	0.0108
16	36	82	119	0.3353	0.4616	0.2033	0.0673	0.089	-0.0448	-0.1081	0.036	0.0287	0.2418	0.1411
17	8	66	166	-0.0441	-0.0624	0.0214	0.00568	0.00644	-0.00585	-0.00581	0.000043	0.000042	0.00393	0.00199
18	34	80	125	0.4442	0.6001	0.2205	0.0688	0.0964	-0.0354	-0.1261	0.0716	0.0558	0.416	0.2531

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.