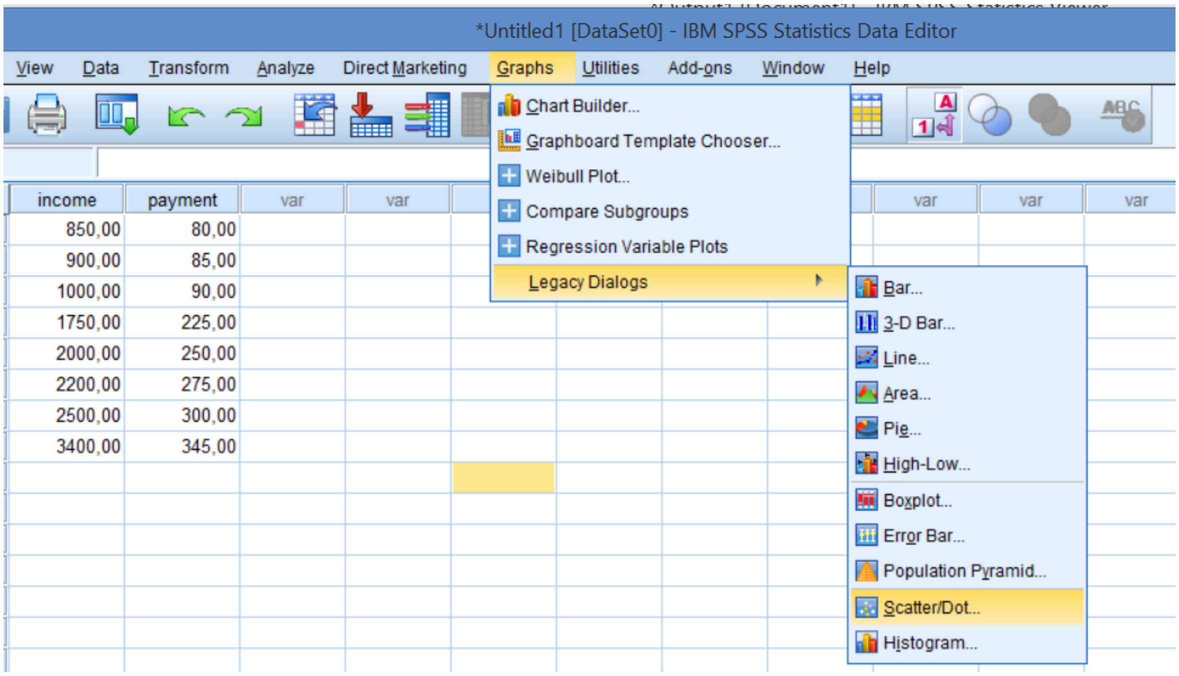
Example 1: The monthly incomes of randomly selected 8 families and their cultural payments were given in table below:

monthly incomes x_i	**cultural payments** y_i
850	80
900	85
1000	90
1750	225
2000	250
2200	275
2500	300
3400	345

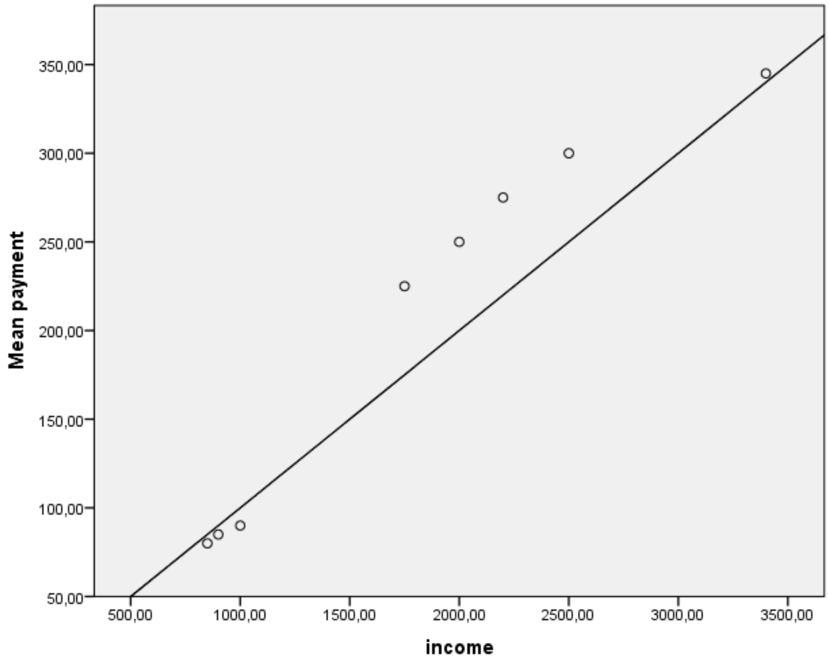
In order to do a regression analysis between two variables, we must look scatter plot to see if there is a linear relationship between these two variables, if not, we cannot conduct a regression analysis. Secondly, an assumption must be checked that whether if dependent variable (here cultural payments) has a normal distribution. If these two assumptions are satisfied, then we can continue to do regression analysis. So here first of all, these are checked and then the regression analysis is performed.

