# LAB 3 INSTRUCTIONS

# SIMPLE LINEAR REGRESSION

In this lab you will first learn how to display the relationship between two quantitative variables with a scatterplot and also how to measure the strength of a linear relationship between two quantitative variables with correlation. Then you will learn how to use the regression tool in SPSS. In particular, you will study how to estimate the slope and intercept of the regression line and the mean of the response variable for each value of the explanatory variable. You will also learn how to apply the diagnostic tools for linear regression available in SPSS.

We will demonstrate some basic features of SPSS using an example of reaction times in a simple computer exercise.

**Example:** Reaction time (RT) is the elapsed time between the presentation of a sensory stimulus to a subject and the subsequent response. RT is often used in experimental psychology to measure the duration of mental operations**.** A computer-based system is used to measure a subject's reaction time. The software displays a small circle at a random location on the computer screen. The subject tries to click in the circle with the mouse as quickly as possible. A new circle appears as soon as the subject clicks the old one. The time required by the subject to click in the new circle is recorded.

The data file *times.sav* gives data for the subject's trials, 20 with each hand. The data can be downloaded to your workstation or a computer by clicking the hyperlink below.

Here is the description of the variables in the data file:

| **Column** | **Variable Name** | **Description of Variable** |
|---|---|---|
| 1 | time | Time required to click in the new circle (in milliseconds), |
| 2 | distance | Distance from the cursor location to the center of the new circle (in units whose actual size depends on the size of the screen), |
| 3 | hand | 1 if right hand, 2 if left hand. |

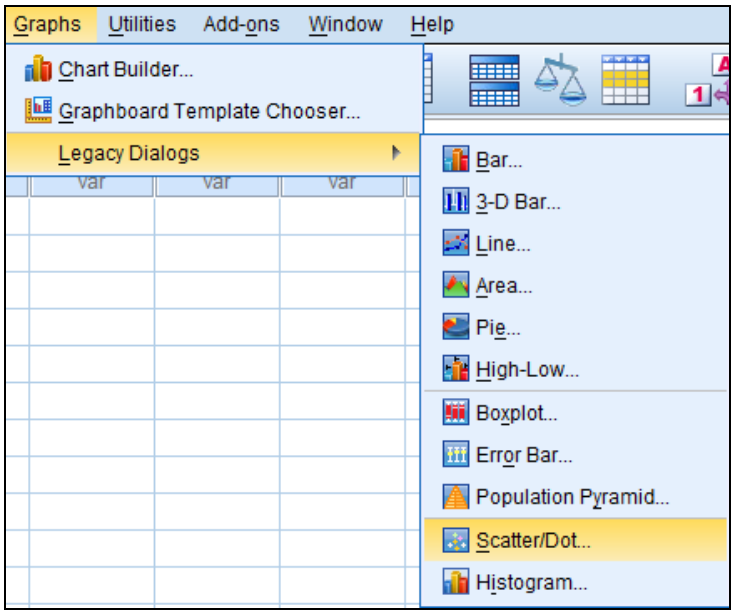
DOWNLOAD DATA

You will use simple linear regression in SPSS to examine the relationship between time and distance for each hand and compare the performance of the right and left hand in the exercise.

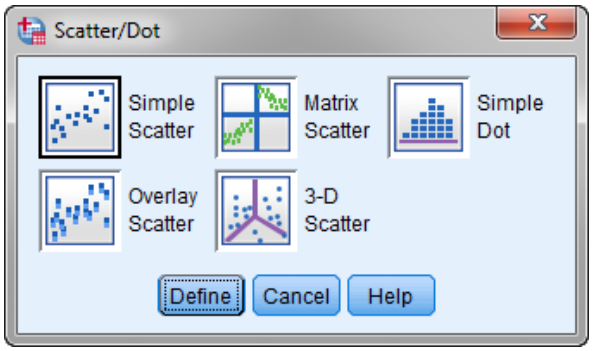
## 1. Scatterplot

Scatterplots allow to visualize the relationship between two quantitative variables and to evaluate the form, direction and strength of the relationship. Scatterplots have already been discussed in the Introductory Lab, Section 11. You learned there how to obtain and edit a scatterplot in SPSS. For your convenience here we will demonstrate how to obtain a scatterplot of reaction times versus distance for the two hands.

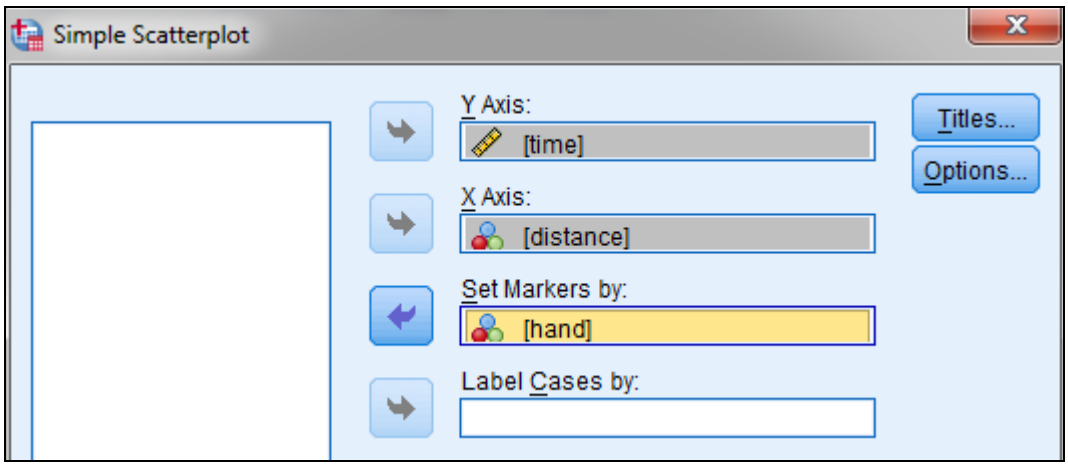
Click the *Graphs→Legacy Dialogs→Scatter/Dot* option in the resulting pull-down menu.




The following Scatter/Dot dialog box is obtained:



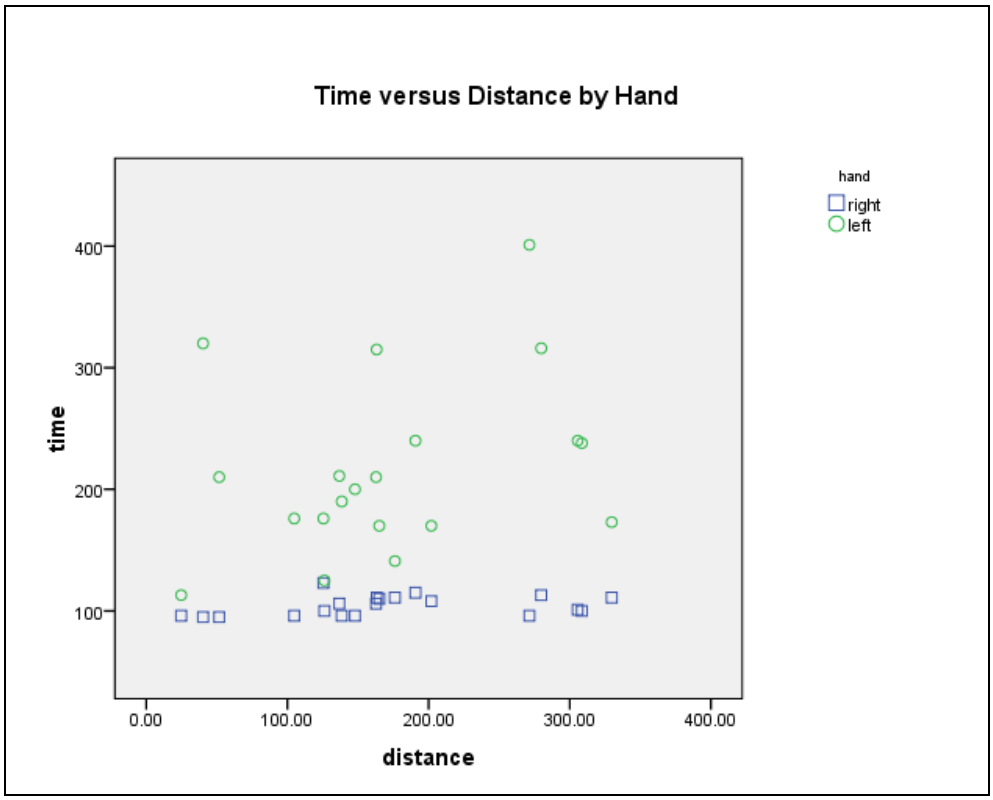

Choose *Simple Scatter* as the form of the desired scatterplot in the dialog box above. Then click *Define*. You will obtain the *Simple Scatterplot* dialog box. In order to obtain a scatterplot of *time* vs. *distance* by *hand*, fill out the dialog box as follows:



Click the *Titles* tab and enter the title of the plot. Click OK. The scatterplot will be displayed in the contents pane of the *Viewer* window.

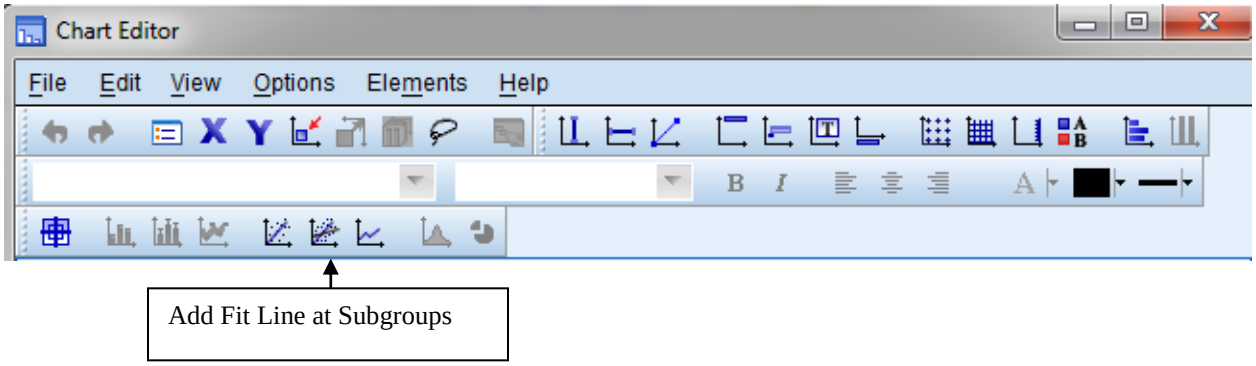
We will replace the small blue open circles for the right-hand observations by blue small squares to better distinguish between right-hand and left-hand observations.

Double-click the scatterplot in the *Viewer* window to open the *Chart Editor* window. Then double-click one of the small open circles corresponding to the right-hand observations in the *Chart Editor* window or equivalently double-click the small circle for the right-hand observations in the legend on the right. This opens the *Properties* window in the *Marker* style window. Change the blue open circles to blue open squares for the right-hand observations. Finally, the following scatterplot is obtained:

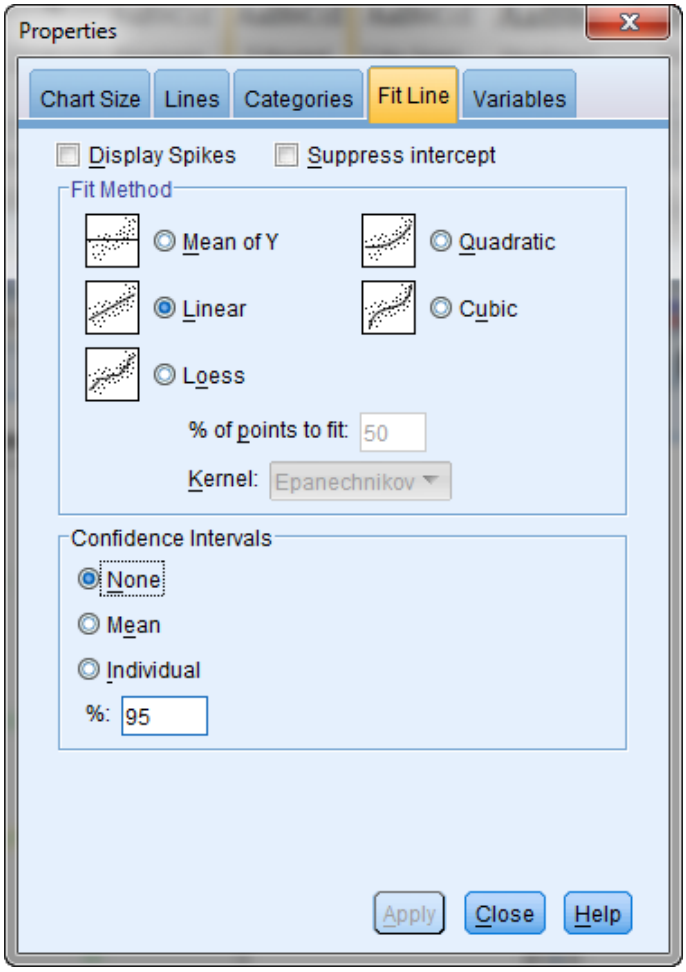



As you can see the right hand observations lie below the left-hand observations. This means the right-hand times are shorter, so the subject is right-handed. There is no striking pattern for the left-hand observations; the pattern for the right-hand observations is obscured because the points are squeezed at the bottom of the plot. Notice a relatively small variation in time values for the right hand. There is a weak positive linear relationship between *time* and *distance* for each hand.

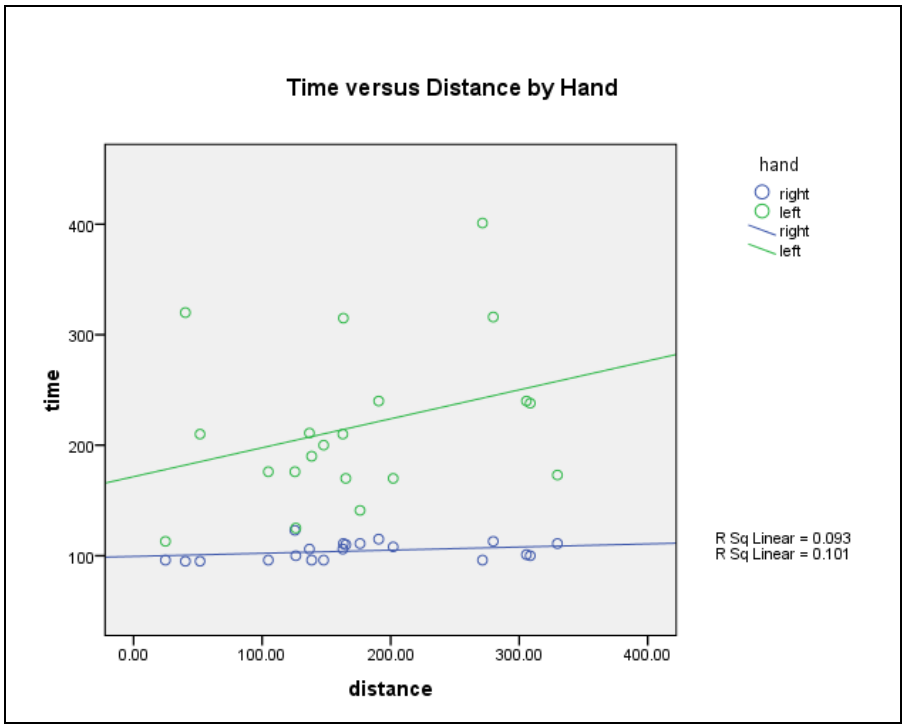
You can edit the scatterplot further by clicking the items in the *Chart Editor* menu (*Edit*, *Options*, or *Elements*) or on the buttons in the toolbar. Place the mouse pointer over each button to see a brief description of the tool. For example, you can fit a line for each hand by clicking click the button in the toolbar labelled "*Add Fit Line at Subgroups*"



