SPSS – PART II

SHORTCOURSE HANDOUT

TEXAS TECH UNIVERSITY
Information Technology Division
Technology Support

Texas Tech University | Carlee M. DeYoung, M.A.

Table of Contents

IBM SPSS Statistics 28 – Part II ShortCourse Handout

Credit: This document is adapted from SPSS (Getting Help, Tutorial, Statistics Coach, and Case Studies) for SPSS 28 and ,with permission, the SPSS 25 Part-II ShortCourse Handout created by Heide Mansouri.

Introduction

IBM SPSS Statistics 28 is a suite of statistical software available for **Windows**, **Macintosh**, and the **UNIX** platforms. IBM SPSS Statistics 28 stores and outputs data files with **.spv** file extensions.

IBM SPSS Statistics 28 provides statistical analysis and data management system in an accessible **point-and-click interface**.

This course will cover **more** common statistical analyses such as **T-tests, ANOVA,** and **Linear Regression**. In this Short Course it is assumed that you are familiar with basic statistics, and you have taken SPSS Statistics – Part I ShortCourse.

Course Objectives

- Perform a One-Sample T-Test;
- Perform a Paired Samples T-test;
- Perform an independent Samples T-Test;
- Perform a One-way Analysis of Variance;
- Perform a Linear Regression.

Accessing IBM SPSS

A desktop download of IBM SPSS Statistics software is offered free to TTU students and at a cost to TTU faculty/ staff on the TTU software license site.

T-Tests

There are three types of **T-Tests**:

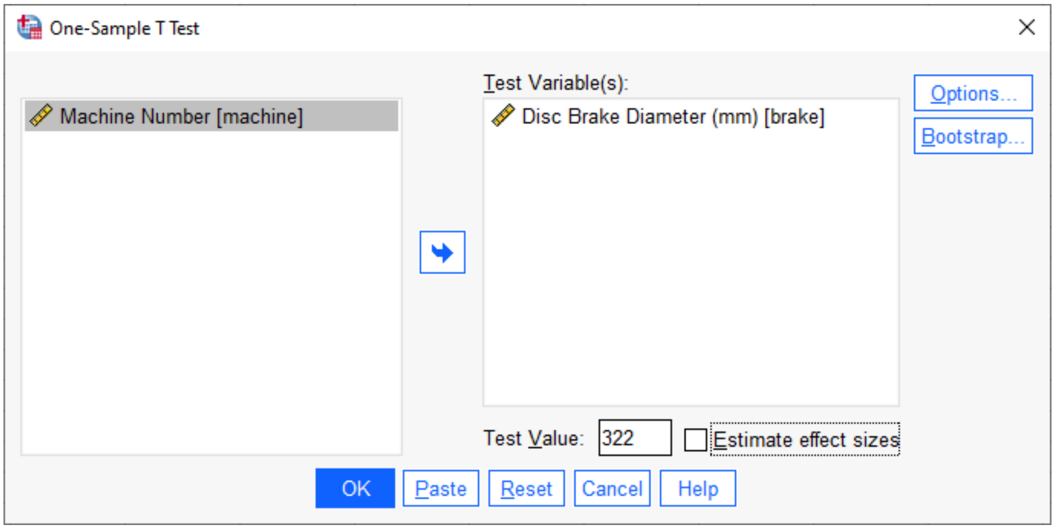
- **One-Sample T-Test**: Compares the mean of one variable with a known or hypothesized value.
- **Paired-samples T-Test:** Compares the means of two variables for a single group. This test is also for matched pairs.
- **Independent-samples T-test**: (two samples t-test) Compares the means of one variable for two groups of cases
- T-tests are used to explore issues such as:
  - Does treatment A yield a higher rate of recovery than treatment B?
  - T-tests always compare two different means or values.

Exercise #1 One Sample T-test

A Manufacturer of a high-performance automobiles produces disc brakes that must measure 322 millimeters in diameter. Quality control randomly draws 16 discs made by each of eight production machine and measures their diameters. Use the **brakes.sav** sample file, and One Sample T-test to determine whether the mean diameters of the brakes in each sample significantly differ from 322 millimeters. In this data set a nominal variable **Machine Number**, identifies the production machine used to make the disc brake. The data from each machine must be tested as a separate sample, so we need to first **split** the data file into groups by **Machine Number.**

Note: Split command, splits the active file into subgroups that can be analyzed separately. All split variables (up to 8 variables) must be numeric.

