Christina Gustafsson

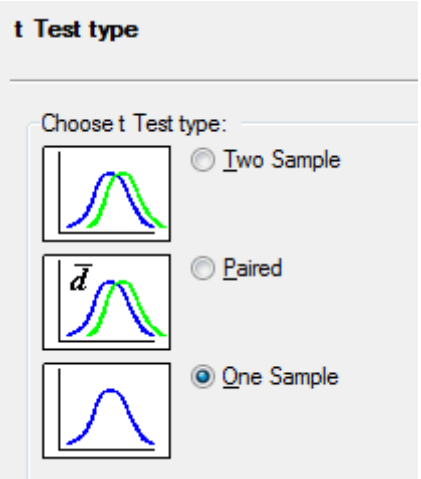
Introductory Guide to SAS Enterprise Guide 6.1 Part IV

10. TESTS OF MEANS AND SOME NONPARAMETRIC TESTS

10.1. Some Tests for One Sample

Let's study a quantitative variable whose distribution is normal distribution (or the distribution is a symmetric one and the number of cases in the sample is at least 40). Now, we are interested in making some conclusions of the population mean that are based on the sample mean. We are going to perform a test called **one sample t test (of mea**n**)**.

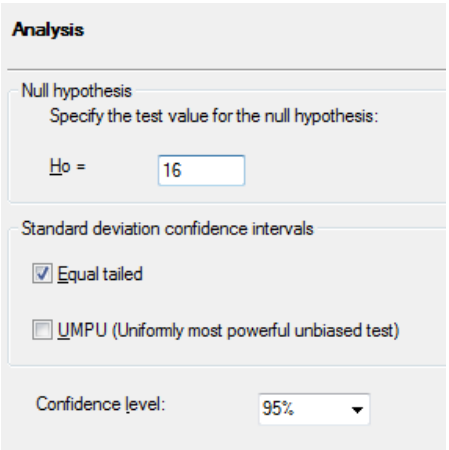
Let's imagine that you assume that the population mean of *attend2* is 16 hours for male students. So, you have a sample of male students and the *attend2*-variable is a quantitative one, and you have checked the normality. You can obtain the one sample t test by selecting **Tasks > ANOVA > t Test**. In the **t Test type** tab you can select the test you want to perform (e.g. **One Sample**).



In the **Data** tab you must assign the quantitative variable to the role **Analysis variables** (e.g. *attend2*). In the **Analysis** tab you can specify the alleged value of the population mean (e.g. *16*) in the **H_0** box. The test hypotheses are now in general form

H_0: The population mean $\mu = \mu_0$.
H_1: The population mean $\mu \neq \mu_0$.

In this case you then replace μ_0 with 16.



With the **Plots** tab you can select the different plots for instance to check the normality of the data if you have not done it yet.

The results in the first table are just basic statistics of *attend2*, for instance the sample mean is 15.2 hours. From the second result table you can see that the 95% confidence interval for population mean is (13.78, 16.62). The third table then shows the test's results. The p-value of the test is 0.2628, and because it is greater than 0.05, we can now accept the null hypothesis.

N	Mean	Std Dev	Std Err	Minimum	Maximum
50	15.2000	4.9939	0.7062	2.0000	28.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
15.2000	13.7808	16.6192	4.9939	4.1716	6.2230

DF	t Value	Pr > |t|
49	-1.13	0.2628

Let's now study a quantitative variable whose distribution is non-normal. Now, we are interested in making some inferences about the population location, but instead of the mean we now study the sample median with the **Wilcoxon signed rank test**.

Let's imagine that you assume that the population median of *attend2* is 16 hours for all students. The variable is a quantitative one, but the distribution is non-normal. Instead of the parametrical t test you must use a nonparametric test. You can perform a suitable test by selecting **Tasks** > **Describe** > **Distribution analysis**. In the **Data** tab must assign the quantitative variable to the role **Analysis variables** (e.g. *attend2*). In the **Tables** tab click the **Tests of location** option and specify the alleged value of population median (e.g. *16*) in the $\mathbf{H_0}$ box. The test hypotheses are in general form

H_0: The population median = Md_0
H_1: The population median ≠ Md_0.

In this case you then replace Md_0 with 16.

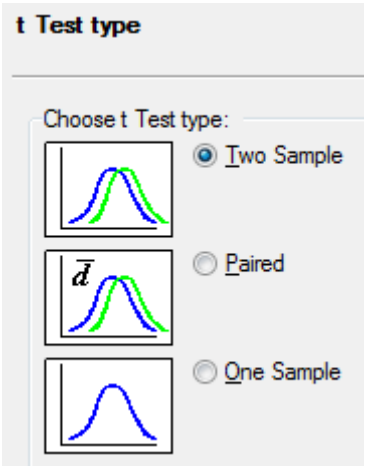
Now, we look through the test results of the (Wilcoxon) **Signed Rank** test in the **Test for Location** table. The p-value of that test is 0.0055. Because the p-value is lower than 0.01, we can now reject H_0 at 1% significance level.

Tests for Location: Mu0=16				
Test	Statistic		p Value	
Student's t	t	3.12878	Pr > |t|	0.0019
Sign	M	13.5	Pr >= |M|	0.1181
Signed Rank	S	3674	Pr >= |S|	0.0055

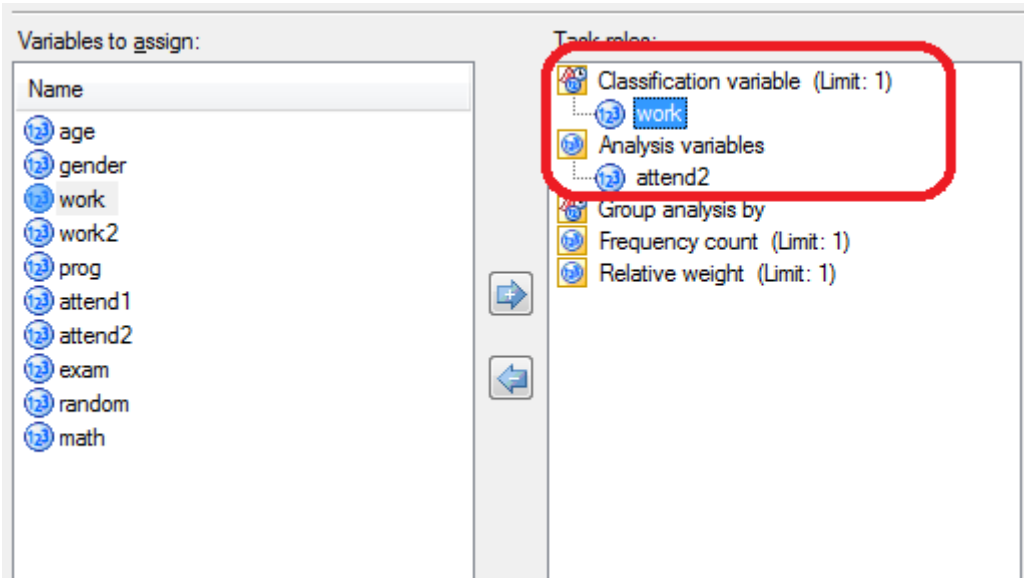
10.2. Some Tests for Two Independent Samples

Let's study a quantitative variable whose distribution is normal distribution (or the distribution a symmetric one and the number of cases in both samples is at least 40) for two different populations. Now, we are interested in making some inferences whether or not the mean of the quantitative variable is the same in the two populations. The test we are going to perform is called **two (independent) samples t test (of mean**).