Graphs → Legacy Dialogs then choose Scatter/Dot then choose Simple Scatter then send the variables Y axis and X axis and obtain the scatter plot.

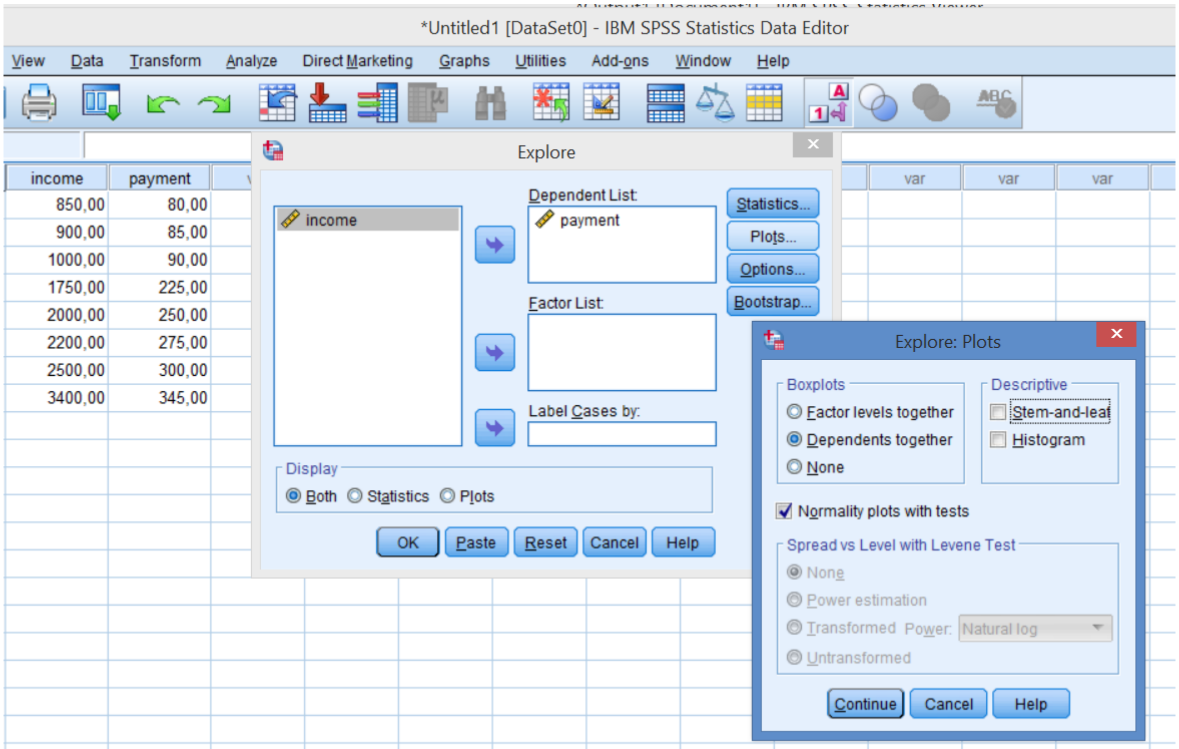


income	payment	var	var	var	var	var
850,00	80,00					
900,00	85,00					
1000,00	90,00					
1750,00	225,00					
2000,00	250,00					
2200,00	275,00					
2500,00	300,00					
3400,00	345,00					

Scatter plot of the data:



From scatter graph it is clear that there is a positive linear relationship between income and payment variables.