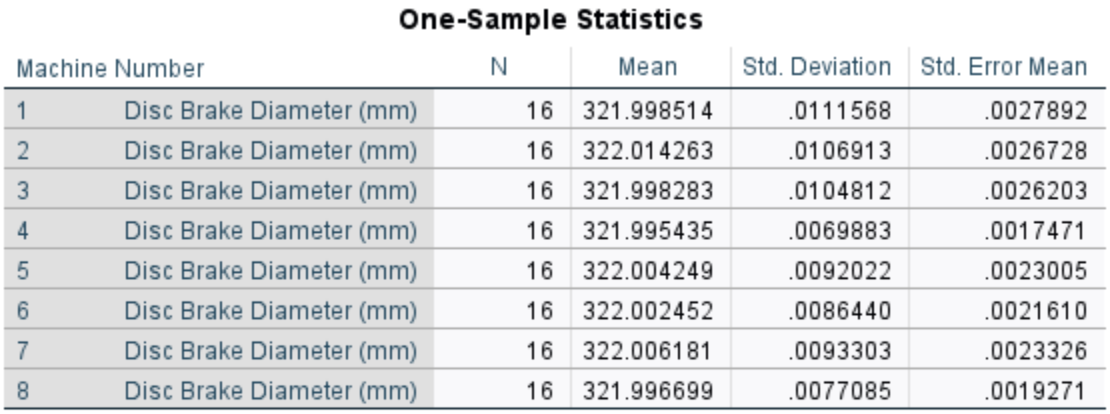
- Open the **brakes.sav** sample file
  - C:\ Program Files\IBM\SPSS\Statistics\28\Sample\English\brakes.sav
- **To split the file**, from menus choose Data -> Split File…
- Select **Compare groups,** radio button.
- Move the Machine Number variable to **Groups Based on:** box.
- Click OK.
- **To begin the One Sample T-test**, from the menus choose **Analyze** -> **Compare means** -> **One Sample T-test…**
- Select Disc Brake Diameter (mm) as the test variable(s).
- Type 322 as the test value, and then click **Options…**
- Type 90 as the confidence interval percentage and then click **Continue**.
- Unselect the “estimate effect sizes” box.
- Click OK.



Machine Number [machine]

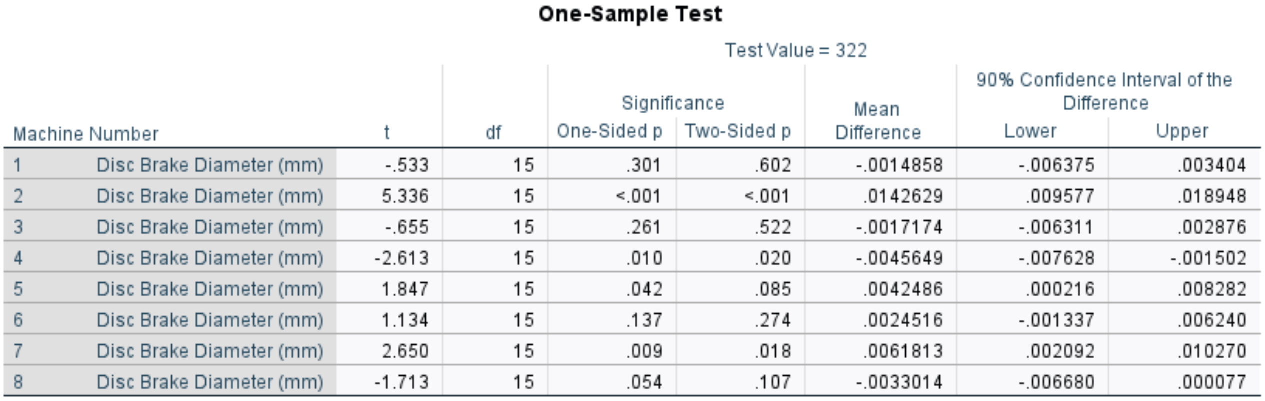
Test Variable(s):
Disc Brake Diameter (mm) [brake]

Test Value: 322 □ Estimate effect sizes



One-Sample Statistics

Machine Number		N	Mean	Std. Deviation	Std. Error Mean
1	Disc Brake Diameter (mm)	16	321.998514	.0111568	.0027892
2	Disc Brake Diameter (mm)	16	322.014263	.0106913	.0026728
3	Disc Brake Diameter (mm)	16	321.998283	.0104812	.0026203
4	Disc Brake Diameter (mm)	16	321.995435	.0069883	.0017471
5	Disc Brake Diameter (mm)	16	322.004249	.0092022	.0023005
6	Disc Brake Diameter (mm)	16	322.002452	.0086440	.0021610
7	Disc Brake Diameter (mm)	16	322.006181	.0093303	.0023326
8	Disc Brake Diameter (mm)	16	321.996699	.0077085	.0019271



