

SPSS 4: Inferential Statistics

After completing this exercise you will be able to:

Use the following inferential statistical techniques in SPSS to generalise from a sample to the population:

- Chi Square test
- T-Test
- Correlation
- Regression

Background Information

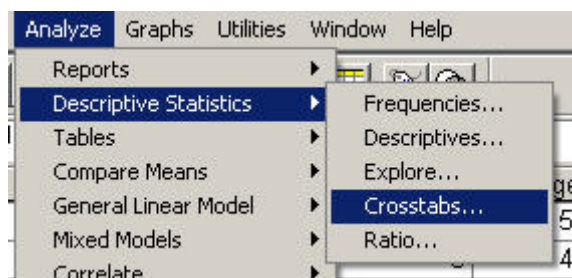
The data for this exercise is taken from the US Current Population Survey of 1985 (http://lib.stat.cmu.edu/datasets/CPS_85_Wages). You can download it as an SPSS file workData.sav from the module webpage for use with this exercise.

Chi-Square Test

A Chi-Square test tests the independence of two categorical variables.

For example, if the Gender and Occupational Category variables are independent we would expect to find a similar proportion of females and males in each Occupational Category.

we will make the assumption that Gender and Employment Category are independent and use a Chi-Square test to test that hypothesis.

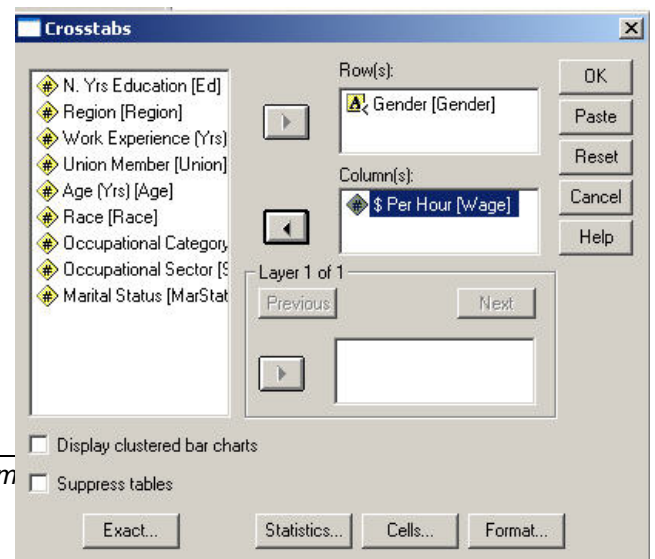


Start by selecting Analyze > Descriptive Statistics > Crosstabs...

This will open the Crosstabs dialog.

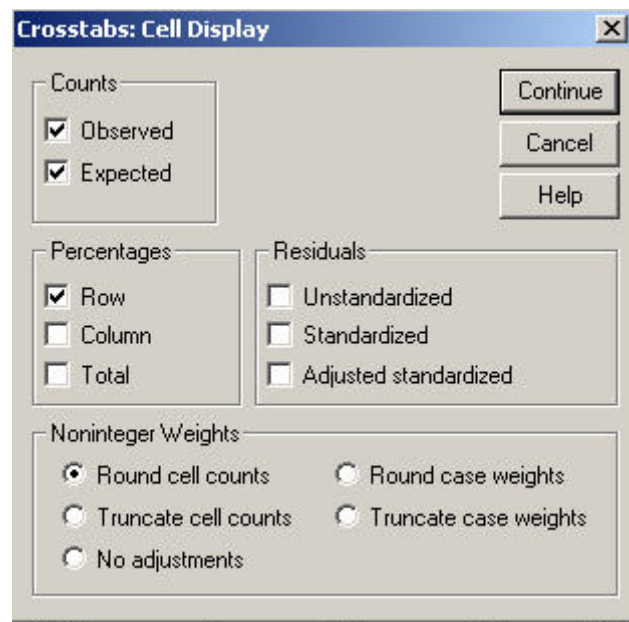
Select the variables you want to compare.