*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

income	payment
850,00	80,00
900,00	85,00
1000,00	90,00
1750,00	225,00
2000,00	250,00
2200,00	275,00
2500,00	300,00
3400,00	345,00

Explore

Dependent List: payment

Factor List:

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

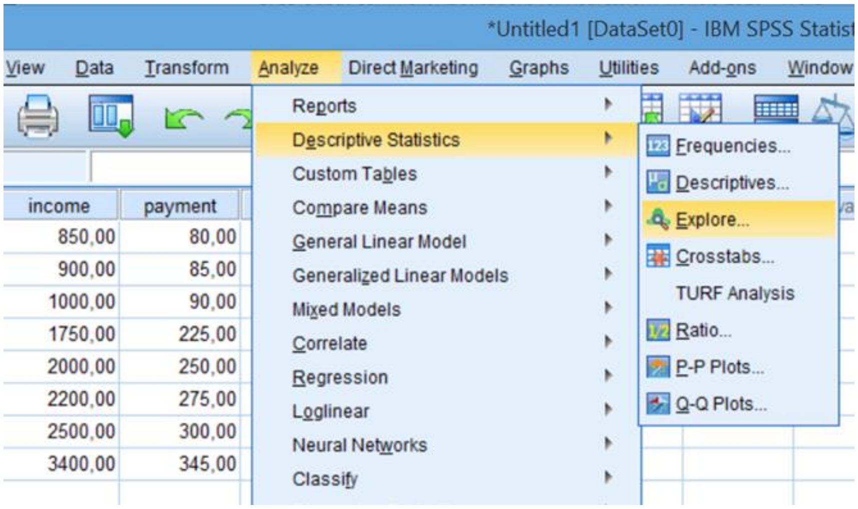
Explore: Plots

Boxplots: ○ Factor levels together ● Dependents together ○ None

Descriptive: □ Stem-and-leaf □ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test: ● None ○ Power estimation ○ Transformed Power: Natural log ○ Untransformed



*Untitled1 [DataSet0] - IBM SPSS Statist

income	payment
850,00	80,00
900,00	85,00
1000,00	90,00
1750,00	225,00
2000,00	250,00
2200,00	275,00
2500,00	300,00
3400,00	345,00

Normality Test Results:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
payment	.238	8	.200*	.868	8	.144

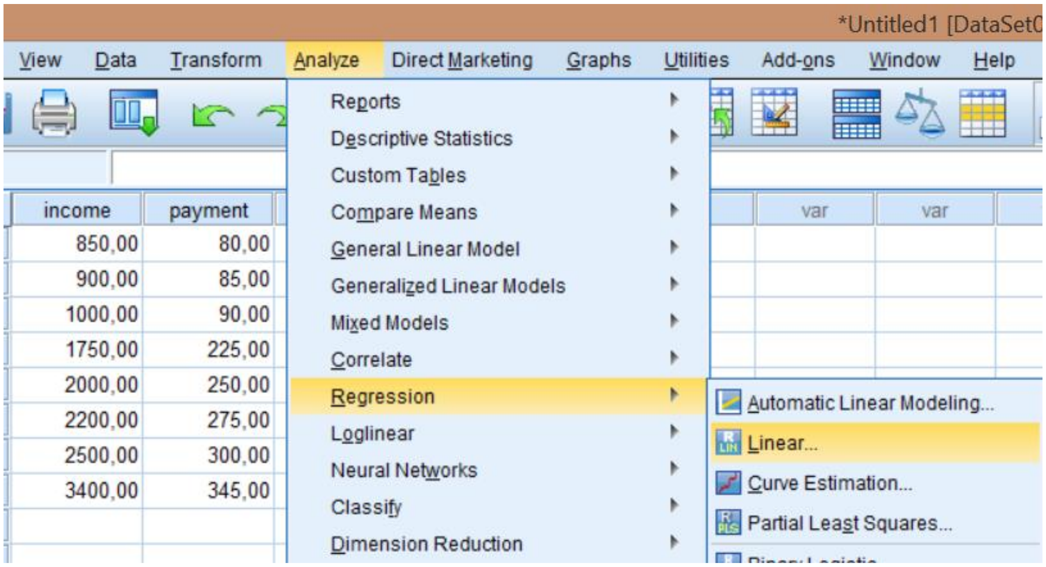
a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