One-Sample Test

		Test Value = 322						
				Significance		Mean Difference	90% Confidence Interval of the Difference	
Machine Number		t	df	One-Sided p	Two-Sided p		Lower	Upper
1	Disc Brake Diameter (mm)	-.533	15	.301	.602	-.0014858	-.006375	.003404
2	Disc Brake Diameter (mm)	5.336	15	<.001	<.001	.0142629	.009577	.018948
3	Disc Brake Diameter (mm)	-.655	15	.261	.522	-.0017174	-.006311	.002876
4	Disc Brake Diameter (mm)	-2.613	15	.010	.020	-.0045649	-.007628	-.001502
5	Disc Brake Diameter (mm)	1.847	15	.042	.085	.0042486	.000216	.008282
6	Disc Brake Diameter (mm)	1.134	15	.137	.274	.0024516	-.001337	.006240
7	Disc Brake Diameter (mm)	2.650	15	.009	.018	.0061813	.002092	.010270
8	Disc Brake Diameter (mm)	-1.713	15	.054	.107	-.0033014	-.006680	.000077

Results

- The **One Sample Statistics** table displays descriptive statistics for each of the eight machines (samples).
- The **t column** on **One sample Test** table displays the observed statistics for each sample, calculated as the ratio of the mean difference divided by the standard error of the sample mean.
- The **df column** displays degrees of freedom (number of cases in each group minus 1).
- The column labeled **Sig. (2-tailed)** displays a probability from the t distribution with 15 degrees of freedom. The value listed is the 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.
- The **Mean Difference** is obtained by subtracting the test value (322) form each sample mean.
- The 90% **Confidence Interval of the Difference** provides an estimate of the boundaries between which the true mean difference lies in 90% of all possible random samples of 16 disc brakes produced by this machine.
- Since their confidence intervals lie entirely above 0.0, you can safely say that **machines 2, 5, and 7** are producing discs that are significantly wider than 322mm on the average.
- Similarly, because its confidence interval lies entirely below 0.0, **machine 4** is producing discs that are not wide enough.

- The one-sample t-test can be used whenever sample means must be compared to a known test value. As with all tests, the one sample t-test assumes that the data is reasonably, normally distributed, especially with respect to skewness.
- Extreme or outlying values should be carefully checked using **boxplots** for example.

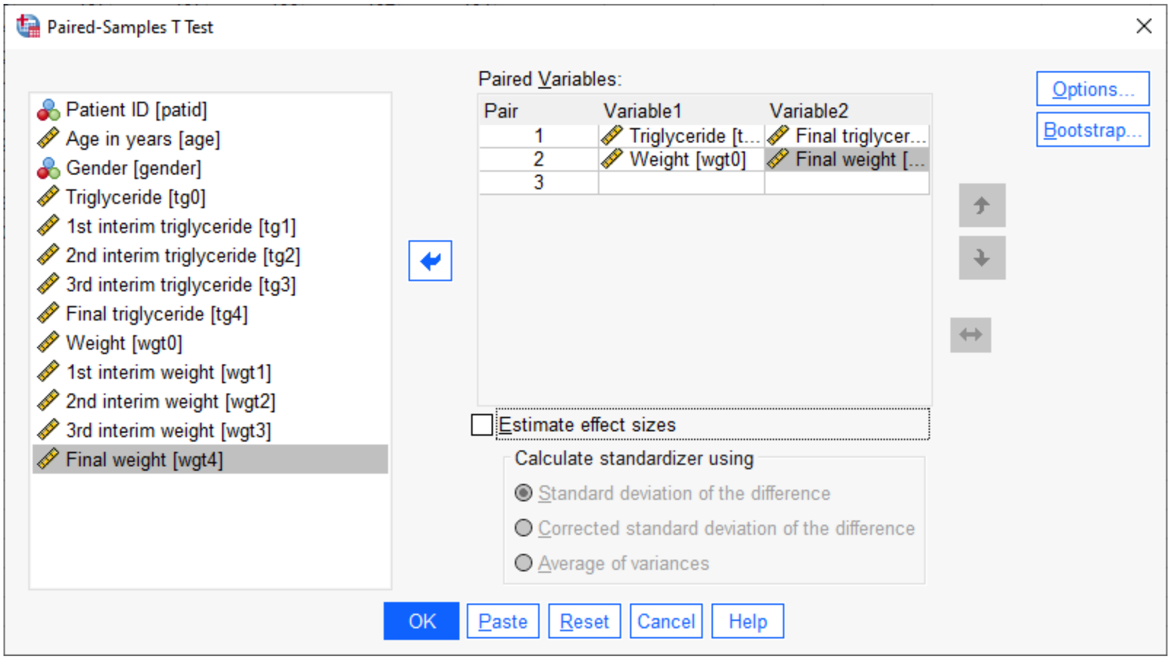
Paired-Sample T-test (Pre - Post Design)

The Paired-Samples T-test procedure is used to test the hypothesis of no difference between two variables. The data may consist of two measurements taken on the same subject or one measurement taken on a matched pair of subjects.