Let's imagine that you assume that the population mean of *attend2* is the same for students who work during the semester (= population 1) and for students who don't work during the semester (= population 2). Now, you have two independent samples of students. And let's assume that you have checked the normality of *attend2* in both samples. You can obtain the suitable t test by selecting **Tasks > ANOVA > t Test**. In the **t Test type** tab you select then the **Two Sample** option.



In the **Data** tab you must assign the variable that classifies your sample into two groups to the role **Classification variable** (e.g. *work*) and the quantitative variable to the role **Analysis variables** (e.g. *attend2*).



In the **Analysis** tab you can specify for the test of means the alleged difference between the population means (e.g. *0*) in the **H_0** box. The test hypotheses are now

H_0: $\mu_1 = \mu_2$ (the means of the two populations are equal)
H_1: $\mu_1 \neq \mu_2$.

With the **Plots** tab you can select the different plots to for instance check the normality if you haven't done it yet.

The results in the first table are just basic statistics of *attend2* for the two groups of *work*, for instance the sample means are 16.2 and 17.1 hours, so it seems that those students who work during semester tend to attend less to lectures than those who do not work. From the second result table you can see that the 95% confidence intervals for the means of different populations.

work	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	73	16.2329	5.0594	0.5922	6.0000	32.0000
2	243	17.1276	5.2755	0.3384	2.0000	33.0000
Diff (1-2)		-0.8947	5.2267	0.6976		

work	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		16.2329	15.0524	17.4133	5.0594	4.3510	6.0455
2		17.1276	16.4609	17.7942	5.2755	4.8444	5.7914
Diff (1-2)	Pooled	-0.8947	-2.2673	0.4779	5.2267	4.8480	5.6702
Diff (1-2)	Satterthwaite	-0.8947	-2.2448	0.4554			

The third and fourth tables then include the results of some tests. The fourth table presents results for the test of **Equality of Variances**. The hypotheses for this test are

H_0: $\sigma_1^2 = \sigma_2^2$ (the variances of the two populations are equal)
H_1: $\sigma_1^2 \neq \sigma_2^2$.

Because the p-value of the variance test is 0.6876, and it's greater than 0.05, we can now accept the H_0. And now, because the variances of the two populations are equal, you can look through the third table and read the results of the **Pooled** test of means (where the variances are assumed equal). The p-value of the test is 0.2006, and because it is greater than 0.05, we can now accept the null hypothesis of the t test. If the variances would have been equal then you should look through the result of the **Satterthwaite** test.

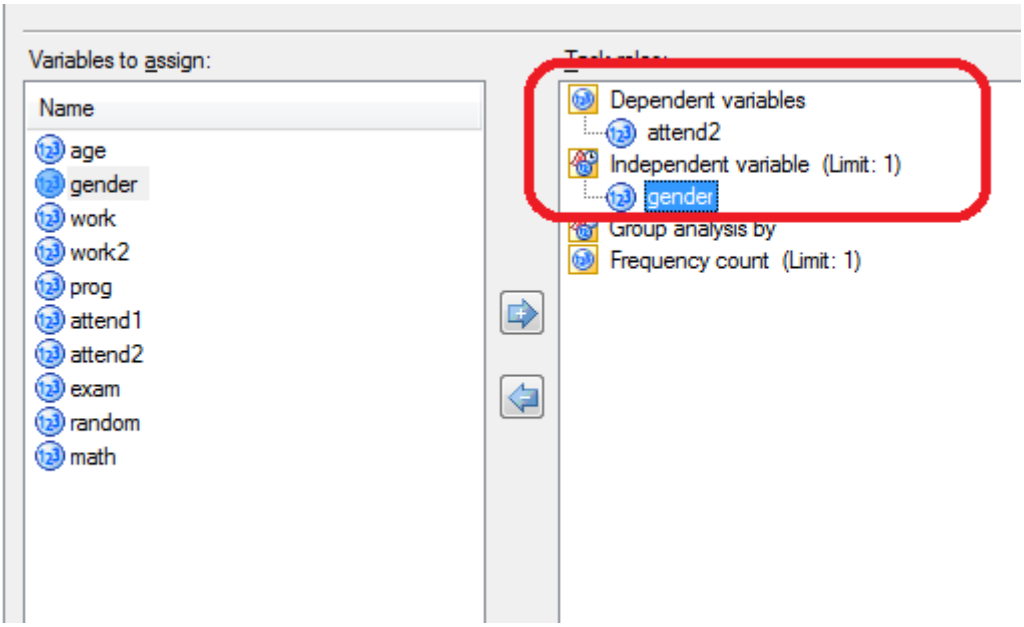
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	314	-1.28	0.2006
Satterthwaite	Unequal	122.82	-1.31	0.1920

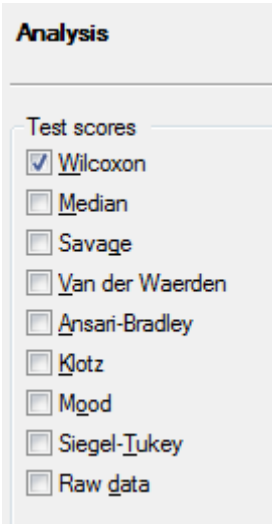
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	242	72	1.09	0.6876

Now, let's study a quantitative variable whose distribution is non-normal for two different populations. Again, we are interested in making some inferences about the locations of the populations. Now, we are going to perform the **Wilcoxon two-sample test**.

Let's imagine that you assume that the population location of *attend2* is the same for both male and female students. You have now two independent samples of students. The *attend2* is a quantitative variable, but the distribution of it isn't normal for female students. Instead of the parametrical t test you must choose a nonparametric test. You can perform a suitable test by selecting **Tasks** > **ANOVA** > **Nonparametric One-Way ANOVA**. In the **Data** tab you then assign the quantitative variable to the role **Dependent variables** (e.g. *attend2*) and the grouping variable to the role **Independent variable** (e.g. *gender*). In the **Analysis** tab check the **Wilcoxon** option. The **Wilcoxon two-sample** test is quite the same as a test called Mann-Whitney U -test. You can calculate the exact p-value of your test, if you check the **Wilcoxon** option in the **Exact p-values** tab, too. That latter option is useful only for small datasets. The test hypotheses are

H_0: The distributions of the two populations are the same.
H_1: The distributions of the two populations are not the same.





In the first results table you can see that the mean scores for *attend2* is almost 163 for female students and 127 for male students, so it seems that female students tend to attend more to lectures than male students.

Wilcoxon Scores (Rank Sums) for Variable attend2 Classified by Variable gender					
gender	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
female	263	42797.50	41291.0	583.772932	162.728137
male	50	6343.50	7850.0	583.772932	126.870000
Average scores were used for ties.					

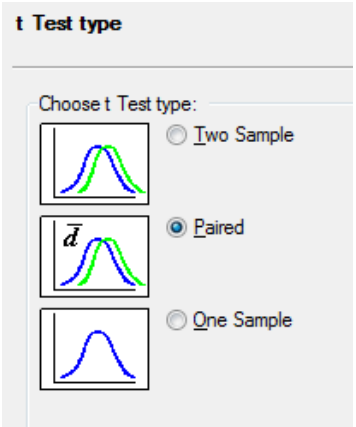
In the second results table you can see different approximation of the p-value. These results tend to be the same: now the p-value for two-sided test is approximately 0.0099 and because it is less than 0.01, we can now reject the null hypothesis at 1% significance level.