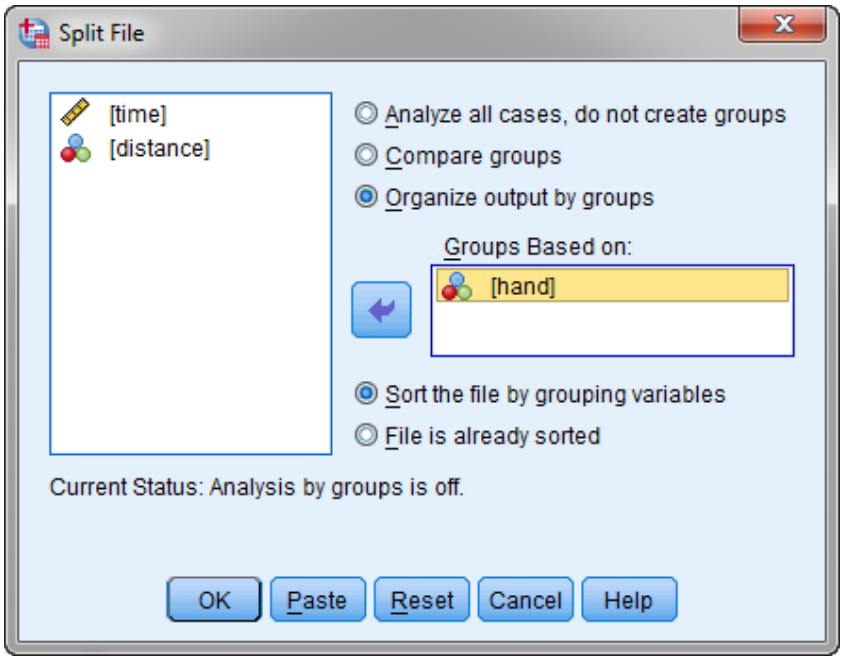

Clicking on the button "*Add Fit Line at Subgroups*" opens the following *Properties* dialog box.



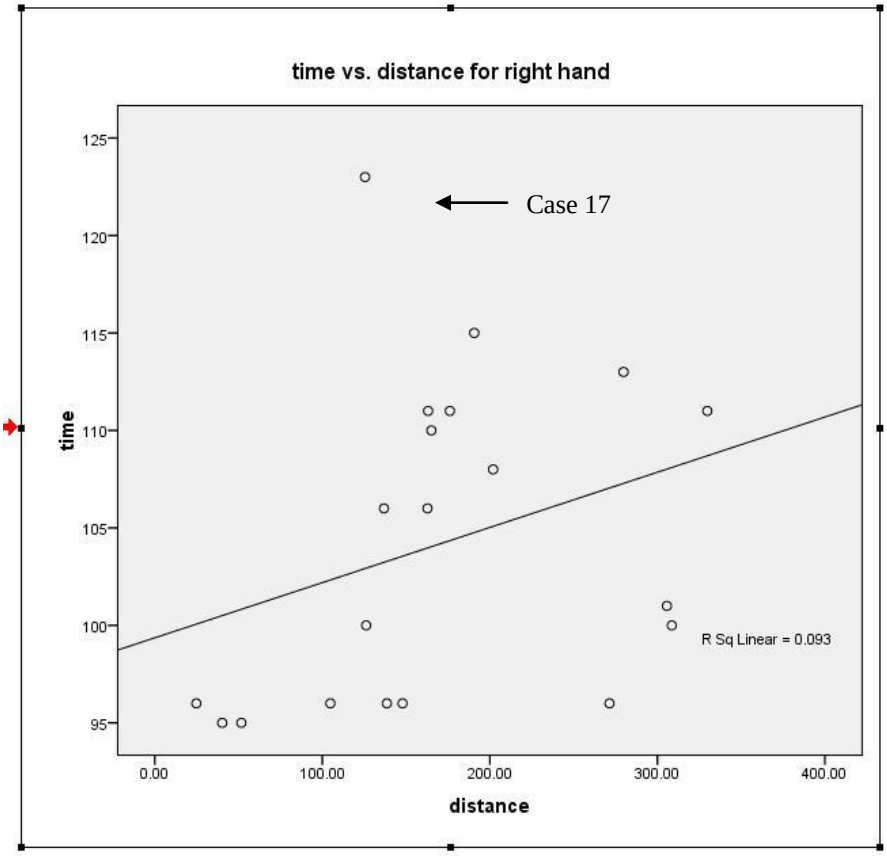

Make sure that the *Fit Line* tab is active (highlighted) in the window and check the *Linear* radio button. The new scatterplot with a fitted line for each group will be obtained.




As the ranges of the variable *time* for the two hands are very different, you may also obtain a separate scatterplot of *time* versus *distance* for each hand to better evaluate the relationship (you will have to split the file first; this operation can be performed with the *Split File* feature in SPSS discussed already in the *Introductory Lab*). This is especially important for the right hand as in this case the response variable time varies in a relatively small range.




The scatterplot of time vs. distance for the right hand with the estimated regression line is shown on the next page.




Notice that the case 17 at the top of the scaterplot is likely to produce a large residual (studentized residual). There are no influentail observations in the plot.

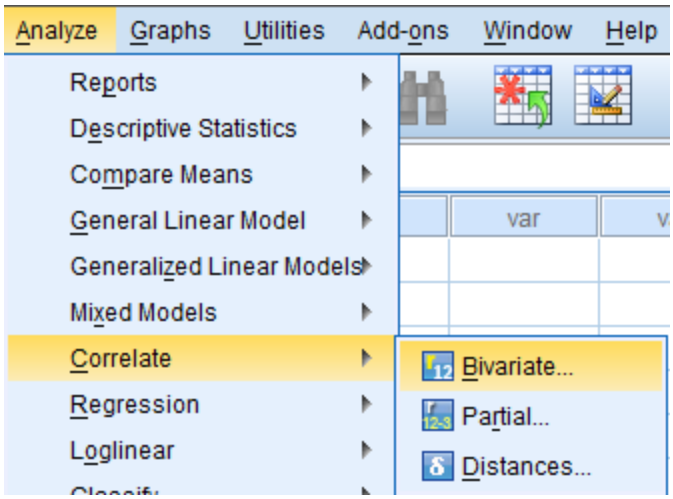
## 2. Correlation

The correlation r describes the direction of a linear relationship between two quantitative variables (positive or negative) and measures the strength of the relationship. It is a number between -1 and +1. The closer the value of correlation is to ±1, the close the data points fall to a straight line, and the stronger is the linear association.