Move the Gender variable to the Row(s) box and the Occupational Category variable to the Column(s) box. The resulting crosstab table will then have two rows (one for each Gender value) and six columns (one for each Occupational Category).

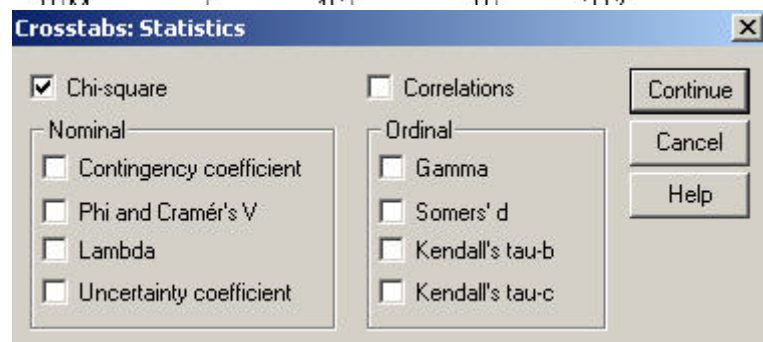


*School of Computer Science & Information Systems
Birkbeck College
Feb 2007*

Click the Cells... tab to open the Crosstabs: Cell Display dialog. Here we can select the values to display for each Gender/Occupational Category combination.



Check the boxes for Observed and Expected Counts and Row and Column Percentages and then click Continue to return to the Crosstabs dialog box.



Now click the Statistics... tab to open the Crosstabs: Statistics dialog and check the Chi Square box. Again, click Continue to return to the Crosstabs dialog box.

Finally, click OK in the crosstabs box to display the output.

Three tables will be displayed in the Outputwindow. The first of these - Case Processing Summary – is of no particular concern here.

The second table is a crosstab showing the actual and expected counts and the percentages for Female and Male within each Occupational Category, as shown below.

Gender * Occupational Category Crosstabulation

		Occupational Category						Total	
		Managem ent	Sales	Clerical	Service	Professio nal	Other		
Gender	F	Count	21	17	76	49	52	30	245
		Expected Count	25.2	17.4	44.5	38.1	48.2	71.6	245.0
		% within Gender	8.6%	6.9%	31.0%	20.0%	21.2%	12.2 %	100.0%
M		Count	34	21	21	34	53	126	289
		Expected Count	29.8	20.6	52.5	44.9	56.8	84.4	289.0
		% within Gender	11.8%	7.3%	7.3%	11.8%	18.3%	43.6 %	100.0%
Total		Count	55	38	97	83	105	156	534
		Expected Count	55.0	38.0	97.0	83.0	105.0	156.0	534.0
		% within Gender	10.3%	7.1%	18.2%	15.5%	19.7%	29.2 %	100.0%

The final column (Total) shows that the proportion of Female and Male employees in the data is very similar. However, Females seem to be over-represented in the Clerical category. The actual count (76) is almost double the expected count (44.5) and the proportion of all Females employed in this category (31%) is far higher than that for Males (7.3%)

What other observations can you make about the data based on this table?

Although these observations hold true for our sample data we need to test the statistical significance of our findings before we can generalise them to the population.

To do this we use the results displayed in the Chi-Square test table (below).

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	93.486 ^a	5	.000
Likelihood Ratio	99.269	5	.000
N of Valid Cases	534		

^a 0 cells (.0%) have expected count less than 5. The minimum expected count is 17.43.

In this table, the large Chi-Square statistic and the low significance indicate that there is a relationship between Gender and Employment Category.

Now test the hypothesis that Gender is not related to Race – i.e. we would expect a similar number of males and females in each ethnic group.

You can probably also think of other hypotheses to test – but remember that you must use Ordinal categorical variables for the Chi-Square test.

