İST292 STATISTICS LESSON 8

REGRESSION ANALYSIS IN SPSS

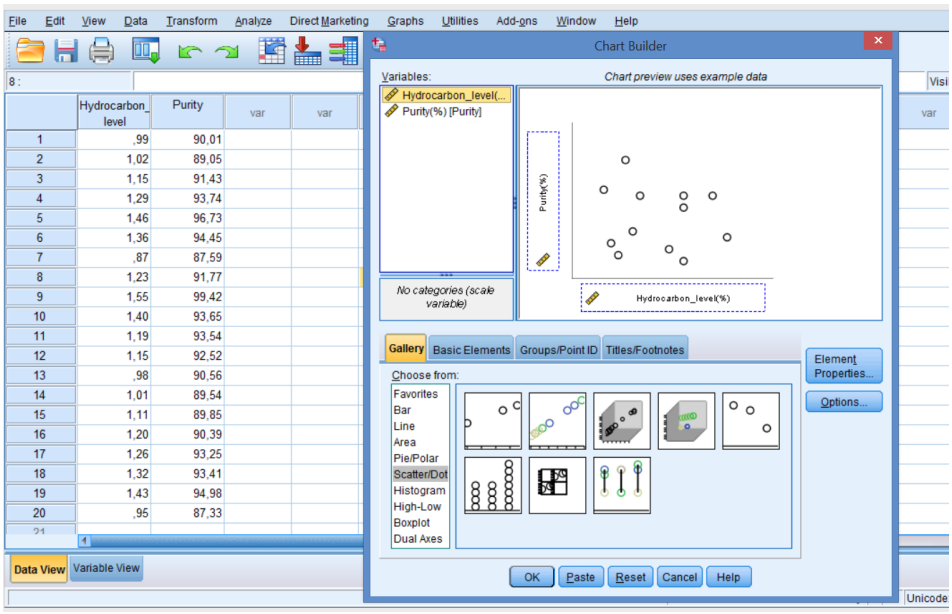
This is an example of SPSS analysis of *oxgyen purity data.*

Table 1. Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level	Purity y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

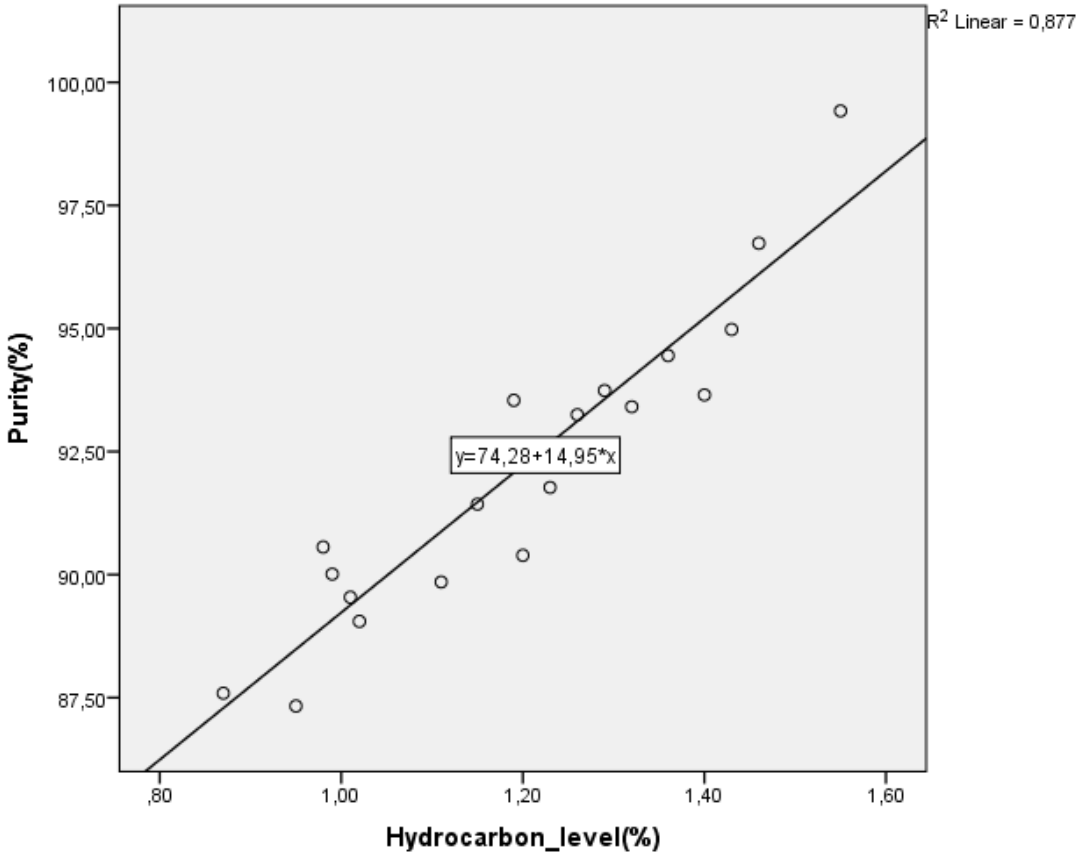
Scatter plot and normality test of the data.