Wilcoxon Two-Sample Test	
Statistic	6343.5000
Normal Approximation	
Z	-2.5798
One-Sided Pr < Z	0.0049
Two-Sided Pr > |Z|	0.0099
t Approximation	
One-Sided Pr < Z	0.0052
Two-Sided Pr > |Z|	0.0103
Z includes a continuity correction of 0.5.	

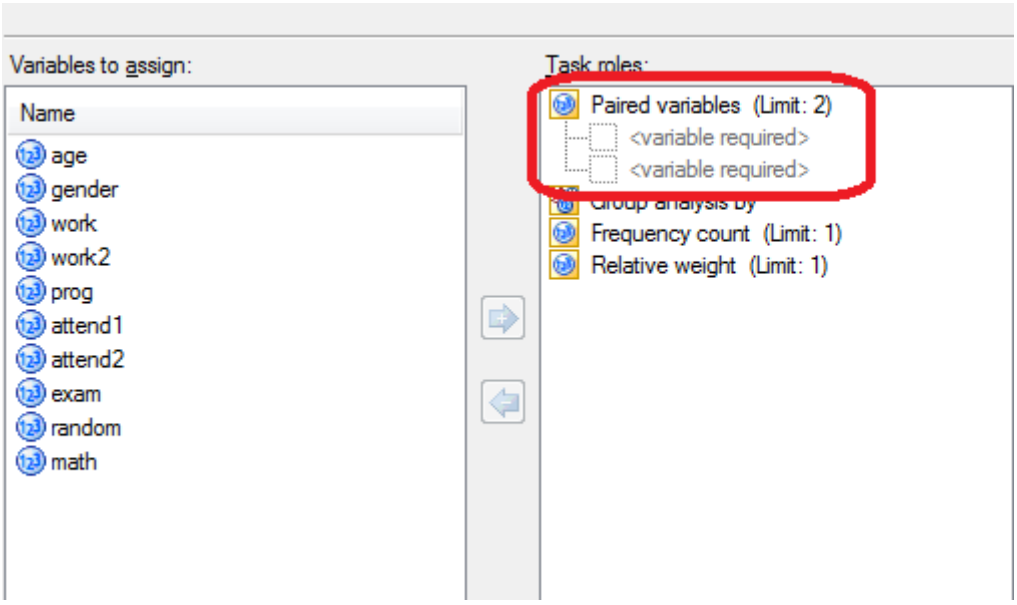
10.3. Some Tests for Paired Samples

Let's now study two quantitative normally distributed variables of the same measurement that are made under two different conditions. Now, we are interested in making some inferences whether or not the means of the quantitative variables are the same. The test that we are going to perform is based on the paired differences between the two variables. The **paired samples t test** is in fact an application of the one sample t test.

You can obtain the suitable t test by selecting **Tasks > ANOVA > t Test**. In the **t Test type** tab you select the **Paired** option.

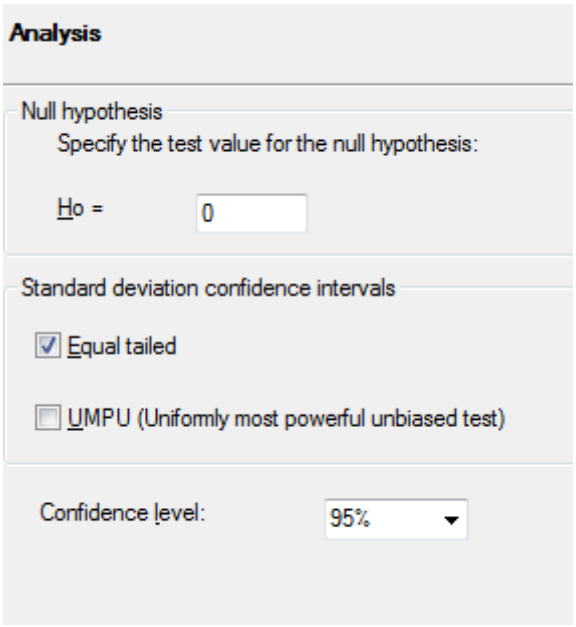


In the **Data** tab you must assign both of the variables to the role **Paired.**



In the **Analysis** tab you can specify the alleged difference of the population mean (e.g. *0*) in the **H_0** box. The test hypotheses are now in general form

H_0: $d = \mu_2 - \mu_1 = 0$. (The mean value of difference is 0; the means of the two populations are equal.)
H_1: $d \neq 0$.



The results tables of the paired samples t test look quite the same as the results of one sample t test and they can be interpreted in the same way.

Let's now study two quantitative variables whose distributions are non-normal. Now, we are interested in making some inferences whether or not the locations of the populations are the same. You can apply the **Wilcoxon Signed Rank test** for one sample to this case too, but first you have to just calculate the difference between the variable values for each case (by using **Query Builder**).

And then you perform the test for the difference the same way you would act with the case of one sample by selecting **Tasks > Describe > Distribution analysis**. In the **Data** tab you must assign the difference to the role **Analysis variables**. In the **Tables** tab click the **Tests of location** option and specify the alleged value of the median of the difference in the **H_0** box. The general form of hypotheses are

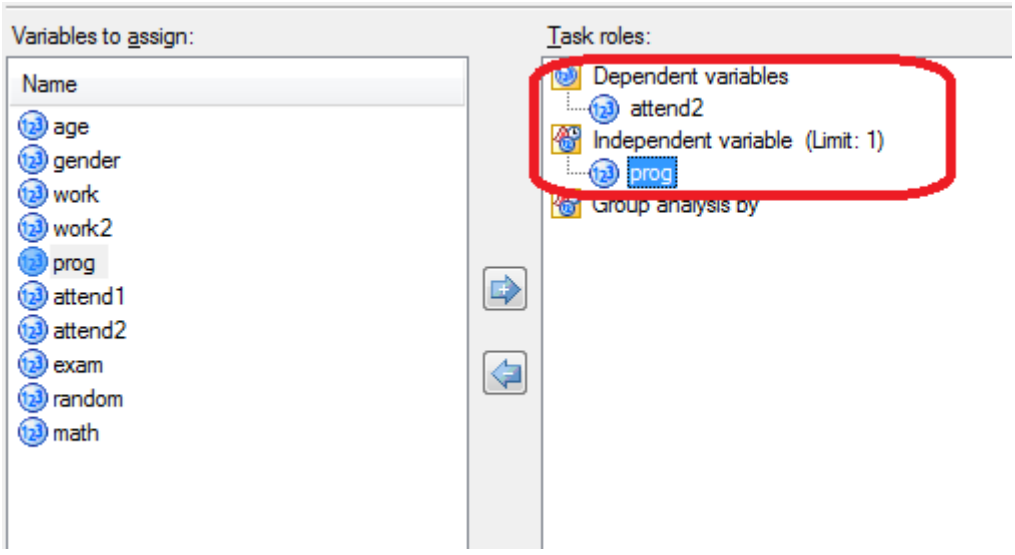
H_0: The median value of difference is 0.
H_1: The median value of difference is not 0.