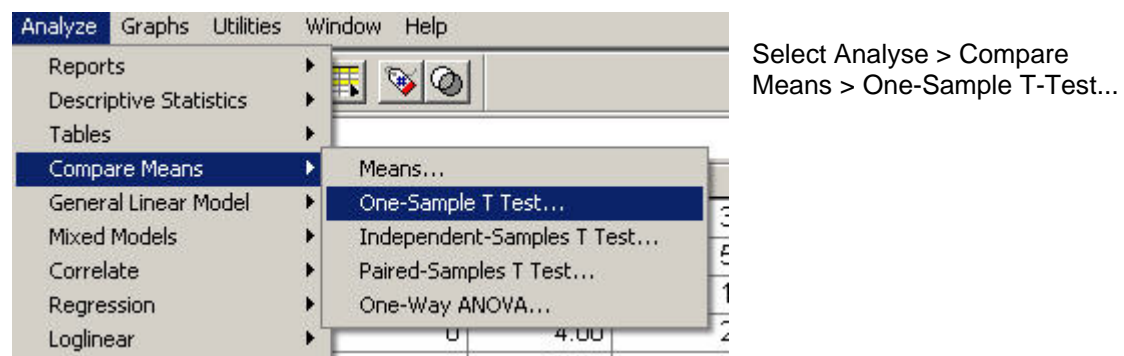
T-Test

A T-Test compares the mean values of two sets of numbers. This comparison allows you to evaluate whether the difference between two means is statistically significant. There are three types of T-Test:

- One Sample
Compares a single sample with a population value
- Independent Sample
Compares scores of two groups on the same variable
- Paired-samples
Compares the means of two variables within the same group

One-sample T-Test

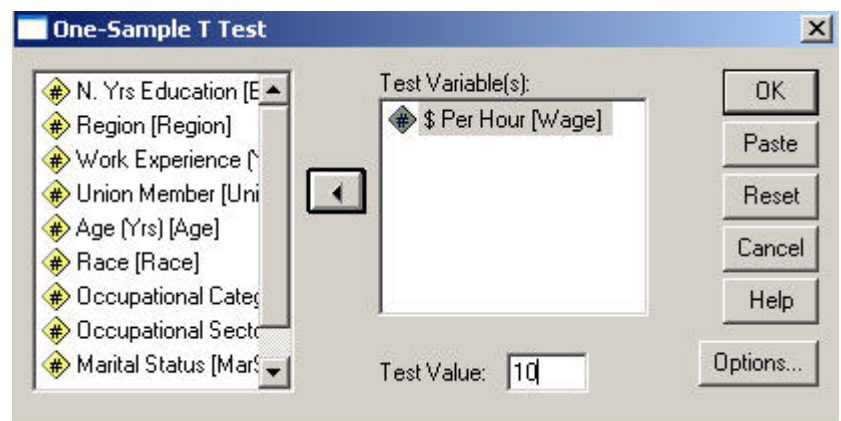
You will use a One-sample test to compare average hourly wages with national average.



Move the Wage variable into the Test Variable(s) box.

For this test we want to compare the mean of this value with the national average hourly wage. As we don't actually know what that was at the time this data was collected we will imagine that it was \$10. You can experiment with other values for this later.

Enter the value 10 in the Test Value box and click OK.



Two tables will be displayed in the Output window (below). The first shows descriptive statistics for the data you are analysing. From it we can see that the mean of our data is just over 9, slightly less than the test value of 10 that we entered.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
\$ Per Hour	534	9.0241	5.13910	.22239

The second table shows the statistics produced for our T-Test.

One-Sample Test

	Test Value = 10					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper	
\$ Per Hour	-4.388	533	.000	-.97594	-1.4128	-.5391	

observed t statistic for each sample, calculated as the ratio of the mean difference divided by the standard error of the sample mean.

degrees of freedom - number of cases minus 1.

sample mean - test value

probability of obtaining an absolute value greater than or equal to the observed t statistic, if the difference between the sample mean and the test value is purely random.

estimate of the boundaries between which the true mean difference lies in 95% of all possible random samples

As the confidence intervals both lie below 0.0 you can safely say that the mean of the wages in the sample data is significantly lower than that of the national average.

The low significance level tells us that the probability that there is no difference between our sample wages and the national average is very small: specifically, less than one time in a thousand would we obtain a mean difference of around a dollar or less between these wages if there were really no difference.