Exercise #2 Paired-Sample T-test (Pre – Post Design)

A physician is evaluating a new diet for her patients with a family history of heart disease. To test the effectiveness of this diet, 16 patients are placed on the diet for 6 months. Their weights and triglyceride levels are measured before and after the study, and physicians want to know if either set of measurements has changed significantly. Use the dietstudy.sav sample file and Paired-samples t-test to determine whether there is a significant difference between the pre- and post-diet weights and triglyceride levels of these patients.

- Open the **dietstudy.sav** sample file
    - C:\ Program Files\IBM\SPSS\Statistics\28\Sample\English\dietstudy.sav
- From the menus choose **Analyze** -> **Compare Means** -> **Paired-Samples T-test…**
- Select **Triglyceride** (TG 0) and **Final Triglyceride** (TG 4) as the first set of paired variables.
- Select **Weight** (wgt 0) and **Final Weight** (wgt 4) as the second pair.
- Click OK.

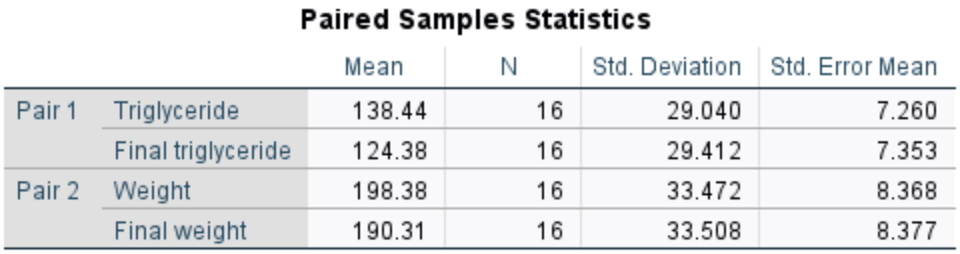

Paired-Samples T Test
Patient ID [patid]
Age in years [age]
Gender [gender]
Triglyceride [tg0]
1st interim triglyceride [tg1]
2nd interim triglyceride [tg2]
3rd interim triglyceride [tg3]
Final triglyceride [tg4]
Weight [wgt0]
1st interim weight [wgt1]
2nd interim weight [wgt2]
3rd interim weight [wgt3]
Final weight [wgt4]
Options...
Bootstrap...
Calculate standardizer using
OK
Paste
Reset
Cancel
Help

Paired Variables:

Pair	Variable1	Variable2
1	Triglyceride [t...	Final triglycer...
2	Weight [wgt0]	Final weight [...
3		

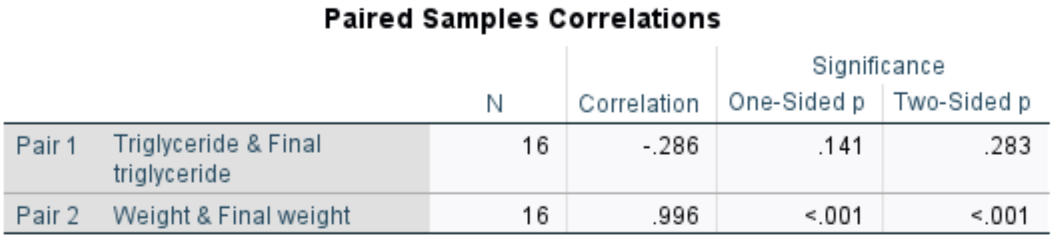
□ Estimate effect sizes

● Standard deviation of the difference
○ Corrected standard deviation of the difference
○ Average of variances



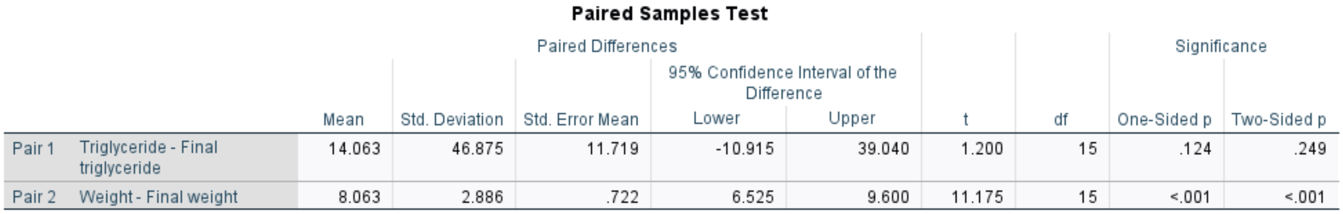
Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Triglyceride	138.44	16	29.040	7.260
	Final triglyceride	124.38	16	29.412	7.353
Pair 2	Weight	198.38	16	33.472	8.368
	Final weight	190.31	16	33.508	8.377