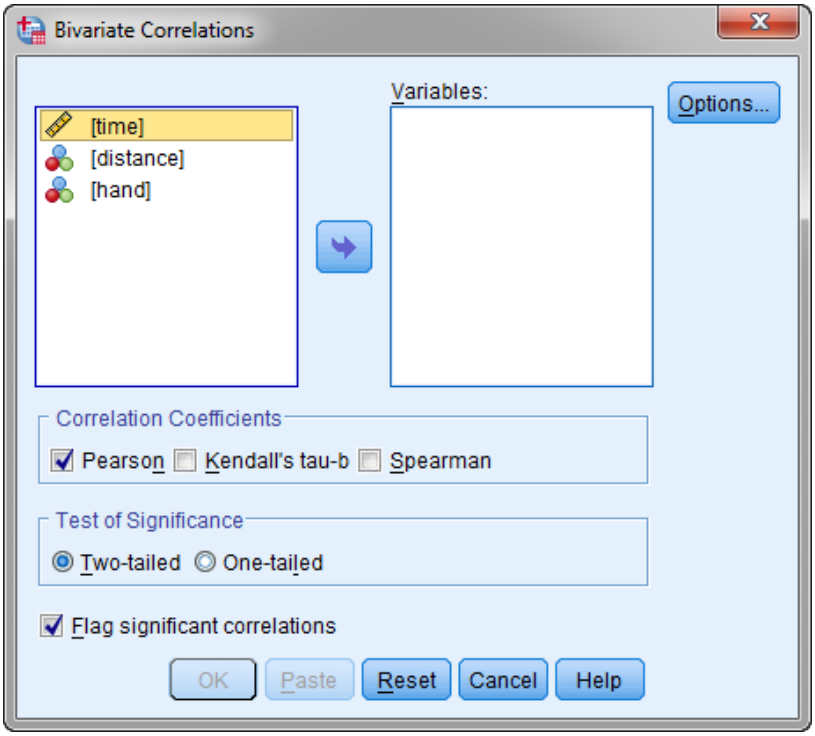
We will demonstrate how to obtain the correlation between *time* and *distance* in SPSS, separately for right and left hand. In order to obtain the correlation coefficients for the two variables separately for right and left hand, it is necessary to split the original file into two subsets corresponding to the observations for each hand. This operation can be performed with the *Split File* feature in SPSS discussed already in the *Introductory Lab*. Your file is already splitted (see the dialog box on page 5) if you obtained separate scatterplots of time versus distance in the previous section.

Click *Data* in the main menu bar, then on *Split File* from the pull-down menu. This opens the *Split File* dialog box as below. Select the *Organize output by groups* radio button, and then select and move the *Hand* variable to the *Groups Based on* box. Click *OK.*

Now we are ready to obtain the correlation coefficients for each hand separately. Select *Analyze→Correlate→Bivariate* option.




The following dialog box is obtained:




Move the variables *time* and *distance* (the variables for which correlation is to be computed) into the *Variables* box. Also make sure that the *Pearson* box is checked (Pearson coefficient is a common measure of the correlation between two quantitative variables) and the *Flag significant correlations* box is checked too (Pearson coefficient is a statistic which estimates the correlation of the two given random variables).

The following output (one for each hand) will be displayed:

**= right**

**Correlations[a]**

| | | | |
|---|---|---|---|
| | Pearson Correlation | 1.000 | .305 |
| | Sig. (2-tailed) | | .191 |
| | N | 20.000 | 20 |
| | Pearson Correlation | .305 | 1.000 |
| | Sig. (2-tailed) | .191 | |
| | N | 20 | 20.000 |

a. = right

**= left**

**Correlations[a]**

| | | | |
|---|---|---|---|
| | Pearson Correlation | 1.000 | .318 |
| | Sig. (2-tailed) | | .171 |
| | N | 20.000 | 20 |
| | Pearson Correlation | .318 | 1.000 |
| | Sig. (2-tailed) | .171 | |
| | N | 20 | 20.000 |

a. = left

The positive values of the correlation coefficients confirm positive linear relationship between time and distance for each hand.

### 3. Linear Regression in SPSS

In this section you will learn how to express the linear relationship between time and distance by fitting a regression model to the data and make predictions about time for a given distance. Let us regress the variable *time* on *distance* for each hand separately. Separate regression line for each hand is justified by the different pattern of the relationship between *time* and *distance* for each hand observed in the scatterplot. The linear regression model for our data can be defined as follows

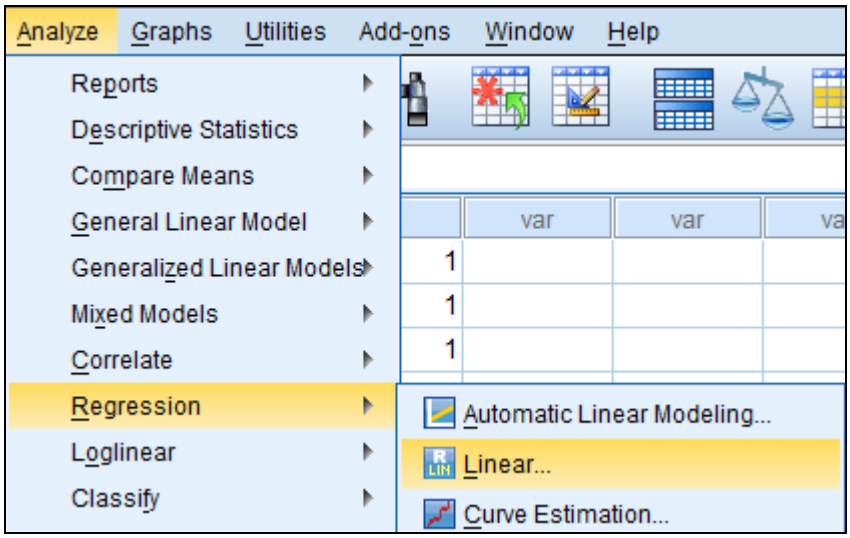
$$\mu(time) = \beta_0 + \beta_1 \cdot \text{distance}.$$

The model can also be rewritten in the equivalent form as

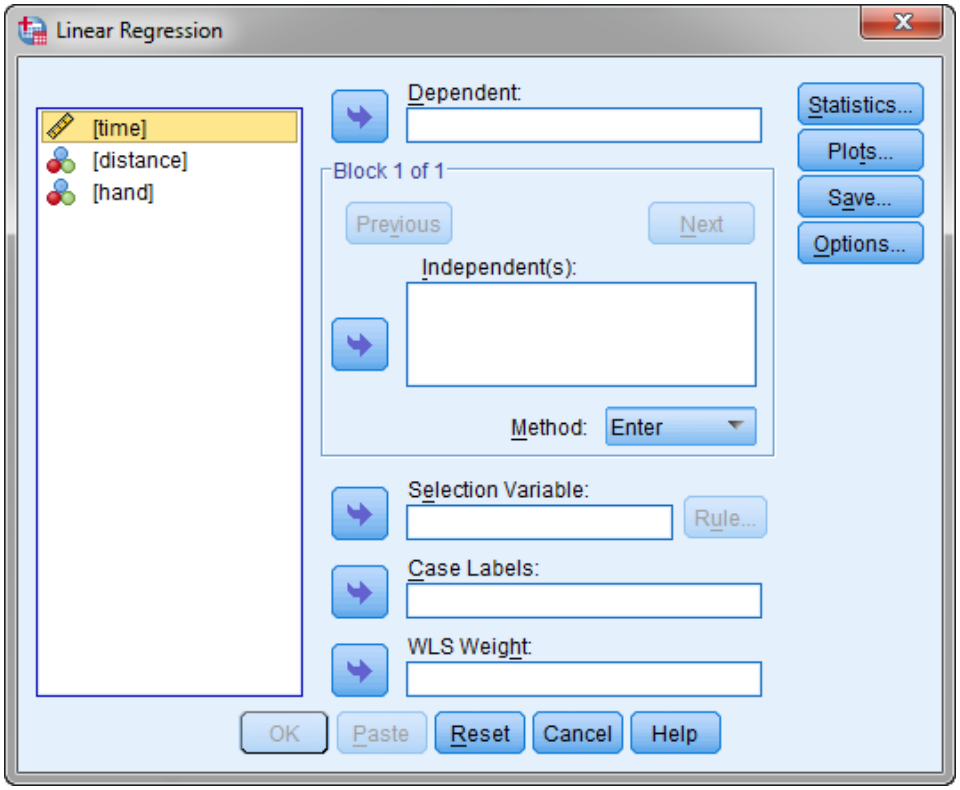
$$time = \beta_0 + \beta_1 \cdot \text{distance} + ERROR.$$

We assume here that *ERROR* follows a normal distribution for each value of *distance*, and the mean of the variable *ERROR* is zero. More, the standard deviation σ the *ERROR* variable is constant for each value of the explanatory variable *distance*.