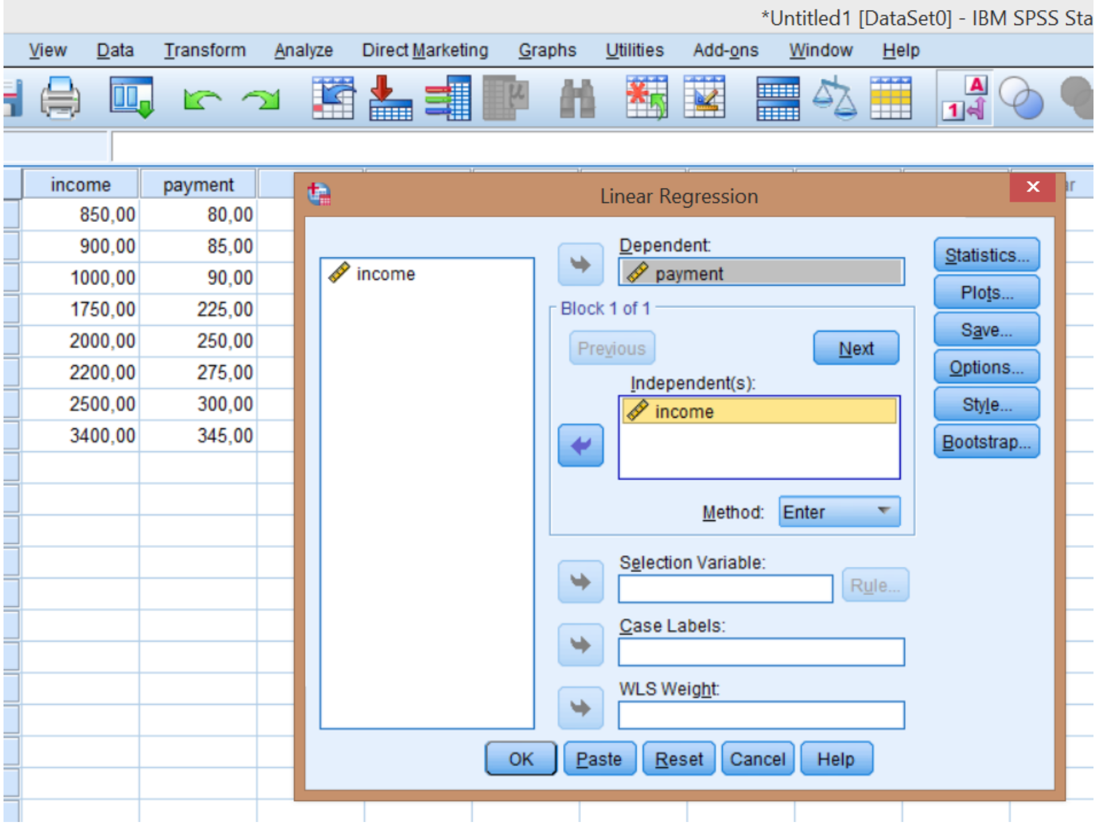
From **Test of Normality Table**

H_0: Data follow a normal distribution.
H_1: Data do not follow a normal distribution

For both Kolmogorov-Smirnov normality test (p-value=0.200) and Shapiro-Wilk normality test (p-value=0.144) p-values are greater than 0.05 so that H_0 cannot be rejected. Data follow a normal distribution. Hence assumption of normality is satisfied, we can continue the regression analysis.



income	payment
850,00	80,00
900,00	85,00
1000,00	90,00
1750,00	225,00
2000,00	250,00
2200,00	275,00
2500,00	300,00
3400,00	345,00



income	payment
850,00	80,00
900,00	85,00
1000,00	90,00
1750,00	225,00
2000,00	250,00
2200,00	275,00
2500,00	300,00
3400,00	345,00