Graphs → Chart Builder then choose Scatter/Dot

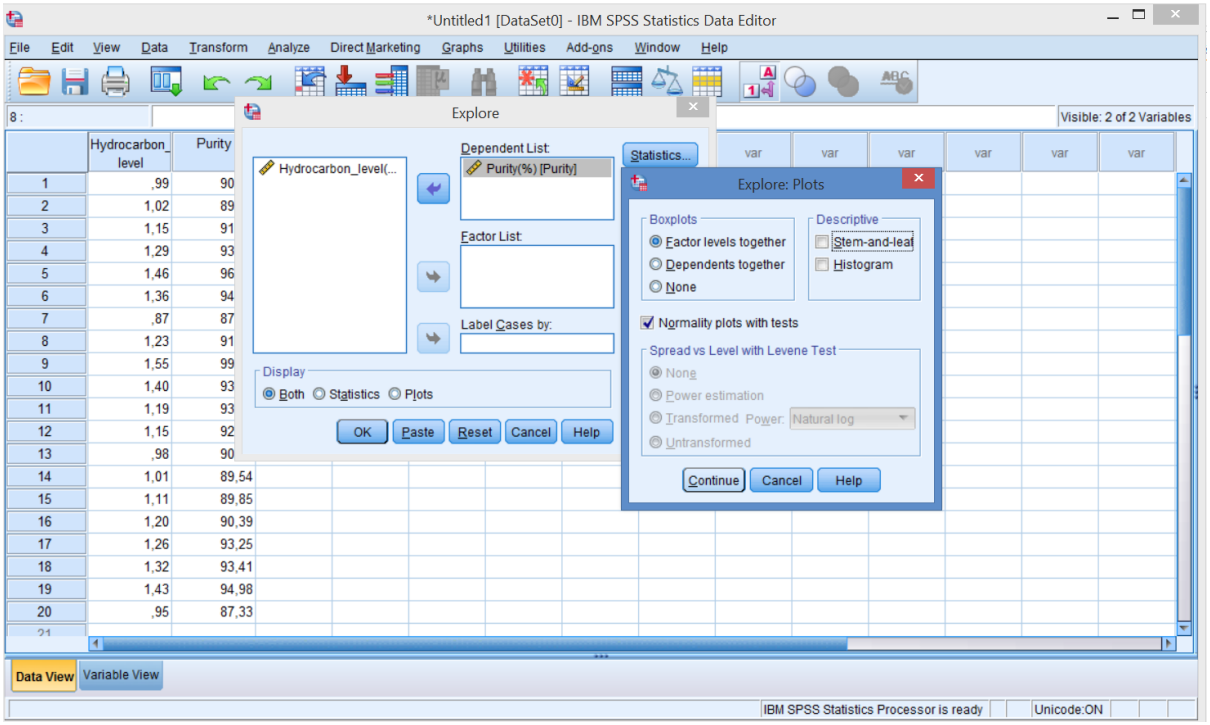


	Hydrocarbon level	Purity
1	,99	90,01
2	1,02	89,05
3	1,15	91,43
4	1,29	93,74
5	1,46	96,73
6	1,36	94,45
7	,87	87,59
8	1,23	91,77
9	1,55	99,42
10	1,40	93,65
11	1,19	93,54
12	1,15	92,52
13	,98	90,56
14	1,01	89,54
15	1,11	89,85
16	1,20	90,39
17	1,26	93,25
18	1,32	93,41
19	1,43	94,98
20	,95	87,33

Scatter plot of the data:



Analyze→ Explore then send Purity variable under Dependent List. Click Statistics then choose Normality plots with tests.