Paired Samples Correlations

				Significance	
		N	Correlation	One-Sided p	Two-Sided p
Pair 1	Triglyceride & Final triglyceride	16	-.286	.141	.283
Pair 2	Weight & Final weight	16	.996	<.001	<.001



Paired Samples Test

		Paired Differences							Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	One-Sided p	Two-Sided p
Pair 1	Triglyceride - Final triglyceride	14.063	46.875	11.719	-10.915	39.040	1.200	15	.124	.249
Pair 2	Weight - Final weight	8.063	2.886	.722	6.525	9.600	11.175	15	<.001	<.001

Results:

- The **Paired Samples Statistics** table displays descriptive statistics for both groups.
- Across all 16 subjects, triglyceride level dropped about 14 points on average after 6 months of the new diet.
- The subjects lost weight over the course of the study; 8lbs on average.
- The standard deviations for pre- and post-diet measurements revel that subjects were more variable with respect to weight than to triglyceride levels.
- At -.0286, the correlation between the baseline and six-moth triglyceride levels is not statistically significant.
- The Pearson correlation between the baseline and six-month weight measurements is 0.996, almost a perfect correlation. Unlike the triglyceride levels, all subjects lost weight.
- The mean column in the **Paired Samples Test** table displays the average difference between triglycerides and weight measurements before the diet and six months into the diet.
- The Std. Deviation column displays the standard deviation of the average difference score.
- The Std. Error Mean column provides an index of the variability one can expect in repeated random samples of 16 patients similar to the one in this study.
- The 95% confidence interval of the difference provides an estimate of the boundaries between which the true mean difference lies in 95% of all possible random samples of 16 patients similar to the ones participating in this study.
- The **t-statistic** is obtained by dividing the mean difference by its standard error.

- The **Sig. (2-tailed)** column displays the probability of obtaining a t statistic whose absolute value is equal to or greater than the obtained t-statistic.
- Since the significance value for change in weight is less than 0.05, you can conclude that the **average loss of 8.063 pounds per patient** is not due to chance variation and can be attributed to the diet.
- However, the significance value greater than 0.10 (Theat is 0.249) for change in triglyceride level shows the diet did not significantly reduce their triglyceride levels.

Independent -Samples T-Test (Two Sample T-Test)

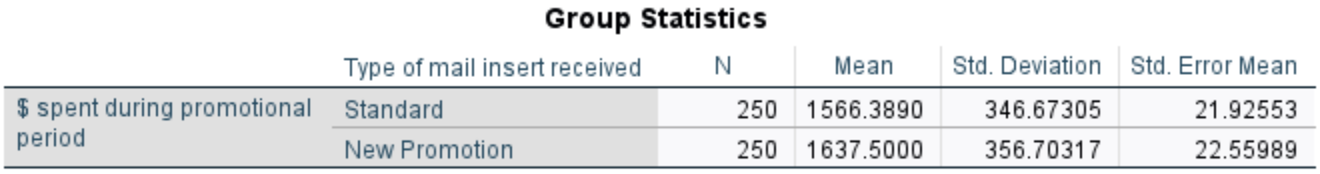
The independent -samples T-test procedure compares means for two groups of cases. Ideally, for this test, the subjects should be randomly assigned to two groups, so that any difference in response is due to treatment (or lack of treatment) and not other factors This is not the case if you compare average income for males and females. A person is not randomly assigned to be a male or female. In such situations, you should ensure that differences in other factors are not masking or enhancing a significant difference in means. Differences in average income may be influenced by factors such as education (and not be gender alone).

Example: Patients with high blood pressure are randomly assigned to a placebo group and treatment group. The placebo subjects receive an inactive pill, and the treatment subjects receive a new drug that is expected to lower blood pressure. After the subjects are treated for two month, the two samples t-test is sued to compare the average blood pressures for the placebo group and the treatment group. Each patient is measured once and belongs to one group only.

Exercise #3 Performing Independent Samples T-Test

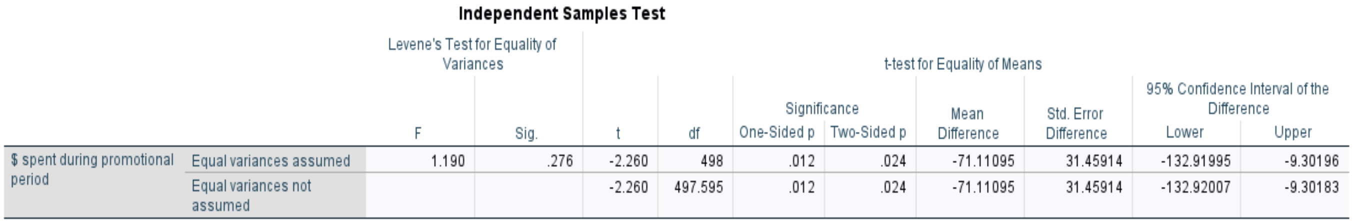
An analyst at a department store wants to evaluate a recent credit card promotion. For this study, 500 cardholders were randomly selected. Half received an ad promoting a reduced interest rate on purchases made over the next three months, and half received a standard seasonal ad. Use the sample file **creditpromo.sav**, to access the data.