Dependent: payment

Independent(s): income

Method: Enter

Linear Regression Analysis Results:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.969[a]	.938	.928	28.56957

a. Predictors: (Constant), income

b. Dependent Variable: payment

From **Model Summary Table** it is clear that for the regression model, the sample correlation coefficient is $R = 0.969$. Since, the coefficient of determination R^2 is just the square of the correlation coefficient between y and x, for the income-payment regression model's $R^2 = 0.938$, that is, the model accounts for 93.8 % of the variability in the data. 93.8 % of variability in the payments of the people for the cultural activities can be explained by the incomes of people. There is a strong and positive relationship between the payments of the people for the cultural activities and their incomes.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	74290.179	1	74290.179	91.017	.000[a]
	Residual	4897.321	6	816.220		
	Total	79187.500	7			

a. Predictors: (Constant), income

b. Dependent Variable: payment

From **ANOVA Table (Analysis of Variance for Testing Significance of Regression for income and payment Data).**

We will use the analysis of variance approach to test the significance of regression model between income and payment. From ANOVA Table it is clear that ***total sum of squares of the dependent variable*** $\mathrm{SS_T} = 79187.500$, ***the regression sum of squares*** is $\mathrm{SS_R} = 74920.179$, ***the error sum of squares*** is $\mathrm{SS_E} = \mathrm{SS_T} - \mathrm{SS_R} = 79187.500 - 74290.179 = 4897.321$ and ***the estimate of*** σ_ε^2 ***for the income and payment data*** $\hat{\sigma}_\varepsilon^2 = \dfrac{\mathrm{SS_E}}{\mathrm{n}-2} = \dfrac{4897.321}{6} = 816.220$.

1) $\mathrm{H_0}$: The linear regression model between income and payment is not statistically significant. $(\beta_1 = 0)$

$\mathrm{H_1}$: The linear regression model between income and payment is statistically significant. $(\beta_1 \neq 0)$

F test value in the ANOVA Table is used for the **hypotheses test given above.**

$$f = \frac{MS_R}{MS_E} = \frac{74290.179}{816.220} = 91.017$$

The test value is compared by the F table value $f_{0.05,1,6} = 5.99$. Since $f > f_{0.05,1,6}$, the H_0 hypothesis is rejected. We can test this hypothesis also by using p-value (Sig.), since p-value=0.000<0.05, $H_0 : \beta_1 = 0$ is rejected and so the variability in payments of the people (spending) for cultural activities can be explained by the variability in the incomes of the people.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-3.951	24.238		-.163	.876	-63.259	55.357
	income	.115	.012	.969	9.540	.000	.086	.145

a. Dependent Variable: payment

2) We will test the significance of parameters in the fitted regression model to the data. The hypotheses are:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

and we will use $\alpha = 0.05$.

For these hypotheses, t test statistic is used:

$$t = \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} = \frac{0.115}{0.012} \cong 9.540.$$

The test value is compared with the Student's t distribution value $t_{0.025,6} = 2.447$, since the value of the test statistic falls into the critical region, clearly, then the hypothesis that the intercept is zero is rejected. This hypothesis can also be tested by using p-value (Sig.), since p-value=0.000<0.05, $H_0 : \beta_1 = 0$ is rejected. The linear regression model between income and payment is statistically significant.

For testing the hypothesis $\begin{matrix} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{matrix}$, we will use $\alpha = 0.05$.

For these hypotheses, t test statistic is used: $t = \frac{\hat{\beta}_0}{se\left(\hat{\beta}_0\right)} = \frac{-3.951}{24.238} \cong -0.163$.

Taking absolute of the test statistics value $|t| = 0.163$ is compared with the Student's t distribution value $\mathrm{t}_{0.025,6} = 2.447$, since $|t| < t_{0.025,6}$ then the hypothesis with the intercept cannot be rejected. This hypothesis can also be tested by using p-value (Sig.), since p-value=0.876>0.05, $\mathrm{H}_0 : \beta_0 = 0$ cannot be rejected. *It means that we do not need the β_0 term in the fitted model and the fitted linear regression line will pass through the origin.*

** **Note that the analysis of variance procedure for testing for significance of regression is equivalent to the t-test. That is, either procedure will lead to the same conclusions.**

3) From **Coefficients Table,** ***the fitted simple linear regression model is***

$$\hat{y}_i = -3.951 + 0.115 x_i, \quad i = 1, 2, \ldots, 8$$

A unit increment in family income results in the 0.115 TL of increment in their payment for cultural activities.

If family's income is $x = 3000$ TL, their expected/estimated payment for cultural activities would be 341.049TL. Here $x = 3000$ TL is the range of the data.