Normality Test Results:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Purity(%)	,102	20	,200*	,968	20	,713

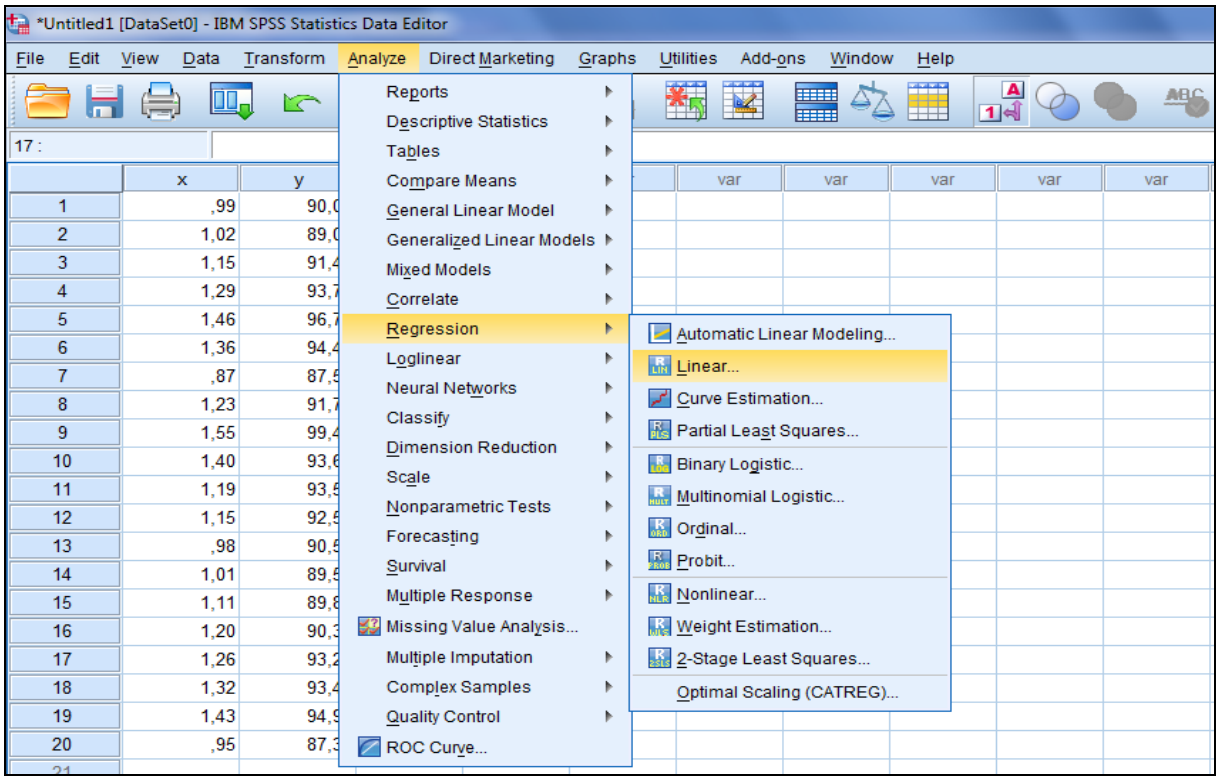
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