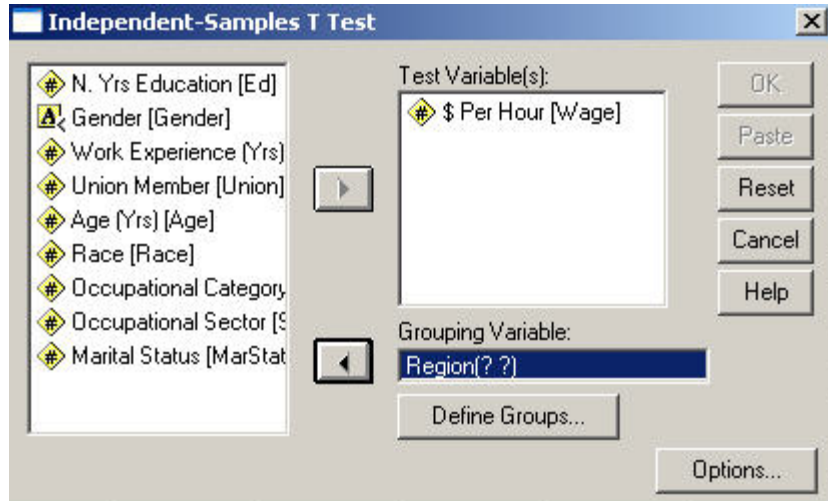
Try this T-Test again. First use 9 (the mean of our sample data) as your test value. What effect would you expect this to have on the lower and upper boundaries?

Now try the test again with a value that is lower than that of the sample data (say 4). What effect would you expect this to have on the boundaries?

Independent Samples T-Test

You will now use an Independent-Samples T Test to compare the wages of workers in the South with workers elsewhere.

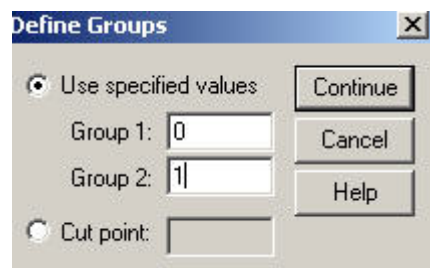
Select Analyze > Compare Means > Independent-Samples T Test to open the Independent-Samples T Test dialog.



Insert the Variable whose values you want to compare into the Test Variable(s) box. In this case we want to compare the wages of the two groups of workers so the test variable is Wages.

Insert the variable containing the groups you want to compare into the Grouping Variable box. Note that before you define the groups to be compared the variable name will be followed by (??).

Click Define Groups,,, to open the Define Groups dialog and insert the values of the groups



you want to compare into the Group 1 and Group 2 boxes. The Region variable has a value of 1 for workers coming from the South and 0 for those from elsewhere so in this case you simply enter 0 and 1 into the boxes.

If there were more than two possible values you would need to choose the two you wanted to compare as the Independent-Samples T Test is only used to compare two groups.

Click Continue to return to the previous dialog box and then click OK to display the statistics from your T Test in the Output Window.

Two tables will be displayed.

Group Statistics

	Region	N	Mean	Std. Deviation	Std. Error Mean
\$ Per Hour	Other	378	9.4892	5.22893	.26895
	South	156	7.8969	4.74435	.37985

The second shows the statistics for the T-Test.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference		
										Lower	Upper
\$ Per Hour	Equal variances assumed	.879	.349	3.286	532	.001	1.59231	.48461	.64032	2.54430	
	Equal variances not assumed			3.421	316.642	.001	1.59231	.46543	.67659	2.50803	

The table contains two sets of two sets of analyses: the first assumes equal variances and the second does not. In this example, it makes little difference which sets of scores we use as they are very similar. However, SPSS includes Levene's test for Equality of Variances to enable us to assess which set of scores is most appropriate for the analysis we are performing. In this case, since it assesses the significance level of the F-Score to be > 0.05 (the significance level with 95% confidence) we can use the Equal variances assumed scores.

As both the lower and upper bounds are > 0 we can conclude that wages in the South are lower than those in the rest of the country.