Again, the results can be interpreted the same way than in the case of one sample signed rank test.

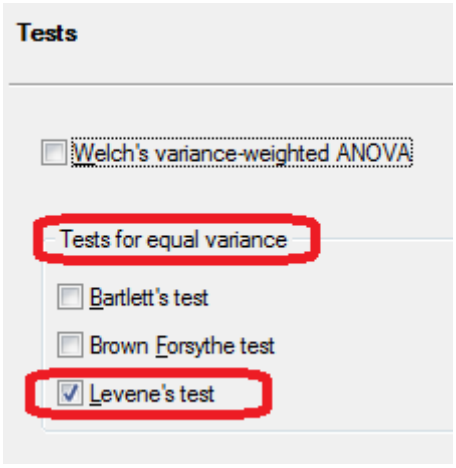
10.4. One-way Analysis of Variance (ANOVA) and Kruskal-Wallis -test

The **one-way analysis of variance** allows you to compare whether or not all the means of the same quantitative variable are equal in three or more different populations. So, in a way, it generalizes two-sample t test to more than two groups. Again, the distribution of the quantitative variable should be normal distribution in every population and even the variances of the quantitative variable should be equal in all the populations. In the case of one-way ANOVA you have one independent variable that classifies you data into three or more groups.

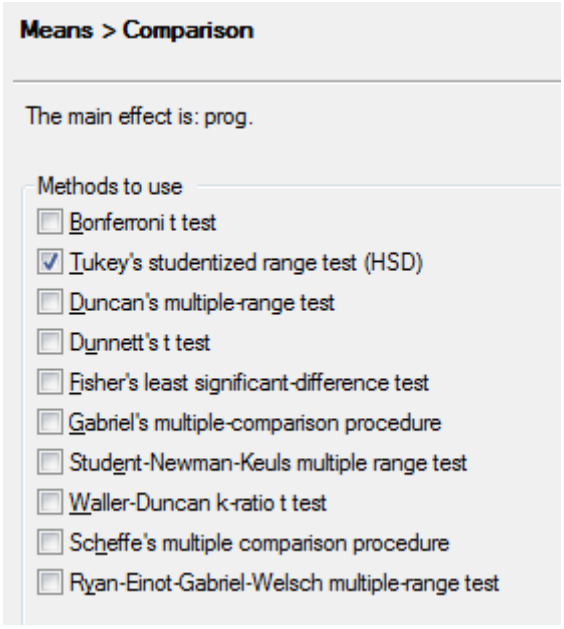
Let's imagine that you assume that the population mean of *attend2* is the same for every group of *prog*. Now, you have five independent samples of students. And let's assume that you have checked the normality of *attend2* in every sample. You can obtain the one-way ANOVA by selecting **Tasks > ANOVA > One-way ANOVA**. In the **Data** tab you must assign the variable that classifies your sample into three or more groups to the role **Independent** (e.g. *prog*) and the quantitative variable to the role **Dependent variables** (e.g. *attend2*).



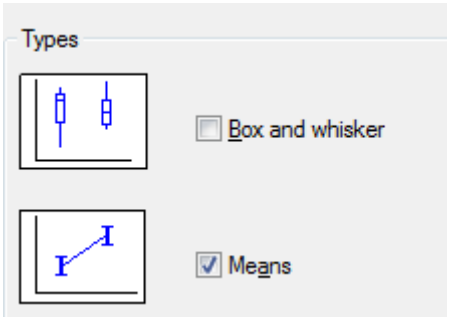
In the **Tests** tab you can check the **Levene's test** option if you want to perform an analysis to figure out whether or not the variances are equal in each group (population). If the results of this test show that the variances are equal, then you do not have to use any other option in this tab. If the results show that the variances are not equal, then you should select also the **Welch's variance-weighted ANOVA** option in order to perform that analysis instead the ordinary ANOVA analysis.



In the **Comparison** tab the options enable you to obtain results of pair wise comparisons of the means. These tests are called **Post Hoc** –tests (I quite often prefer the **Tukey's Studentized range test (HSD)**). And you perform these tests only if you have already obtained such results that the population means are not equal.



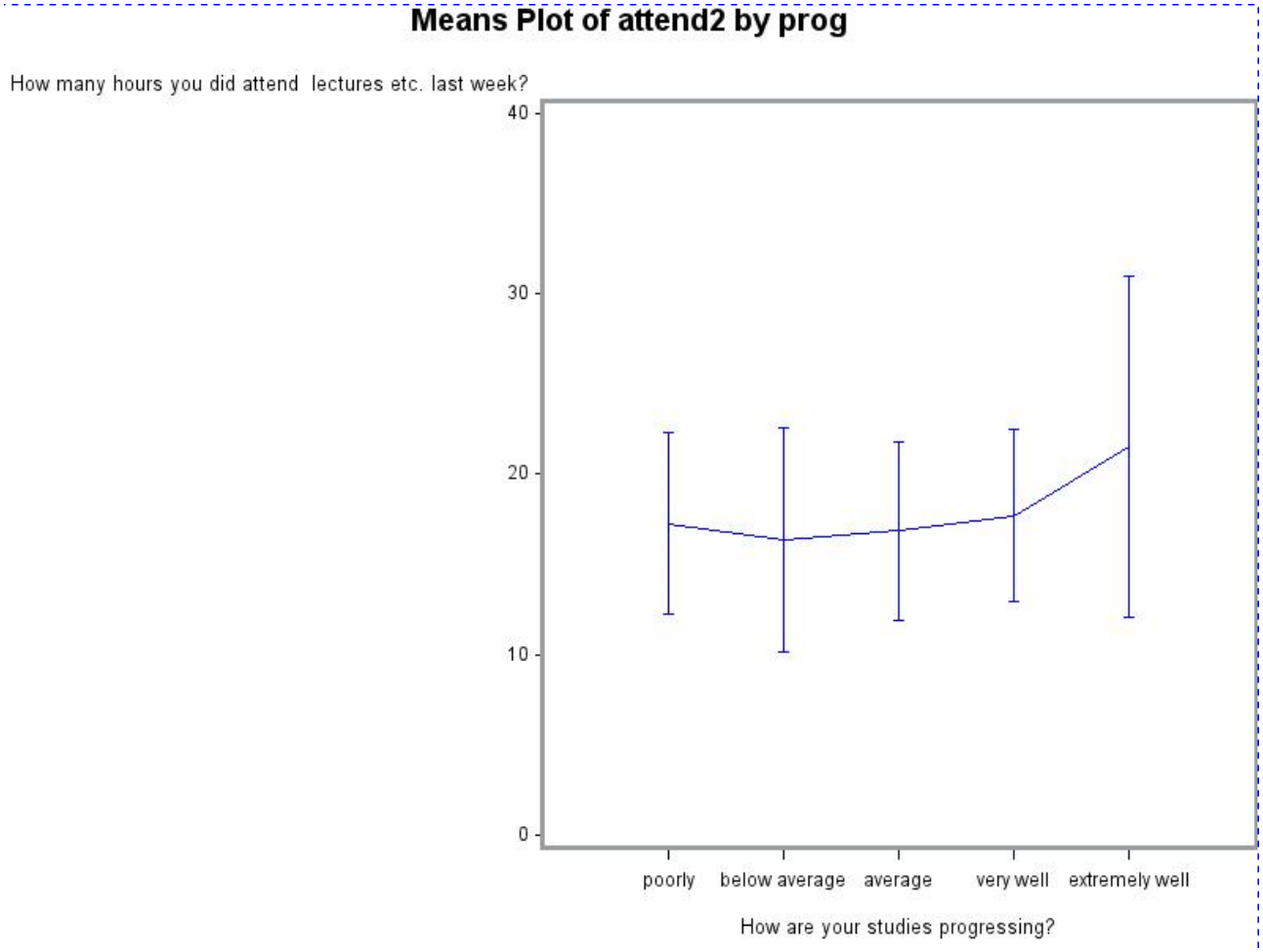
With the **Breakdown** tab you can select which statistics you want to be shown in the results tables. In the **Plots** tab you can then select which plot to include in the results.