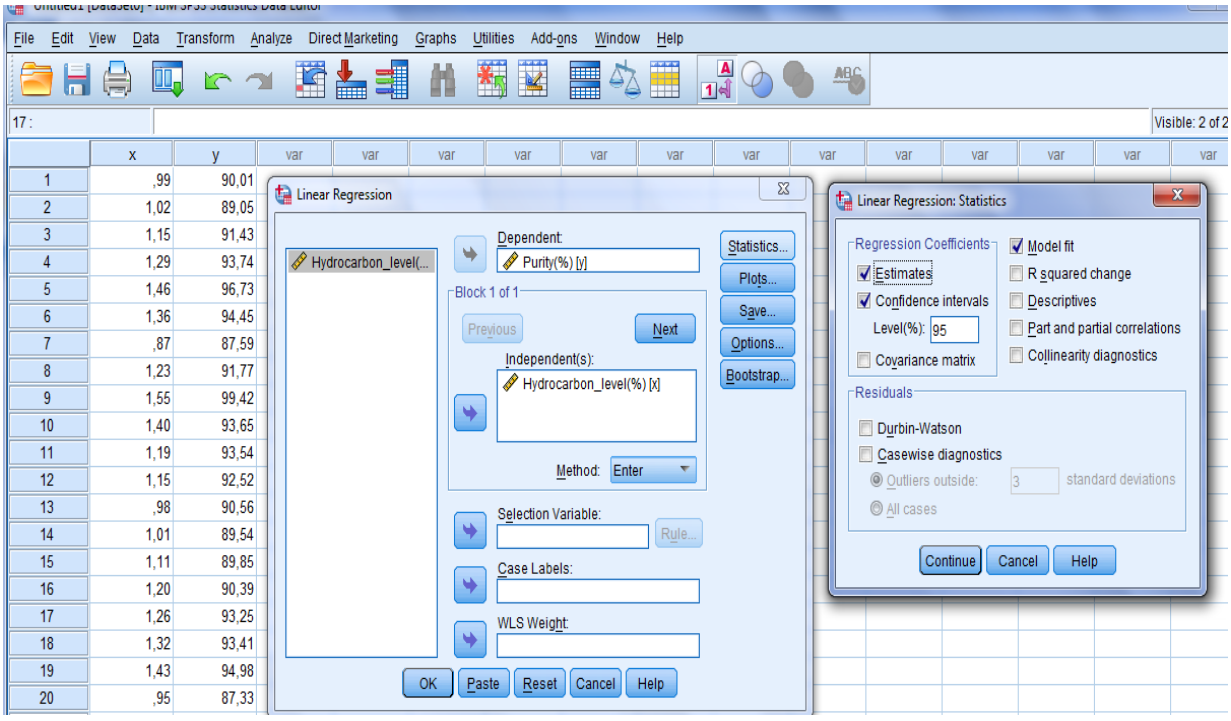
From **Test of Normality Table**
H_0: Data follow a normal distribution.
H_1: Data do not follow a normal distribution

For both Kolmogorov-Smirnov normality test (p-value=0.200) and Shapiro-Wilk normality test (p-value=0.144) 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.

Analyze→ Regression →Linear



On Linear Regression Window Purity (%)-y variable is under Dependent, Hydrocarbon-level (%)- x variable under Independent (s). Click Statistics on this window and then on Statistics window click Estimates, Confidence Intervals, Model fit. Then Click Continue and Ok.



OUTPUTS

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Hydrocarbon_level(%)[b]	.	Enter

a. Dependent Variable: Purity(%)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,937[a]	,877	,871	1,08653

a. Predictors: (Constant), Hydrocarbon_level(%)

From **Model Summary Table** it is clear that for the oxygen purity regression model, the sample correlation coefficient is $R = 0.937$. There is a strong relationship between hydrocarbon level and oxygen purity (%). Since, the coefficient of determination R^2 is just the square of the correlation coefficient between y and x, for the oxygen purity regression model's $R^2 = 0.877$, that is, the model accounts for 87.7 % of the variability in the data. Hydrocarbon level (x) explains 87.7% variation in oxygen purity (y).