Click on *Regression* in the *Analyze* menu, then on *Linear*.




This opens the *Linear Regression* dialog box displayed below.




Select and move the variable *time* into the *Dependent* box, and *distance* variable into the *Independent(s)* box. This performs a basic regression analysis and produces the value of $R^2$, an analysis of variance table for the regression, and the estimates of the regression line coefficients. Some of the output is produced by the two default options *Estimates* and *Model Fit* available in the dialog box obtained by clicking on the *Statistics*…button in the right panel of the window.

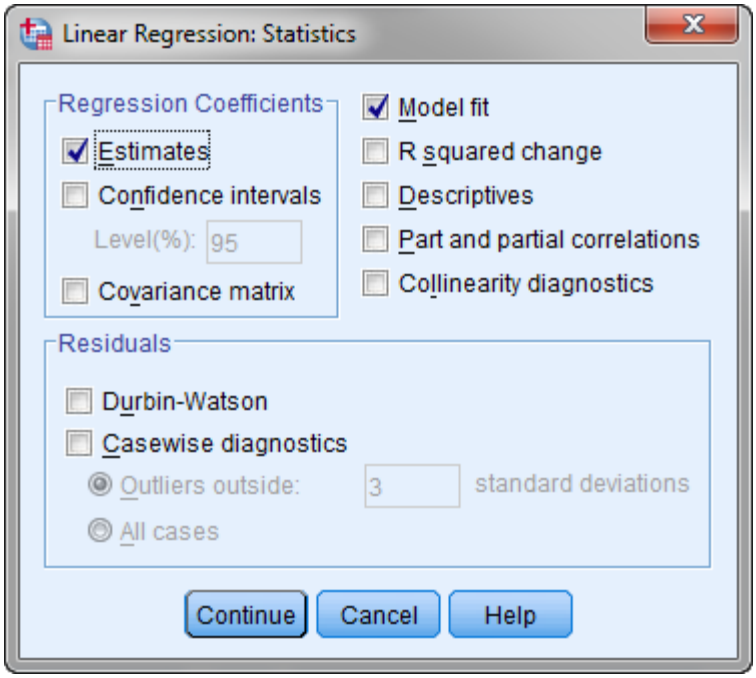
Now we will discuss in detail the additional information on the regression model available by clicking on one of the four buttons in the right panel of the Linear Regression window: *Statistics*…, *Plots*…, *Save*…, and *Options*….

First click the *Statistics* button. This opens the *Linear Regression: Statistics* dialog box. Notice that *Estimates* and *Model Fit* are the default options. With the *Estimates* option checked, the estimates of the regression coefficients $\beta_0$ and $\beta_1$ with their corresponding standard errors, the values of the t statistics and the two-tailed significance level of t are displayed. If the *Model Fit* option is checked, the model summary in the form of $R^2$ value, adjusted $R^2$ value, standard error of the estimate and ANOVA table will be displayed.

The $R^2$ shows the fraction of the variation in the response variable *time* that can be explained by the explanatory variable *distance*. The higher the value of $R^2$, the more powerful is the regression model to predict time given distance.

*Adjusted R Square* value is an estimate of how well your model would fit another data set from the same population. The value of adjusted $R^2$ is always less than or equal to the value of $R^2$ ; since the estimated regression coefficients are based on the values in this particular data set, the model fits the data somewhat better than it would another sample from the same population.

Standard error of the estimate is an estimate of the standard deviation $\sigma$ of the error term in the regression model $time = \beta_0 + \beta_1 \cdot \text{distance} + ERROR.$ The smaller the standard error, the more accurate and trustworthy the predictions based on the model are.



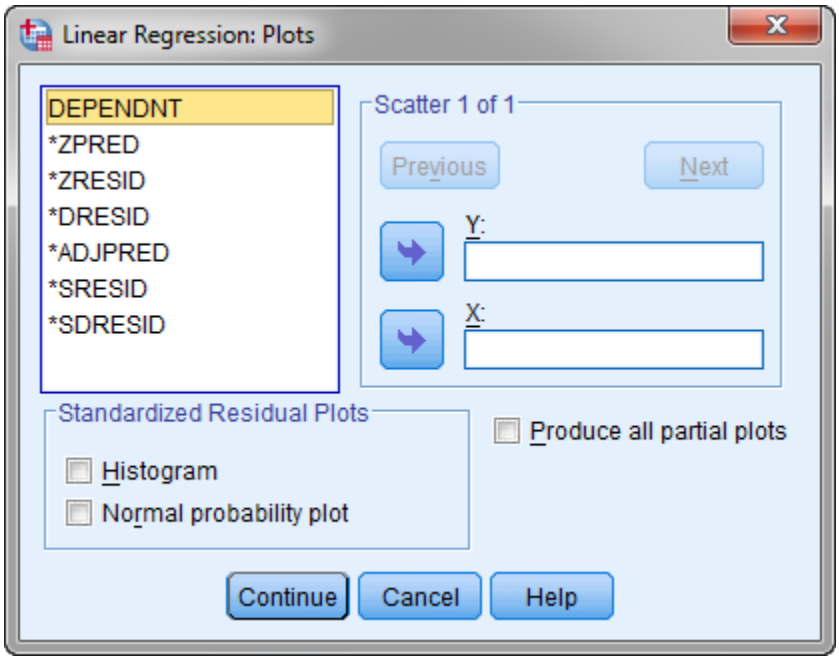

You may also check the *Confidence intervals* option in the above dialog box to obtain the 95% confidence intervals for the regression coefficients.

The option *R squared change* displays the change in $R^2$ when an explanatory variable is added or removed from a regression equation. If the change is substantial, the explanatory variable is a good predictor of the response variable.

If you check the option *Descriptives*, the number of valid cases, the mean, and the standard deviation for each variable in the analysis will be provided. Moreover, a correlation matrix with a one-tailed significance level will also be displayed.

Given the structure of the experiment (the observations are obtained in time order), it is worthy to check for autocorrelation (residuals for case i might correlate with residuals for case i-1) with Durbin-Watson (DW) statistic (the *Durbin-Watson* check box is included in the *Residuals* group in the above dialog box). If successive residuals are uncorrelated, DW=2. There is evidence of autocorrelation if either DW is substantially smaller than 2 or substantially larger than 2.

Now click the *Plots* button. This opens the *Linear Regression: Plots* dialog box.




You can use the dialog box to obtain scatterplots with any combination of the dependent variable (DEPENDNT) and any of the predicted values (*ZPRED, *ADJPRED) or residuals (*ZRESID, *DRESID, *SRESID, *SDRESID) listed in the left panel of the window.