- Open the **creditpromo.sav** sample file
    - C:\ Program Files\IBM\SPSS\Statistics\28\Sample\English\creditpromo.sav
- From **Analyze** menu, choose **Compare Means** -> **Independent -Samples T test…**
- Move the **$ spent during promotional period** variable to the Test variable(s) box.
- Move the **Type of mail insert received** variable to the Grouping Variable box
- Click **Define Groups…**
- Type: 0 as the **Group 1** value and 1 as the **Group 2** value and then click **Continue.**
- Click OK.



Group Statistics

	Type of mail insert received	N	Mean	Std. Deviation	Std. Error Mean
$ spent during promotional period	Standard	250	1566.3890	346.67305	21.92553
	New Promotion	250	1637.5000	356.70317	22.55989



Independent Samples Test

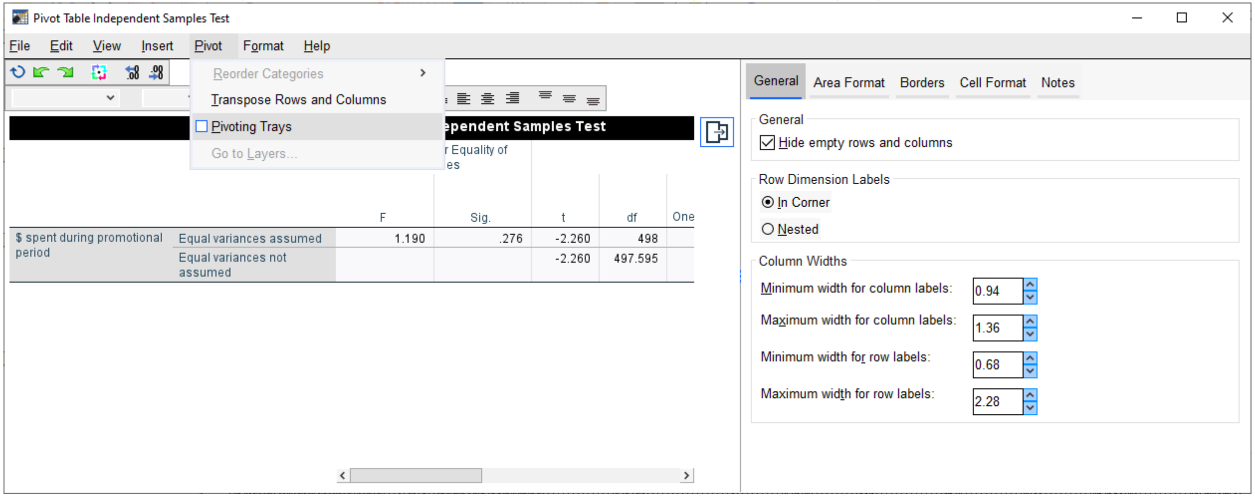
		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance		Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p			Lower	Upper
$ spent during promotional period	Equal variances assumed	1.190	.276	-2.260	498	.012	.024	-71.11095	31.45914	-132.91995	-9.30196
	Equal variances not assumed			-2.260	497.595	.012	.024	-71.11095	31.45914	-132.92007	-9.30183

Results:

- The **Group Statistics** table displays the descriptive statistics for both groups. On average, customers who receive the interest-rate promotion charged bout $71 more than the comparison group, and they vary a little more around their average.
- The procedure produces two tests of the difference between the two groups. One test assumes that the variances of the two groups are equal. The **Levene** statistic tests this assumption.
- In this example, the significance value of the statistic is 0.276 (for F=1.19). Because this value is greater than 0.10, you can assume that the groups have equal variances and ignore the second test (null hypothesis is not rejected).
- Using the pivoting trays, you can change the default layout of the table so that only the "equal variances" test is displayed.

Exercise #4 Using the Pivoting Trays

- On the output in Exercise #3, double click the **Independent Samples Test** table to activate it and use the Pivoting Trays.
- From the **Pivot Table** viewer menus choose **Pivot** -> **Pivoting Trays**
- Move Assumptions from the row to the **Layer**.
- **Close** the pivoting trays window. Test table is pivoting so that assumptions in the layer, the **Assumptions = Equal variances assumed** panel is displayed.