When you look through the results, it's a good idea to start the interpreting with the descriptive statistics results table or the Means plot. The table (or the graph, too) shows that *prog* group

"extremely well" has the highest mean of *attend2* and the group "below average" the lowest mean, so the sample means seem to be different.

Level of prog	N	attend2	
		Mean	Std Dev
average	202	16.8465347	4.96968050
below average	52	16.3653846	6.19939544
extremely well	4	21.5000000	9.43398113
poorly	11	17.2727273	5.04164475
very well	43	17.6976744	4.77859754



In the next stage, you should interpret the results of the Levene's test for Homogeneity of Variance. The hypotheses for this test are

H_0: The variances for all the populations are equal.
H_1: The variances for all the populations are not equal.

The p-value of the Levene's test is now 0.1087, and because it's greater than 0.05, we can now accept the H_0.

Levene's Test for Homogeneity of attend2 Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
prog	4	14567.6	3641.9	1.91	0.1087
Error	307	585560	1907.4		

Because the variances are equal, you can now interpret the results in the ordinary ANOVA table. The hypotheses for the ordinary ANOVA F test in the next table are

H_0: The means for all the populations are equal.
H_1: The means for all the populations are not equal.

The p-value of the F test is now 0.3249, and because it is greater than 0.05, we can now accept the H_0.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	127.906481	31.976620	1.17	0.3249
Error	307	8404.551852	27.376390		
Corrected Total	311	8532.458333			

If you have to reject the H_0 of the Levene's test for variances, then instead of interpreting the basic ANOVA F test you should interpret the Welch's test. The hypotheses for this test are the same as for the ordinary ANOVA F test.

Welch's ANOVA for attend2			
Source	DF	F Value	Pr > F
prog	4.0000	0.58	0.6822
Error	17.5275		

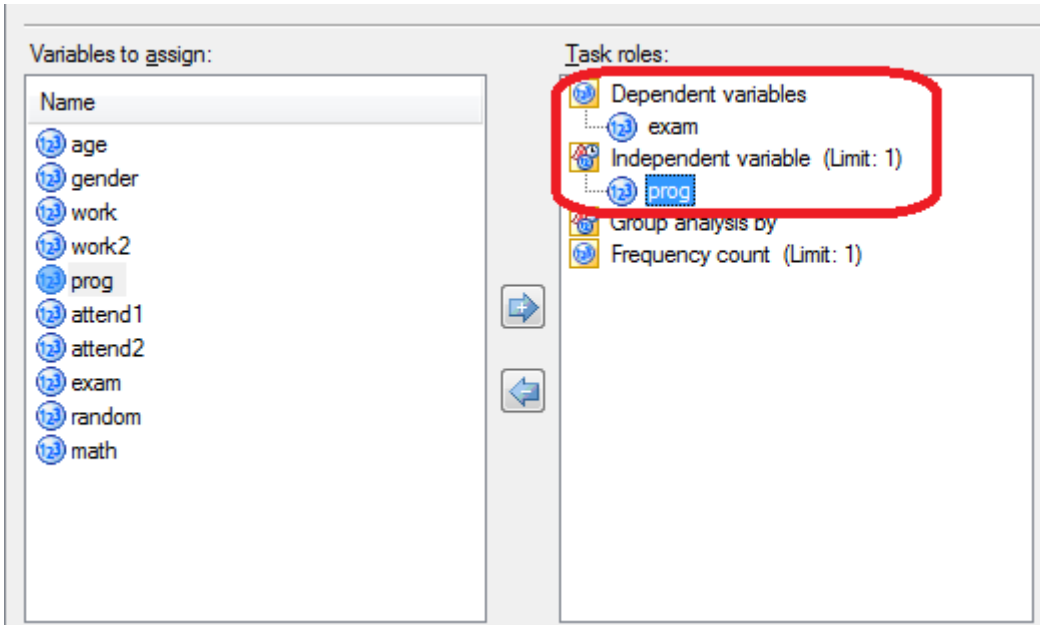
If you have to reject the H_0 of the ordinary ANOVA F test, then you could do the pair wise comparisons of the group means by looking through the results of the Tukey's test (in the next table just few of the pair wise comparisons are shown). If there is significant difference at the 0.05 significance level between two group means, it will be indicated by ***.

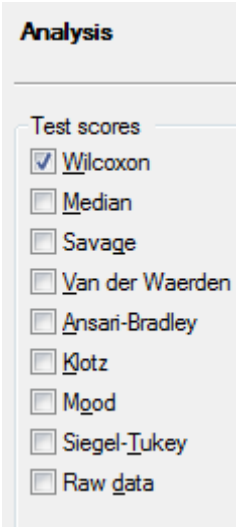
Comparisons significant at the 0.05 level are indicated by ***.			
prog Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
extremely well - very well	3.8023	-3.7029	11.3076
extremely well - poorly	4.2273	-4.1558	12.6103
extremely well - average	4.6535	-2.5961	11.9030
extremely well - below average	5.1346	-2.3152	12.5844
very well - extremely well	-3.8023	-11.3076	3.7029
very well - poorly	0.4249	-4.4262	5.2761
very well - average	0.8511	-1.5602	3.2625

Now, let's study a quantitative variable whose distribution is non-normal for three or more different populations. Again, we are interested in making some inferences about the locations of the populations. The test we are going to perform is called **Kruskal-Wallis test**.

Let's imagine that you assume that the population location of *exam* the same for every group of *prog*. Now, you have five independent samples of students. The *exam* is a quantitative variable, but the distribution of it is not normal in the samples. Instead of the parametrical ANOVA you must use a nonparametric test. You can perform a suitable test by selecting **Tasks** > **ANOVA** > **Nonparametric One-Way ANOVA**. In the **Data** tab you then assign the quantitative variable to the role **Dependent variables** (e.g. *exam*) and the grouping variable to the role **Independent variable** (e.g. *prog*). In the **Analysis** tab check the **Wilcoxon** option. This option enables you to obtain the results of the Kruskal-Wallis test. You can calculate the exact p-value of your test, if you check the **Wilcoxon** option in the **Exact p-values** tab, too. That latter option is again useful only for small datasets. The test hypotheses are

H_0: The distributions (especially the locations) of the populations are the same.
H_1: The distributions of the populations are not the same.





In the first results table you can see that the mean scores of *exam* is just 92 in *prog* group "poorly" and 219 in *prog* group "extremely well", so the distributions seem to be quite different for these extreme groups.