To obtain a scatterplot of any two of these, select one of them and move into the Y (vertical axis) box and then select the other one and move it into the X (horizontal axis) box.

The following variables listed in the left panel are available:

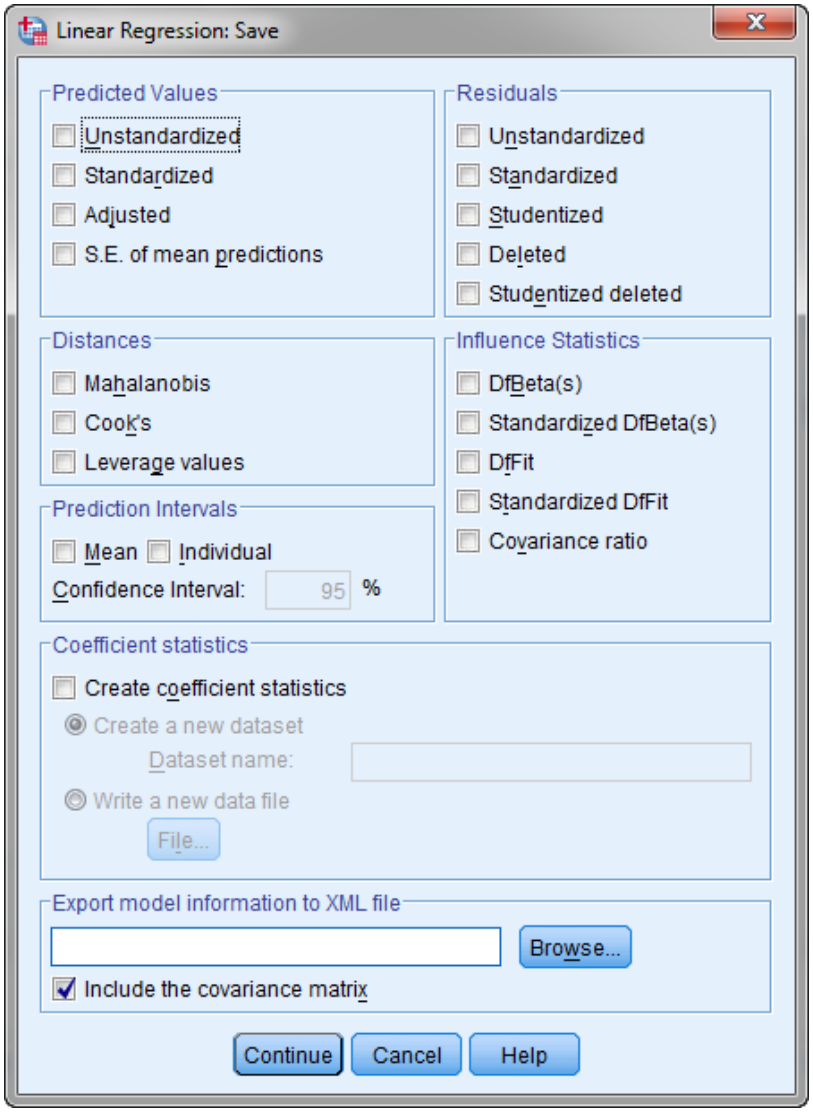
*ZPRED: The standardized predicted values of the dependent variable,
*ZRESID: The standardized residuals (the standardized residual is the residual divided by the estimated standard deviation of the residuals; note that there is no need to subtract the mean from the residual, since the mean of the residuals is 0),
*DRESID: Deleted residuals, (the residuals for a case when it is excluded from the regression computations),
*ADJPRED: Adjusted predicted values, the predicted value for a case when it is excluded from the regression computations,
*SRESID: Studentized residuals, (residuals divided by the corresponding estimate of its standard deviation that varies from case to case),
*SDRESID: Studentized deleted residuals, (the studentized residuals for a case when it is excluded from the regression computations).

The plot of residuals versus fitted (predicted) values can be used to detect nonlinearity (curved plot), non-constant variance (a fan-shaped or double-bow pattern) and the presence of outliers.

To obtain a plot of standardized residuals versus standardized predicted values, move **ZRESID* into the *Y:* box and **ZPRED* into the *X:* box.

To obtain a normal probability plot of standardized residuals, simply select the *Normal probability plot* check box. Moreover, a histogram of the standardized residuals can be obtained to help you check whether they are normally distributed. Click *Continue* to close the dialog box.

Click the *Save* button in the *Linear Regression* dialog box; this opens the *Linear Regression: Save* dialog box displayed below.




For each option you select in the above dialog box window, one or more variables will be added to the data file in the *Data Editor*.

The dialog box window is divided into several groups labelled *Predicted Values*, *Residuals*, *Distances*, *Influence Statistics*, *Prediction Intervals* and *Coefficient statistics*. We will discuss here the most important options within each group. Note that the statistics can be provided for any linear regression model not just model with one explanatory variable.

The *Predicted Values* group includes *Unstandardized* (the value predicted by the regression model for the response variable), *Standardized* (the predicted value for the response variable standardized to have a mean of 0 and a standard deviation of 1), *Adjusted* (the predicted value for a case if that case is excluded from the calculation of the regression coefficients) and *S.E. of mean predictions* (an estimate of the standard error of the mean predicted value).

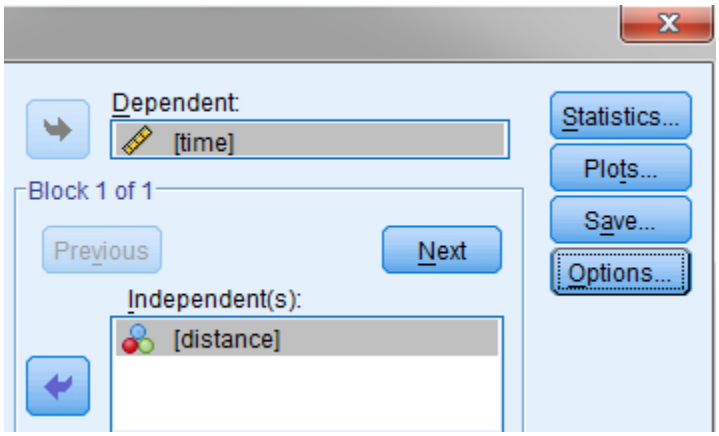
The *Distances* group includes *Mahalanobis* (a measure of the distance of a case from the means of all of the explanatory variables; this measure provides an indication of whether or not an observation is an outlier with respect to the explanatory variable values), *Cook's* (measures overall influence, the effect of omitting a case on the estimated regression coefficients; a large Cook's indicates that excluding the case from computation of the estimated regression model changes the coefficients substantially; a value close to or larger than 1 indicates a large influence), and *Leverage values* (measures the influence of a case on the fit of the regression; used to flagging cases with unusual explanatory variable values and potential influential cases).

The *Prediction Intervals* group includes *Mean* (two new variables representing lower and upper bounds for the prediction interval of the mean value of the response variable, for all cases with the given values of the explanatory variable) and *Individual* (tow new variables representing lower and upper bounds for the prediction interval for the response variable for a case with the given values of the explanatory variables). Below the two options, you can specify the level for the confidence interval by entering a percentage value greater than zero and less than 100. The prediction intervals are wider than the corresponding confidence interval for the mean values.