General

☑ Hide empty rows and columns

Row Dimension Labels

● In Corner

○ Nested

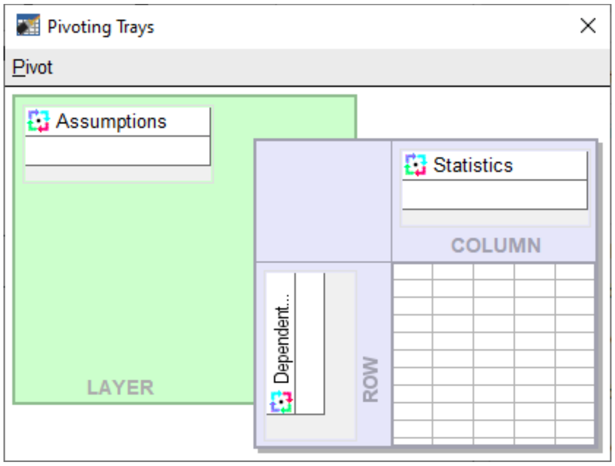
Column Widths

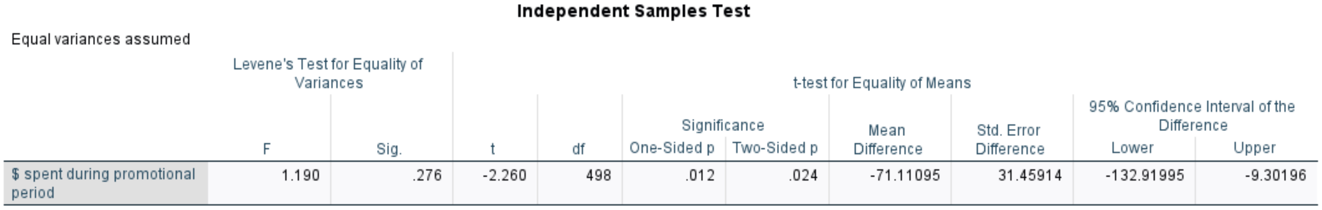
Minimum width for column labels: 0.94

Maximum width for column labels: 1.36

Minimum width for row labels: 0.68

Maximum width for row labels: 2.28





Independent Samples Test

Equal variances assumed

	Levene's Test for Equality of Variances		t-test for Equality of Means							
					Significance		Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
	F	Sig.	t	df	One-Sided p	Two-Sided p			Lower	Upper
$ spent during promotional period	1.190	.276	-2.260	498	.012	.024	-71.11095	31.45914	-132.91995	-9.30196

Results:

- The t column displays the observed t statistic (t=-2.260) for each sample, calculated as the ratio of the difference between sample means divided by the standard error of the difference.
- The df column displays the degrees of freedom. For the independent samples t test, this equals the total number of cases in both samples minus 2.
- The column labeled Sig. (2 tailed) displays a probability from the t distribution with 498 degrees of freedom. The value listed is the probability of obtaining an absolute value greater than or equal to the observed statistic, if the difference between the sample means is purely random.
- The Mean Difference is obtained by subtracting the sample mean for group 2 (the New Promotion group) from the sample mean for group 1.
- The 95% confidence interval of the difference provides an estimate of the boundaries between which the true mean difference lies in 95% of all possible random samples of 500 cardholders.
- Since the significance value of the test (.024) is less than 0.05, you can safely conclude that the average of 71.111 dollars more spent by cardholders receiving the reduced interest rate is not due to chance alone. The store will now consider extending the offer to all credit customers.

Comparing Groups

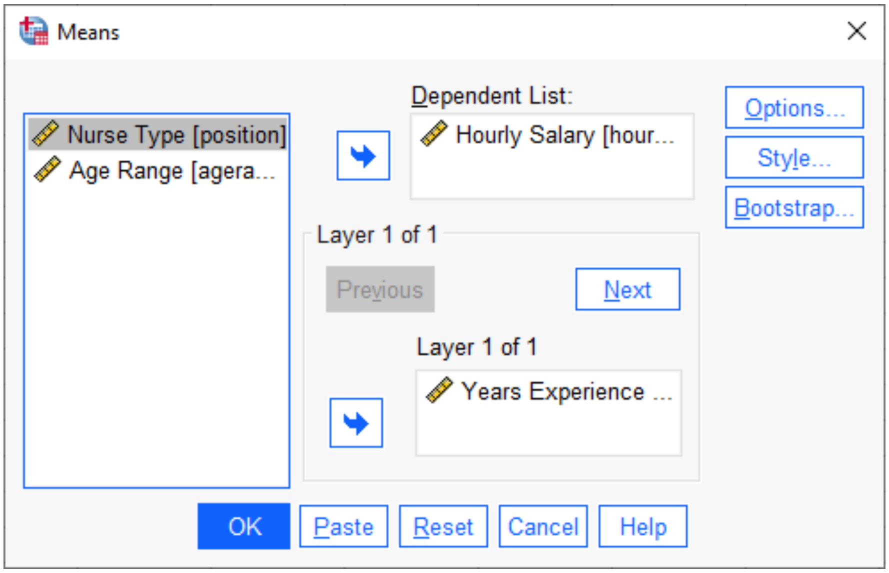
The Means procedure calculates subgroups measn and related univariate statistics for dependent variables within categories of one or more independent variables. You can test for differences between group means using **one-way ANOVA.** The one-way ANOVA in means provides you with linearity tests and association measures to help you understand the structure and strength of the relationship between the groups and their means.