Use an independent samples T test to test whether there is any significant difference in the wages earned by union and non-union members.

What if you wanted to compare wages in different occupational sectors? There are three possible values –manufacturing, construction or other. You can only compare two at a time. Use the independent samples T Test to compare the manufacturing and construction sector wages.

In all the examples above, the Grouping variable has been ordinal or nominal. In other words it has always contained a number of predefined groups. What if we wanted to use a scale variable as the grouping variable? For example, we now want to conduct a T Test to find out whether workers aged 30 or over earn higher wages than those aged less than 30. To do this, construct the T Test as for the previous examples except this time use the Cut Point option of the Define Groups dialog. Enter a value of 30 to divide your data into two groups. In the Group Statistics table you will notice that the data has been divided into two groups: ≥ 30 and < 30 . What conclusions can you draw from the T Test statistics about the wages of these two groups?

Paired-samples T Test.

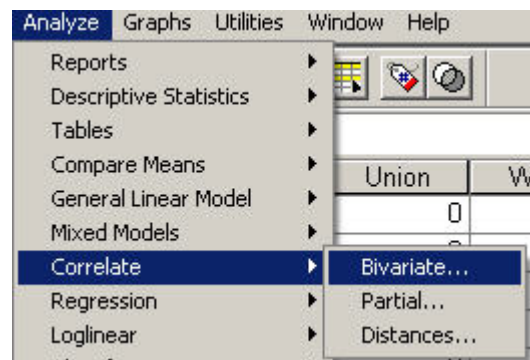
There is no data really suitable for a paired-samples test in this data. However, you could try using it to test the hypothesis that there is no difference between people's age and their years of work experience.

Correlation

Correlation measures the linear relationship between two variables. We used a scatterplot to display this for pairs of variables in the previous exercises

A correlation coefficient has a value ranging from -1 to 1. Values that are closer to the absolute value of 1 indicate that there is a strong relationship between the variables being correlated whereas values closer to 0 indicate that there is little or no linear relationship. The sign of a correlation coefficient describes the type of relationship between the variables being correlated. A positive correlation coefficient indicates that there is a positive linear relationship between the variables: as one variable increases in value, so does the other. A negative value indicates a negative linear relationship between variables: as one variable increases in value, the other variable decreases in value.

We would intuitively expect that the values for years of education, work experience, wages, would increase proportionally with each other.

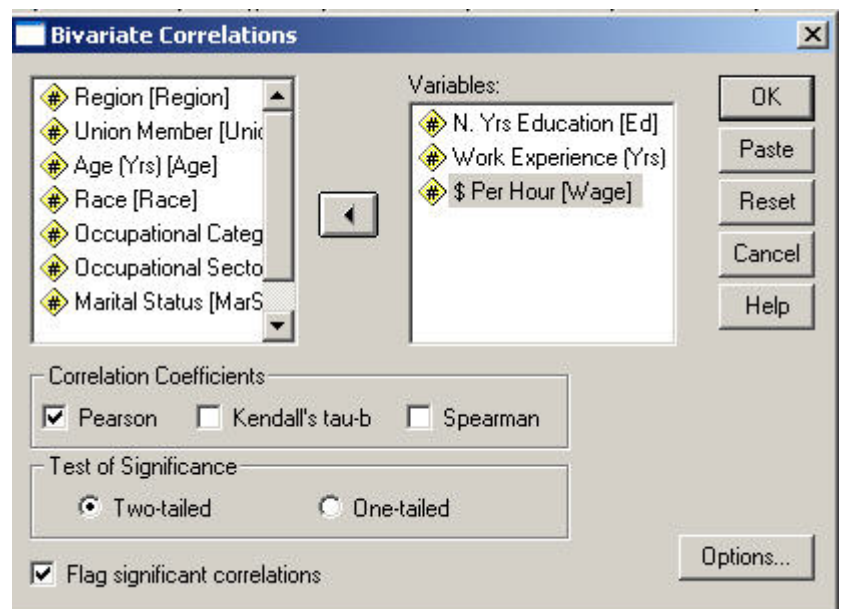


Select Analyze > Correlate > Bivariate to open the Bivariate correlations dialog box.