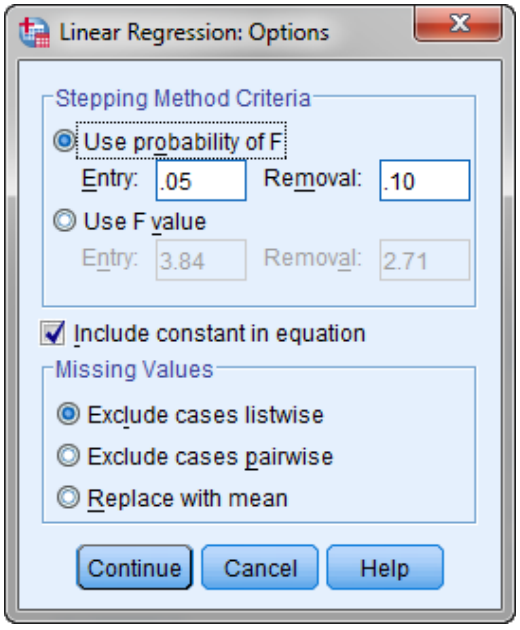
The *Residuals* group includes *Unstandardized* (the value of the response variable minus its predicted value from the fitted regression model), *Standardized* (the residual divided by an estimate of its standard error), *Studentized* (the residual divided by an estimate of its standard deviation that varies from case to case, useful for flagging outliers), and *Deleted* (the residual if the current case were excluded from the calculation of the regression coefficients), and *Studentized deleted* (the deleted residual divided by an estimate of its standard deviation).

In general, the *Influence Statistics* group display change in the regression coefficients and predicted values that results from the exclusion of a particular case. The group includes *DfBeta(s)* (a new variable for each explanatory variable in the regression model, including the constant, containing the change in the coefficient for that term if the current case were omitted from the calculations; it measures how much impact each observation has on a particular explanatory variable), *Standardized DFBeta(s)* (*DfBeta* value divided by an estimate of its standard deviation; you may want to flag cases with absolute values greater than 2 divided by the square root of N, where N is the number of cases.), and *DfFit* (the change in the predicted value of the response variable if the current case is omitted from the calculations).

In the *Linear Regression* dialog box, click *Options* button.



The *Linear Regression: Options* dialog box opens.




The first group in the dialog box is *Stepping Method Criteria*. These criteria do not apply to simple linear regression model. They will be discussed in detail in *Lab 4 Instructions*.

In some situations the regression line has to be forced through the origin so the regression model has a meaningful interpretation. For example, regression of return on investment must be forced through zero as return must be zero if investment is zero. Similarly, the regression line of alcohol blood content on number of drinks must pass through the origin. Leave the option *Include constant in equation* selected for an ordinary model with a constant in the regression equation (default option). Deselect it if you want to constrain the constant term to equal 0. This leads to regression through the origin. However, you should only force a regression line through the origin when there is a strong justification to do so in the underlying theory or the model.

The last group in the *Linear Regression: Options* dialog box provides the available treatments of missing data in the regression procedure. By default, SPSS excludes all cases that have missing values for any of the variables in the regression from the computation of the regression statistics. This is known as *listwise* deletion (*Exclude cases listwise* option). When pairwise deletion of missing data is selected (*Exclude cases pairwise*), then cases will be excluded from any calculations involving variables for which they have missing data.

The SPSS output for the right hand is shown below. Remember that the data were split to allow for separate regression for each hand.

**Model Summary[b,c]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .305[a] | .093 | .042 | 8.068 |

a. Predictors: (Constant),

b. = right

c. Dependent Variable:

The value of R=0.305 is the correlation coefficient between time and distance. Thus there is a weak correlation between *time* and *distance* for the right hand. As the *R Square* value here is 0.093, less than 10% of the variation in time can be explained by distance.

**ANOVA**[b,c]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 119.939 | 1 | 119.939 | 1.842 | .191[a] |
| | Residual | 1171.811 | 18 | 65.101 | | |
| | Total | 1291.750 | 19 | | | |

a. Predictors: (Constant),

b. = right

c. Dependent Variable:

**Coefficients**[a,b]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 99.364 | 4.026 | | 24.678 | .000 | 90.905 | 107.823 |
| | | .028 | .021 | .305 | 1.357 | .191 | -.016 | .072 |

a. = right

b. Dependent Variable:

The above ANOVA table contains the results of the F-test that examines the overall utility of the regression model. The test compares the simple linear regression model $\mu(time \mid \text{distance}) = \beta_0 + \beta_1 \cdot \text{distance}$ to the equal-means model $\mu(time \mid \text{distance}) = \beta_0.$ If the slope $\beta_1$ is zero, the simple linear regression model reduces to the equal-means model (the explanatory variable *distance* in no use as the predictor of *time*). The F test of the overall significance of the simple linear regression model is therefore equivalent to testing the null hypothesis about the slope $\beta_1$, $H_0$: $\beta_1 = 0$ versus the alternative $H_a$: $\beta_1 \neq 0$.

The F statistic for our regression model is equal to 1.842. The statistic F follows an F distribution with 1 degree of freedom for the numerator and 18 degrees of freedom for the denominator. The p-value of the two-sided test is 0.191. Therefore *distance* is not useful in predicting *time*.

In the *Coefficients* table that follows the ANOVA output the estimates $\hat{\beta}_0, \hat{\beta}_1$ of the regression coefficients and their standard errors are provided. Moreover, the values of the corresponding t statistics to test the regression coefficients $H_0$: $\beta_i = 0$ versus the alternative $H_a$: $\beta_i \neq 0$ for i=0, 1 and the two-sided p-values of the tests are also displayed.

The p-value of the two-sided t test about the slope $H_0$: $\beta_1 = 0$ versus the alternative $H_a$: $\beta_1 \neq 0$ is equal to 0.191. The value is identical to the p-value of the F-test provided in the above ANOVA table. It is also easy to verify that $F=t^2$, where t is the value of the t statistic to test the null hypothesis $H_0$: $\beta_1 = 0$ versus the alternative $H_a$: $\beta_1 \neq 0$.

As requested, the 95% confidence intervals for each regression coefficient are also provided. Notice that the 95% confidence interval for the slope $\beta_1$ contains zero; the result is consistent with the outcome of the t test about the slope. Based on the output, the estimated regression model is

$$\mu(time \mid \text{distance}) = 99.364 + .028 \cdot \text{distance}.$$

The residuals statistics for the model is displayed in the table on the next page.

**Residuals Statistics**[a,b]

| | Minimum | Maximum | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Predicted Value | 100.07 | 108.70 | 104.25 | 2.512 | 20 |
| Std. Predicted Value | -1.666 | 1.772 | .000 | 1.000 | 20 |
| Standard Error of Predicted Value | 1.806 | 3.743 | 2.458 | .703 | 20 |
| Adjusted Predicted Value | 101.06 | 109.80 | 104.47 | 2.624 | 20 |
| Residual | -11.049 | 20.082 | .000 | 7.853 | 20 |
| Std. Residual | -1.369 | 2.489 | .000 | .973 | 20 |
| Stud. Residual | -1.456 | 2.574 | -.013 | 1.020 | 20 |
| Deleted Residual | -12.490 | 21.474 | -.220 | 8.627 | 20 |
| Stud. Deleted Residual | -1.506 | 3.146 | .014 | 1.103 | 20 |
| Mahal. Distance | .002 | 3.139 | .950 | 1.093 | 20 |
| Cook's Distance | .002 | .230 | .050 | .057 | 20 |
| Centered Leverage Value | .000 | .165 | .050 | .058 | 20 |