4) From **Coefficients Table**, we will find a 95% confidence interval on the slope of the regression line for income - payment data;

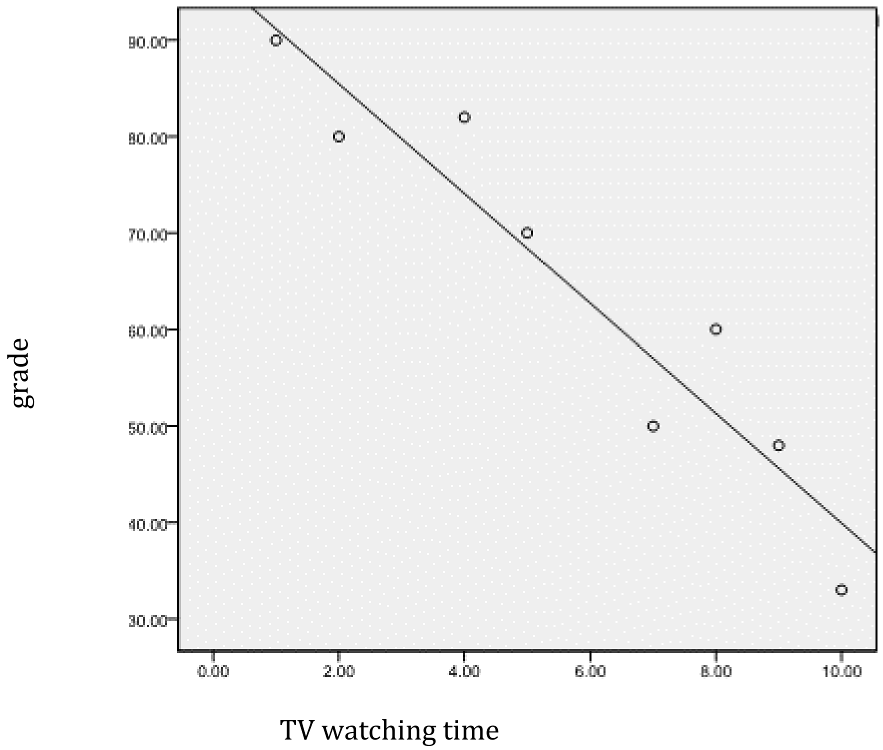
Since β_1 is statistically significant in the model, 95% confidence interval of actual value of the β_1 is (0.086, 0.145).

$$P\left(0.086 < \beta_1 < 0.145\right) = 0.95.$$

Example 2: Statistics course grades of randomly selected 8 students and their TV watching times were given in table below:

TV watching time (hourly) x	Statistics course grade y
1	90
2	80
4	82
5	70
8	60
7	50
9	48
10	33

Scatter plot of the data:



From scatter graph it is clear that there is a negative linear relationship between Statistics course grade and TV watching time variables.

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
grade	,165	8	,200(*)	,958	8	,793

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

From **Test of Normality Table**
H_0: Data follow a normal distribution.
H_1: Data do not follow a normal distribution

For both Kolmogorov-Smirnov normality test (p-value=0.200) and Shapiro-Wilk normality test (p-value=0.793) p-values are greater than 0.05 so that H_0 cannot be rejected. Data follow a normal distribution. Hence assumption of normality is satisfied, we can continue the regression analysis.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.948[a]	.899	.882	6.76653

a. Predictors: (Constant), time

b. Dependent Variable: grade

From **Model Summary Table** it is clear that for the regression model, the sample correlation coefficient is $R = 0.948$. Since, the coefficient of determination R^2 is just the square of the correlation coefficient between y and x, for the grade-TV watching time regression model's $R^2 = 0.899$, that is, the model accounts for 89.9 % of the variability in the data. 89.9 % of variability in the students' grades can be explained by their watching time on TV.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2446.160	1	2446.160	53.426	.000[a]
	Residual	274.715	6	45.786		
	Total	2720.875	7			

a. Predictors: (Constant), time

b. Dependent Variable: grade

From **ANOVA Table (Analysis of Variance for Testing Significance of Regression for grade and TV watching time Data).**

We will use the analysis of variance approach to test the significance of regression model between grade and TV watching time. From ANOVA Table it is clear that ***total sum of squares of the dependent variable*** $SS_T = 2720.875$, ***the regression sum of squares*** is $SS_R = 2446.160$, ***the error sum of squares*** is $SS_E = SS_T - SS_R = 2720.875 - 2446.160 = 274.715$ and ***the estimate of*** σ_ε^2 ***for the income and payment data*** $\hat{\sigma}_\varepsilon^2 = \dfrac{SS_E}{n-2} = \dfrac{274.715}{6} = 45.786$.