Wilcoxon Scores (Rank Sums) for Variable exam Classified by Variable prog					
prog	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
below average	50	7606.50	7775.00	576.852210	152.130000
average	202	31396.50	31411.00	747.274989	155.428218
very well	43	7308.00	6686.50	542.104163	169.953488
poorly	11	1017.00	1710.50	290.151501	92.454545
extremely well	4	877.00	622.00	177.004209	219.250000
Average scores were used for ties.					

In the next table you can see the results of the Kruskal-Wallis test. The p-value of that test is 0.0673, and because it is greater than 0.05, we can no accept the H_0.

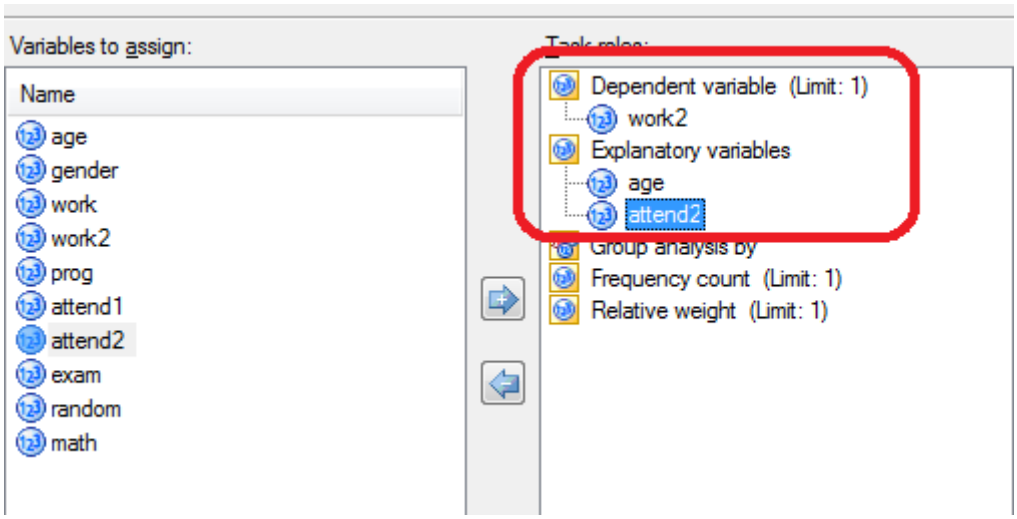
Kruskal-Wallis Test	
Chi-Square	8.7624
DF	4
Pr > Chi-Square	0.0673

You could illustrate the possible difference in locations by using a Box plot chart. With that chart you can see for instance medians, lower quartiles and upper quartiles of the groups.

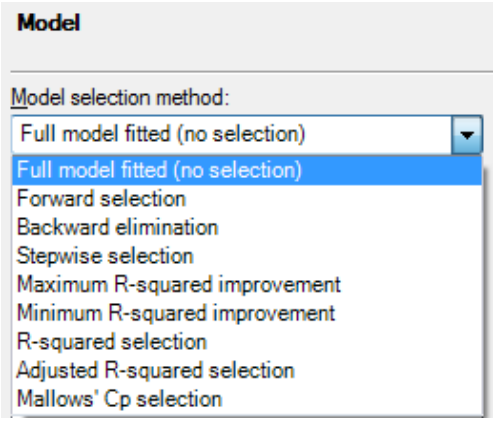
11. LINEAR REGRESSION

Regression analysis includes techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable (denoted usually y) and one or more independent variables (also called explanatory variables, usually denoted by x_i). The idea of linear regression analysis is to illustrate the linear relationship between one (or more) usually quantitative explanatory variable and a quantitative dependent (response) variable.

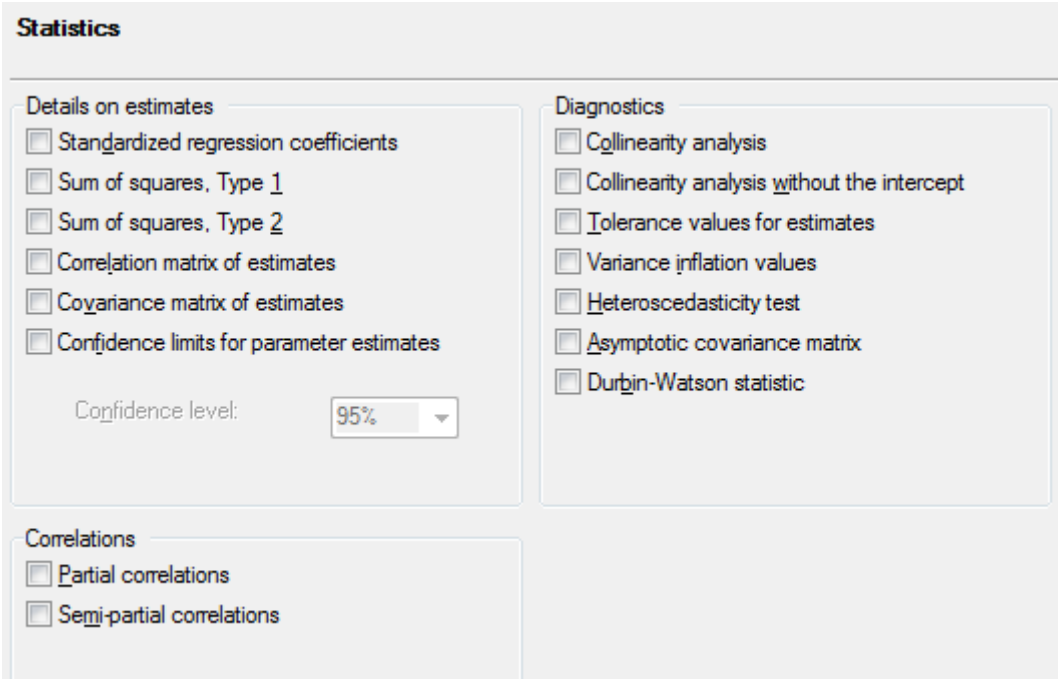
You obtain linear regression analysis by selecting **Tasks > Regression > Linear Regression** (or HP Linear Regression). In the **Data** tab you assign just one **Dependent variable** (e.g. *work2*) and at least one **Explanatory variable(s)** (e.g. *age* and *attend2*).