Move the variables you want to compare into the Variables box. Each variable listed in the *Variables* box will be correlated with every other variable in the box. In this case we are comparing the Years of Education, Work Experience and Wages variables.

We can simply accept the default selections for the other options and click OK to display the correlation table in the Output window.

The Correlation table for these variables is shown below.



Correlations

		N. Yrs Education	Work Experience (Yrs)	\$ Per Hour
N. Yrs Education	Pearson Correlation	1	-.353(**)	.382(**)
	Sig. (2-tailed)	.	.000	.000
	N	534	534	534
Work Experience (Yrs)	Pearson Correlation	-.353(**)	1	.087(*)
	Sig. (2-tailed)	.000	.	.044
	N	534	534	534
\$ Per Hour	Pearson Correlation	.382(**)	.087(*)	1
	Sig. (2-tailed)	.000	.044	.
	N	534	534	534

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

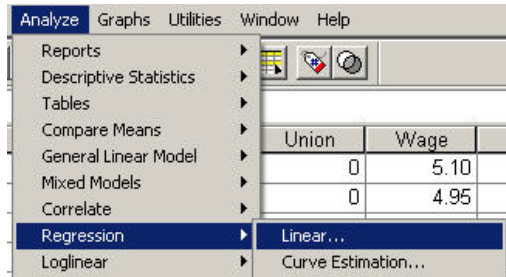
This output gives us a correlation matrix for the three correlations we requested. Note that despite there being nine cells in the above matrix, there are only three correlation coefficients of interest: (1) the correlation between years of education and work experience, (2) the correlation between work experience and wages, and (3) the correlation between wages and years of education. The reason only three of the nine correlations are of interest is because the diagonal consists of correlations of each variable with itself, always resulting in a value of 1.00 and the values on each side of the diagonal replicate the values on the opposite side of the diagonal.

For example, the three unique correlation coefficients show there is a positive correlation between workers' number of years of education and their current wages. This positive correlation coefficient (.382) indicates that there is a statistically significant ($p < .001$) linear relationship between these two variables such that the more education a person has, the greater their wages. Also observe that there is a statistically significant ($p < .001$) negative correlation coefficient (-.353) for the association between education level and work experience, indicating that the linear relationship between these two variables is one in which the values of one variable decrease as the other increases. The third correlation coefficient (.087) also indicates a positive association between workers' wages and their previous work experience, although this correlation is fairly weak.