Note: R^2 indicates the proportion of the variance in the dependent variable (y) that is predicted or explained by linear regression and the independent variable (x).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	152,127	1	152,127	128,862	,000[b]
	Residual	21,250	18	1,181		
	Total	173,377	19			

a. Dependent Variable: Purity(%)

b. Predictors: (Constant), Hydrocarbon_level(%)

From **ANOVA Table (Analysis of Variance for Testing Significance of Regression for Oxygen Purity Data),**

We will use the analysis of variance approach to test for significance of regression using the oxygen purity data model. From ANOVA Table it is clear that ***total sum of squares of the dependent variable*** $SS_T = 173.377$, ***the regression sum of squares*** is $SS_R = 152.127$, ***the error sum of squares*** is $SS_E = SS_T - SS_R = 173.377 - 152.127 = 21.250$ and ***the estimate of*** σ_ε^2 ***for the oxygen purity data*** $\hat{\sigma}_\varepsilon^2 = \dfrac{SS_E}{n-2} = \dfrac{21.250}{18} = 1.181$.

1) We will test for significance of regression using the model for the oxygen purity data. The hypotheses are

H_0 : Model is not significant $(\beta_1 = 0)$
H_1 : Model is significant $(\beta_1 \neq 0)$

and we will use $\alpha = 0.05$.

We would reject H_0 if $f > f_{0.05,1,18}$. The quantities $MS_R = SS_R/1$ and $MS_E = SS_E/18$ are called **mean squares**. In general, a mean square is always computed by dividing a sum of squares by its number of degrees of freedom. The test statistic is $f = MS_R / MS_E = 152.127/1.181 \cong 128.862$ and since $f = 128.862 > f_{0.05,1,18} = 4.41$ we reject H_0, so we conclude that the simple regression model between the Hydrocarbon level(%) and Purity(%) is significant. (Remember $MS_E = \hat{\sigma}_\varepsilon^2$). We can test this hypothesis also by using p-value (Sig.), since p-value=0.000<0.05, H_0 : Model is not significant is rejected. Comment on H_0 : Model is not significant *is rejected, this implies that x is of value in explaining the variability in y. Rejecting* H_0 : Model is not significant *could mean either that the straight-line model is adequate. The regression model is 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.***

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	74,283	1,593		46,617	,000	70,936	77,631
	Hydrocarbon_level(%)	14,947	1,317	,937	11,352	,000	12,181	17,714

a. Dependent Variable: Purity(%)

From **Coefficients Table** we know that $\hat{\beta}_1 = 14.97$ and for testing the hypothesis $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ the t-statistic is $t = \dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)} \cong \dfrac{14.947}{1.317} = 11.352$. Since the reference value of t is $t_{0.025,18} = 2.101$, the value of the test statistic is very far into the critical region, implying that $H_0 : \beta_1 = 0$ should be 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. *The regression model is significant at the significance level of 0.05.*

2) From **Coefficients Table** the t-statistic for testing the hypothesis $\begin{matrix} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{matrix}$ is $t = \dfrac{\hat{\beta}_0}{se\left(\hat{\beta}_0\right)} = \dfrac{74.283}{1.593} \cong 46.617$. Since the reference value of t is $t_{0.025,18} = 2.101$, the value of the test statistic is very far 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, $H_0 : \beta_0 = 0$ is rejected. β_0 is statistically *significant at the significance level of 0.05.*

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

$$\hat{y} = 74.283 + 14.947\,x$$

A unit increment (1% increment) in hydrocarbon level results in the 14.997% of increment in oxygen purity.

Using the regression model, we would predict oxygen purity of $\hat{y} = 104.177\%$ when the hydrocarbon level is $x = 2.00\%$ ($\hat{y} = 74.283 + 14.947.(2) = 104.177$)The purity 104.177 % may be interpreted as an estimate of the true population mean purity when hydrocarbon level is 2.00%, or as an estimate of a new observation when x=2.00 %. These estimates are, of course, subject to error; that is, it is unlikely that a future observation on purity would be exactly 104.177 % when the hydrocarbon level is 2.00 %.

4) From **Coefficients Table**, we will find a 95% confidence interval on the slope of the regression line using the oxygen purity data;

$$P\left(12.181 \leq \beta_1 \leq 17.714\right) = 0.95$$

Comment: This interval contains the actual value of β_1 with 95% confidence.

Similarly, 95 % confidence interval on the intercept β_0 is

$$P\left(70.936 \leq \beta_0 \leq 77.631\right) = 0.95$$

Comment: This interval contains the actual value of β_0 with 95% confidence.