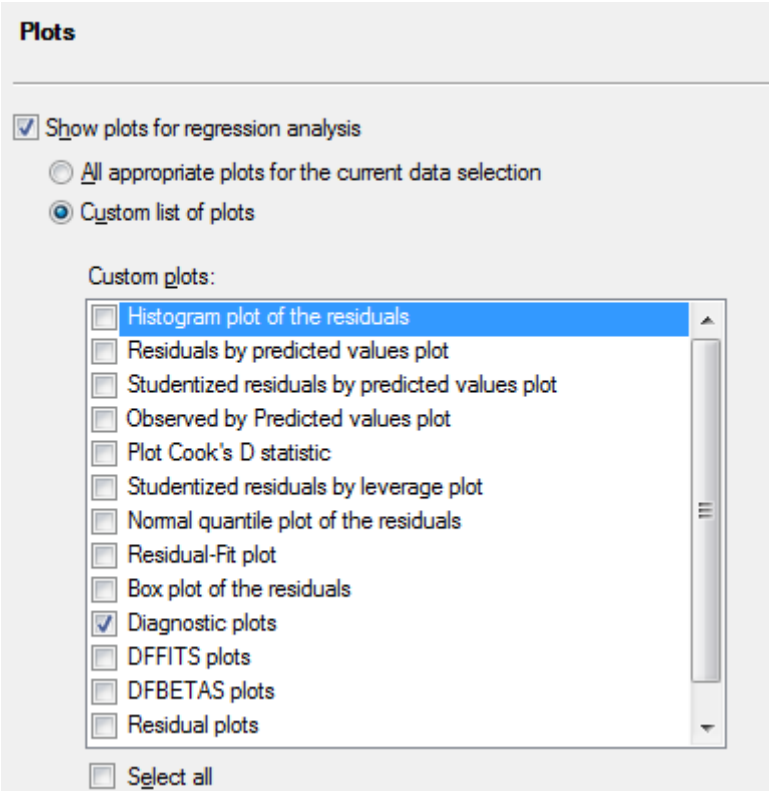
In the **Model** tab you can select the method of modeling. The **Full model fitted** option creates a model with all the variables you have assigned in the **Data** tab. Other quite commonly used options are the three next methods. The **Forward selection** option starts with no variables in the model and adds variables one by one to the model by comparing the p-values. The **Backward selection** option starts with all variables in the model and deletes variables by comparing the p-values. The **Stepwise selection** method is similar to the forward selection method except that variables already in the model do not necessarily stay there. Variables are added or deleted by comparing the p-values.



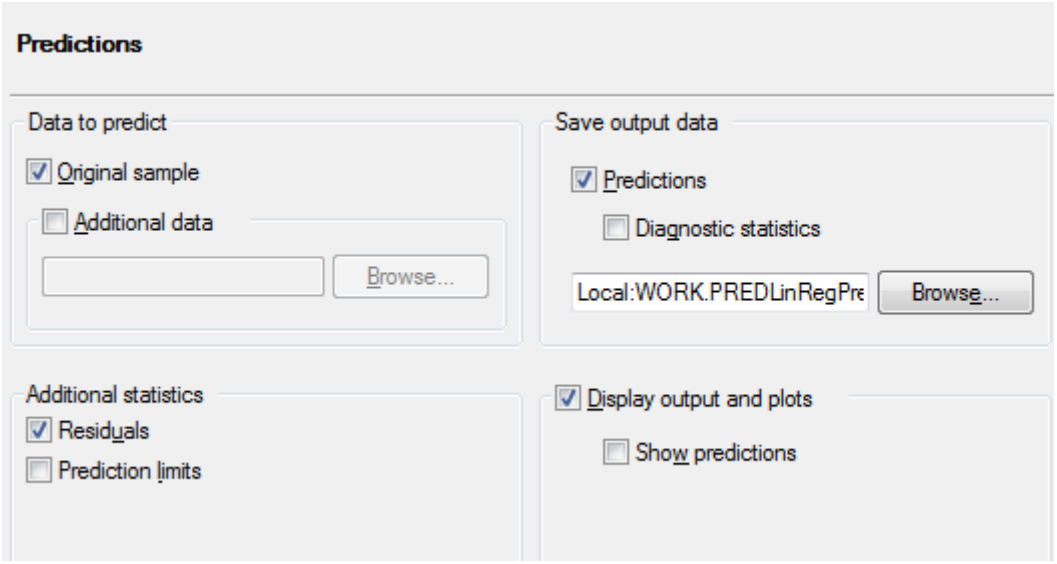
In the **Statistics** tab you can select some statistics for example to detect whether or not the explanatory variables are highly correlated (options **Collinearity analysis**, **Tolerance values for estimates** and **Variance inflation values**).



In the **Plots** tab the default option **All appropriate plots** for **the current data selection** creates a vast amount of different plots to examine the properties of the model. With the **Custom list of plots** option you can select those plots you want to include in the results.



With the **Predictions** tab you can save the predicted values and residuals in a new result data set.



In the first results table the F test measures whether or not there is some sense in the regression model. The hypotheses for this test are

H_0: The regression coefficients β_i are all 0.
H_1: At least one of the regression coefficients is not 0.

The p-value of the F test is now less than 0.0001, so we can now reject the H_0 at 0.1% significance level.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2022.11778	1011.05889	13.74	<.0001
Error	68	5005.06532	73.60390		
Corrected Total	70	7027.18310			

The next table then shows that the coefficient of determination (often denoted R^2) is 0.2878, so almost 29% of the total variation in *work2* can be explained by this regression model.

Root MSE	8.57927	R-Square	0.2878
Dependent Mean	15.69014	Adj R-Sq	0.2668
Coeff Var	54.67938		

The **Parameter Estimates** table shows the estimated regression coefficients, and thus the estimated regression model is now

$$\hat{y} = 0.10192 + 0.95302 \cdot \text{age} - 0.46827 \cdot \text{attend2}\,.$$

So, as every one year increase in *age*, increases the value of *work2* on average by 0.95 hours and every on hour increase in *attend2*, decreases the value of *work2* on average by 0.47 hours.

The t tests test whether or not we can assume a single regression coefficient to be zero. So, the hypotheses for those tests are

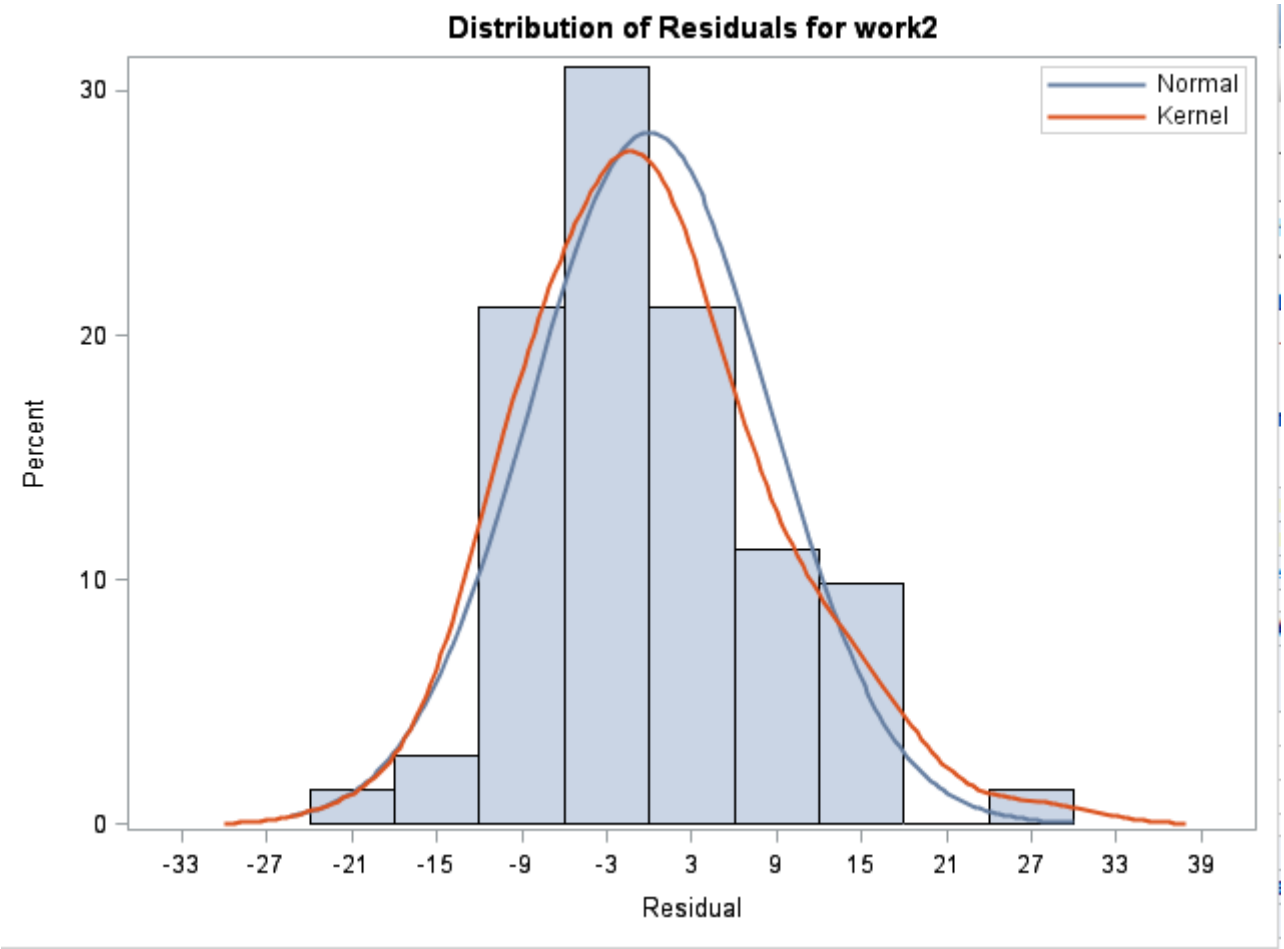
H_0: $\beta_i = 0$.
H_1: $\beta_i \neq 0$.