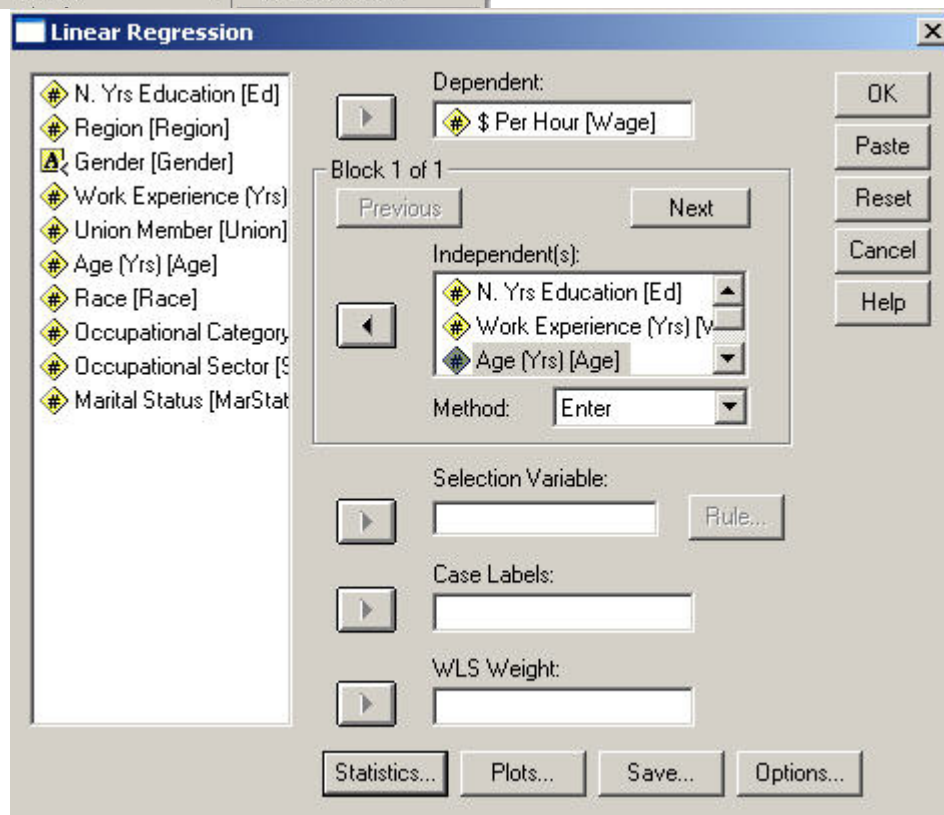
Regression

Regression is a technique that can be used to investigate the effect of one or more predictor variables on an outcome variable. Regression allows you to make statements about how well one or more independent variables will predict the value of a dependent variable.

We will use this technique to examine how well educational experience, work experience and age can predict the wages a worker can expect to earn.



Select Analyze > Regression > Linear... to open the Linear Regression dialog.



The Dependent variable is the variable whose outcome we want to predict – in this case Wage. Move this to the Dependent box.

The independent variables are the predictor variables – the variables whose effect on the dependent variable we want to examine. In this case we want to use Yrs Education, Work Experience and Age. Move these variables into the Independent(s) box.

Click OK to display the regression tables in the output window.

Four tables will be displayed. The first simply summarises the variables you are using for the regression. The next table, the model summary table (shown below) reports the strength of the relationship between the model and the dependent variable. The first statistic, R , is the multiple correlation coefficient between all of the predictor variables and the dependent variable. In this model, the value is .449, which indicates that there is a reasonable degree of variance shared by the independent variables and the dependent variables. The next value, R Square, is simply the squared value of R . This is frequently used to describe the goodness-of-fit or the amount of variance explained by a given set of predictor variables. In this example, the value is .202, which indicates that 20% of the variance in the dependent variable is explained by the independent variables in the model

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.449(a)	.202	.198	4.60370

a Predictors: (Constant), Work Experience (Yrs), N. Yrs Education, Age (Yrs)

The ANOVA table is a useful test of the model's ability to explain any variation in the dependent variable

The second table in the output is an ANOVA table. This is a useful test of the model's ability to explain any variation in the dependent variable. The F statistic represents a test of the null hypothesis that the expected values of the regression coefficients are equal to each other and that they equal zero. Put another way, this F statistic tests whether the R square proportion of variance in the dependent variable accounted for by the predictors is zero. If the null hypothesis were true, then that would indicate that there is not a regression relationship between the dependent variable and the predictor variables. But, instead, it appears that the four predictor variables in the present example are not all equal to each other and could be used to predict the dependent variable, current salary, as is indicated by a reasonably large F value and a small significance level.

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2843.850	3	947.950	44.727	.000(a)
	Residual	11232.849	530	21.194		
	Total	14076.699	533			

a Predictors: (Constant), Work Experience (Yrs), N. Yrs Education, Age (Yrs)

b Dependent Variable: \$ Per Hour

The Regression row shows the variation that is accounted for by our model. The Residual row shows the variation that is not accounted for by our model. In this case, the sum of squares shows that less than a quarter of the variation is accounted for by our model. However, the low significance value of the F statistic indicates that the variation in the model is not due to chance.

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4.770	7.043		-.677	.499
	N. Yrs Education	.948	1.155	.483	.821	.412
	Age (Yrs)	-.022	1.155	-.051	-.019	.985
	Work Experience (Yrs)	.128	1.156	.307	.110	.912

a. Dependent Variable: \$ Per Hour

The coefficients table provides information about the effects of individual predictor variables. From this we can see that Years of Education is the best predictor of current wages.