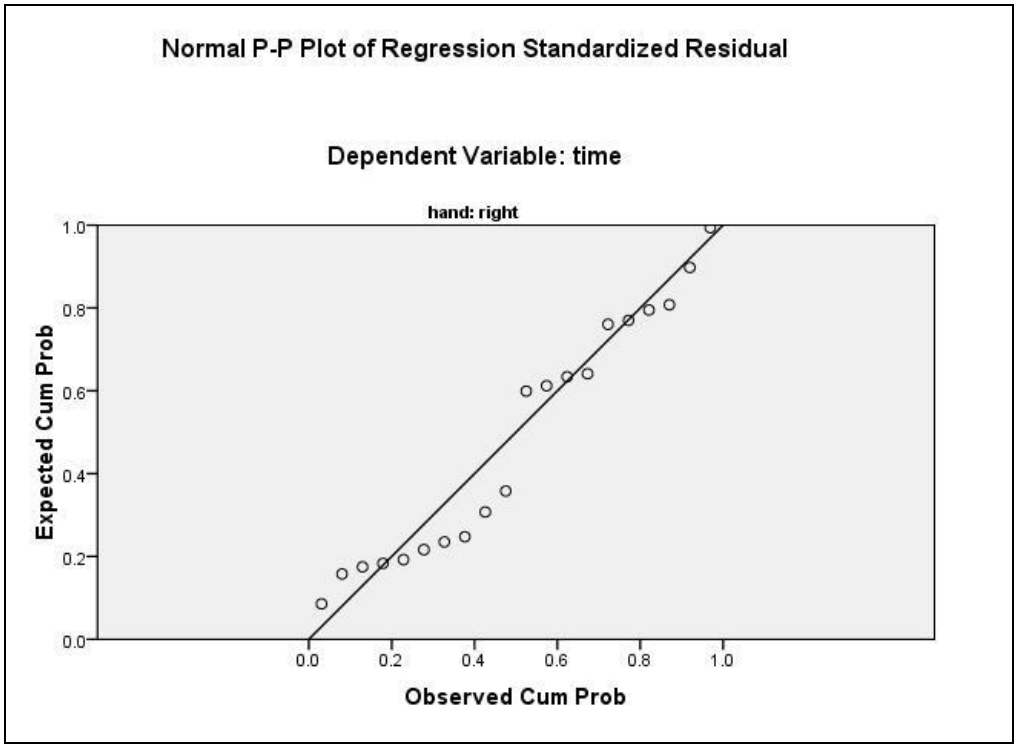
a. = right

b. Dependent Variable:

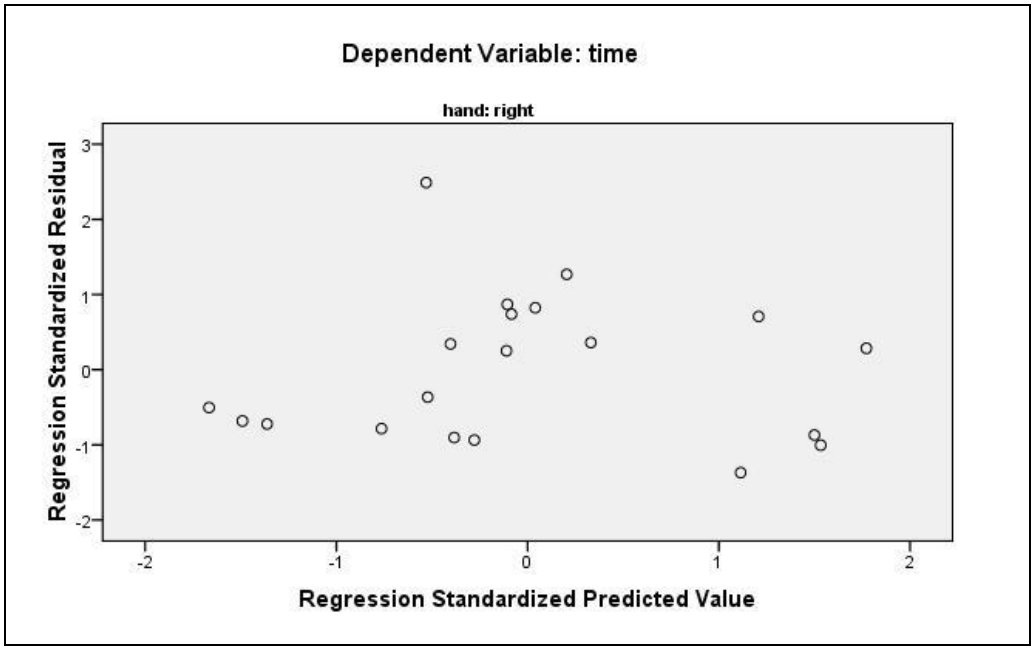
| | distance | hand | PRE_1 | RES_1 | SRE_1 | COO_1 | LEV_1 |
|---|---|---|---|---|---|---|---|
| 1 | 190.70 | 1 | 104.76309 | 10.23691 | 1.30322 | 0.04676 | 0.00219 |
| 2 | 138.52 | 1 | 103.28584 | -7.28584 | -0.93026 | 0.02652 | 0.00775 |
| 3 | 165.08 | 1 | 104.03777 | 5.96223 | 0.75830 | 0.01525 | 0.00038 |
| 4 | 126.19 | 1 | 102.93677 | -2.93677 | -0.37629 | 0.00487 | 0.01438 |
| 5 | 163.19 | 1 | 103.98426 | 7.01574 | 0.89239 | 0.02122 | 0.00059 |
| 6 | 305.66 | 1 | 108.01769 | -7.01769 | -0.95375 | 0.09207 | 0.11836 |
| 7 | 176.15 | 1 | 104.35117 | 6.64883 | 0.84549 | 0.01885 | 0.00009 |
| 8 | 162.78 | 1 | 103.97265 | 2.02735 | 0.25788 | 0.00177 | 0.00064 |
| 9 | 147.87 | 1 | 103.55054 | -7.55054 | -0.96218 | 0.02646 | 0.00408 |
| 10 | 271.46 | 1 | 107.04946 | -11.04946 | -1.45600 | 0.13820 | 0.06534 |
| 11 | 40.25 | 1 | 100.50374 | -5.50374 | -0.74739 | 0.05600 | 0.11701 |
| 12 | 24.76 | 1 | 100.06521 | -4.06521 | -0.56191 | 0.03849 | 0.14601 |
| 13 | 104.80 | 1 | 102.33120 | -6.33120 | -0.81840 | 0.02940 | 0.03070 |
| 14 | 136.80 | 1 | 103.23714 | 2.76286 | 0.35291 | 0.00387 | 0.00855 |
| 15 | 308.60 | 1 | 108.10092 | -8.10092 | -1.10448 | 0.12817 | 0.12364 |
| 16 | 279.80 | 1 | 107.28557 | 5.71443 | 0.75793 | 0.04172 | 0.07683 |
| 17 | 125.51 | 1 | 102.91751 | 20.08249 | 2.57379 | 0.22952 | 0.01480 |
| 18 | 329.80 | 1 | 108.70111 | 2.29889 | 0.32162 | 0.01418 | 0.16519 |
| 19 | 51.66 | 1 | 100.82677 | -5.82677 | -0.78224 | 0.05302 | 0.09770 |
| 20 | 201.95 | 1 | 105.08158 | 2.91842 | 0.37223 | 0.00409 | 0.00577 |

Notice a large residual (20.08249) and the studentized residual (2.57379) for the case 17. However, the leverage and the Cook's distance are relatively small for this case. The case is not influential. There are no influential observations in the data.

The normal probability plot is shown below:




All points in the above plot are reasonably close to a straight line. Thus the residuals follow approximately a normal distribution. The plot of standardized residuals versus standardized predicted values provided by SPSS is shown below:




There is no systematic pattern in the plot; the points seem to be randomly scattered about a horizontal line at zero. There is no evidence of substantial change in the spread of residuals over the range of the predicted values.

Henryk Kolacz
University of Alberta
September 2012