1) H_0 : The linear regression model between TV watching time and course grade is not statistically significant.

H_1 : The linear regression model between TV watching time and course grade is statistically significant.

F test value in the ANOVA Table is used for the **hypotheses test given above.**

$$f = \frac{MS_R}{MS_E} = \frac{2446.160}{45.786} = 53.426$$

The test value is compared by the F table value $f_{0.05,1,6} = 5.99$. Since $f > f_{0.05,1,6}$, the H_0 hypothesis is rejected. We can test this hypothesis also by using p-value (Sig.), since p-value=0.000<0.05, $H_0 : \beta_1 = 0$ is rejected and so the variability in students' grades (success) can be explained by the variability in the their watching time on TV. The linear regression model between TV watching time and course grade is statistically significant at the significance level of 0.05.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	96.854	5.077		19.078	.000	84.432	109.277
	time	-5.692	.779	-.948	-7.309	.000	-7.598	-3.787

a. Dependent Variable: grade

2) We will test the significance of parameters in the fitted regression model to the data. The hypotheses are:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

and we will use $\alpha = 0.05$.

For these hypotheses, t test statistic is used:

$$t = \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} = \frac{-5.692}{0.779} \cong -7.309.$$

Taking absolute value of the test value $|t| = 7.309$ is compared with the Student's t distribution value $t_{0.025,6} = 2.447$, since the value of the test statistic falls into the critical region, clearly, then the hypothesis that the slope is zero is rejected. This hypothesis can also be tested by using p-value (Sig.), since p-value=0.000<0.05, $H_0 : \beta_1 = 0$ is rejected.

The hypotheses are $\begin{matrix} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{matrix}$ and we will use $\alpha = 0.05$.

For these hypotheses, t test statistic is used: $t = \dfrac{\hat{\beta}_0}{se(\hat{\beta}_0)} = \dfrac{96.854}{5.077} \cong 19.078$.

The test statistics value $t = 19.078$ is compared with the Student's t distribution value $t_{0.025,6} = 2.447$, since $t \geq t_{0.025,6}$ then the hypothesis that the intercept is zero is rejected. This hypothesis can also be tested by using p-value (Sig.), since p-value=0.000<0.05, $H_0 : \beta_0 = 0$ is rejected. It means that the β_0 term in the fitted model is statistically significant.

** ***Note that the analysis of variance procedure for testing for significance of regression is equivalent to the t-test. That is, either procedure will lead to the same conclusions.***

3) From **Coefficients Table,** ***the fitted simple linear regression model is***

$\hat{y}_i = 96.854 - 5.692 x_i, \quad i = 1, 2, \ldots, 8$

A unit increment (an hour increment) in the students' spending time on TV results in the 5.692 point of decreasing in the students' grade.

Since β_0 term in the fitted model is statistically significant, also can be interpreted. $\hat{y}$ =96.854 when the students' spending time on TV is x=0. The mean student grade is 96. 854 when no time is spent on watching TV.

If the students are spending 3 hours on TV ($x = 3$ hours), their expected/estimated grade would be 79.778 points. $\hat{y}_i = 96.854 - 5.692 \times (3) = 79.778$. Here $x = 3$ hours is in the range of the data.

For $x_3 = 4$ hours, their expected/estimated grade from the fitted model would be 74.086 points. The difference between the observed grade and the expected/estimated grade is called residual. Here $e_3 = y_3 - \hat{y}_3 = 82 - 74.086 = 7.914$

4) From **Coefficients Table**, we will find a 95% confidence interval on the slope of the regression line for grade and TV watching time:

Since both β_0 and β_1 are statistically significant in the model, 95% confidence intervals of actual values of the β_0 and β_1, respectively:

$$P\left(84.432 < \beta_0 < 109.277\right) = 0.95.$$

$$P\left(-7.589 < \beta_1 < -3.787\right) = 0.95.$$