Because the p-values for the explanatory variables are <0.0001 and 0.0294, and they both are less than 0.05, we can now reject the H_0 for both of the explanatory variables.

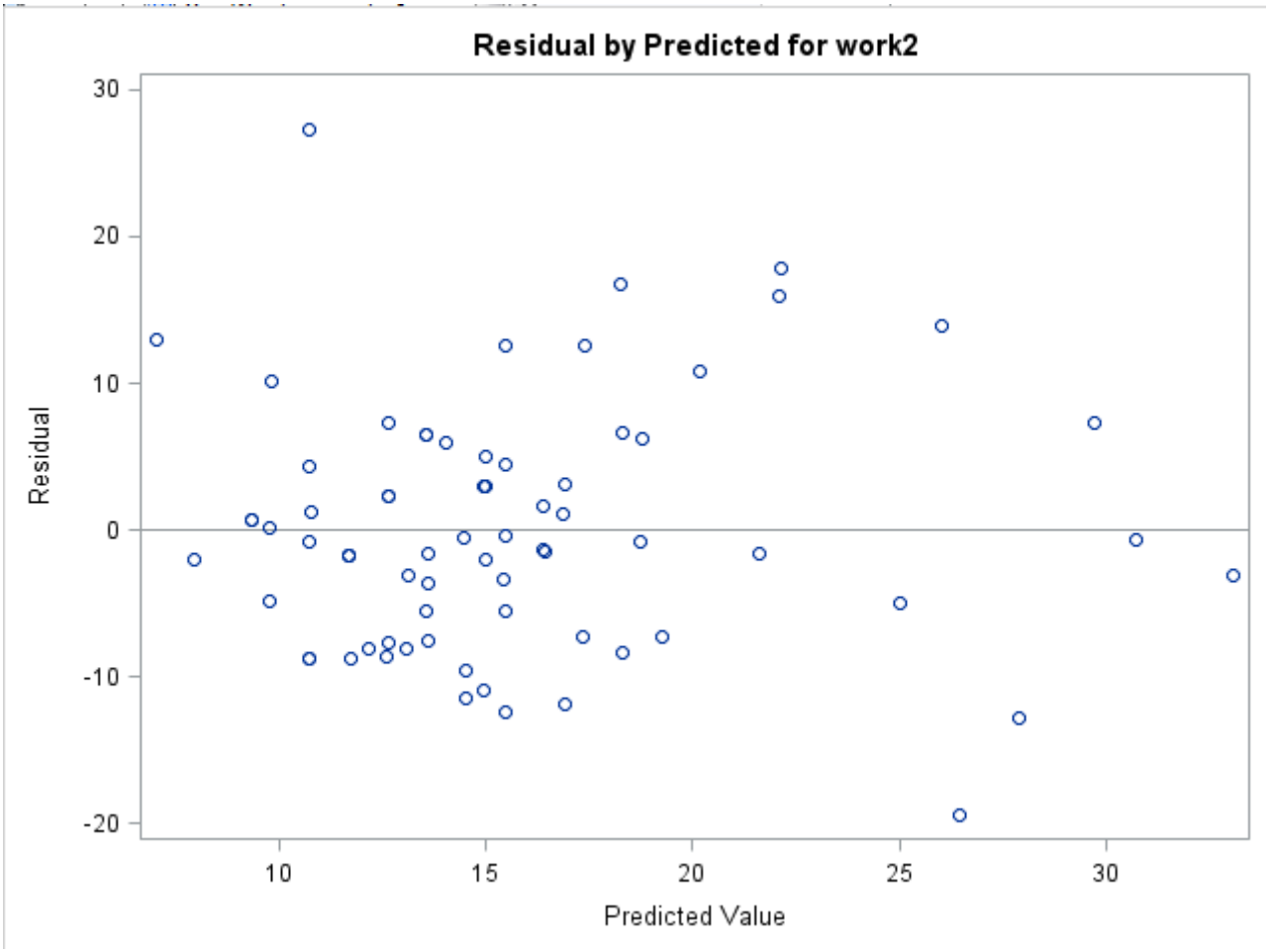
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.10192	6.61416	0.02	0.9878
age	Age in years	1	0.95302	0.21474	4.44	<.0001
attend2	How many hours you did attend lectures etc. last week?	1	-0.46827	0.21046	-2.22	0.0294

In the plot **Distribution of Residuals**, the residuals are the differences between the observed values of the dependent variable (y) and the predicted values ($\hat{y}$). If the model behaves well, the residuals should be (roughly) normally distributed with a mean of 0 and some constant variance. Now, the distribution of residuals seems to be a bit skewed.



In the scatter plot **Residual by Predicted** the data points are scattered randomly about 0, regardless of the size of the predicted value. This means that the residuals have a constant variance and now the residuals can be called homoscedastic .



In ordinary linear regression the explanatory variables should be quantitative ones, but you can run a linear regression with so called dummy variables which represent the classes of a qualitative variable. Dummy variables have always just two levels: 0 and 1. The number of dummy variables should be one less than the number of the classes of the qualitative variable. So let's assume that in our dataset we have a variable "relationship status" where the classes are: 1 = totally single, 2 = dating, 3 = firm relationship and we want to add such a variable as an explanatory variable in a regression model. First we need to create (with Query Builder) two dummy variables: dummysingle and dummydating and the levels for these new dummies are

Dummysingle =	1, if relationship status is 1
	0, if relationship status is 2 or 3
Dummydating =	1, if relationship status is 2
	0, if relationship status is 1 or 3

And then instead of the original " relationship status" the explanatory variables are these two new dummies.