Example: Measuring the average amount of fat absorbed by 3 different types of cooking oil and performing a one-way analysis of variance to see whether the means differ.

Exercise # 5 Using Means to obtain descriptive statistics

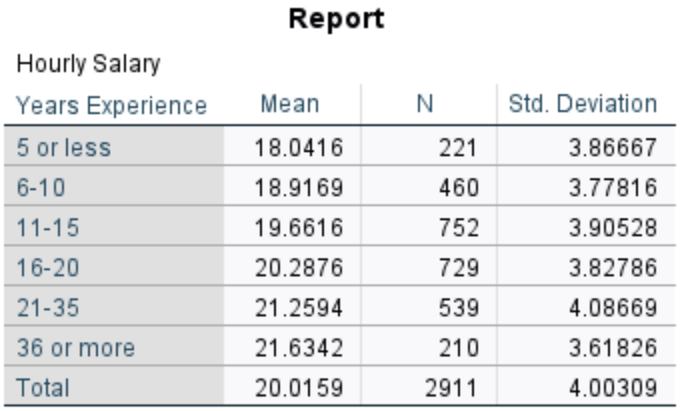
As part of an article on nurses' salaries, a journal gathers information on the hourly wages of nurses from office and hospital positions and with varying levels of experience.
This information is collected in the file **hourlywagedata.sav**. Use **Means Procedure** to examine the relationship between wages, experience, and type of position.

- Open the **hourlywagedata.sav** sample file.
  - C:\ Program Files\IBM\SPSS\Statistics\28\Sample\English\hourlywagedata.sav
- From **Analyze** menu, choose **Compare Means** -> **Means…**
- Select **Hourly Salary** as the dependent variable.
- Select **Years Experience** as the independent variable.
- Click OK.



Dependent List: Hourly Salary [hour...

Layer 1 of 1: Years Experience ...



Report

Hourly Salary

Years Experience	Mean	N	Std. Deviation
5 or less	18.0416	221	3.86667
6-10	18.9169	460	3.77816
11-15	19.6616	752	3.90528
16-20	20.2876	729	3.82786
21-35	21.2594	539	4.08669
36 or more	21.6342	210	3.61826
Total	20.0159	2911	4.00309

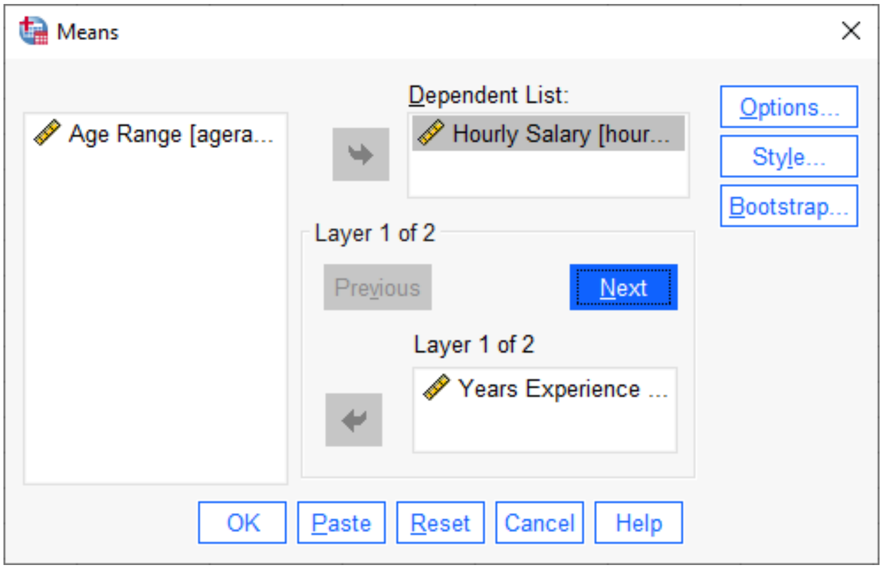
Results:

The results table displays the default statistics for salary at each experience level. Hourly salary increases in regular increments across all Years Experience categories.

Exercise #6 Layering Variables in exercise #5

Using the Means procedure, you can layer position within experience level and observe how salaries differ.

- Recall the **Means** dialog box, and then Click **Next** to increment the Layer from 1 to 2.
- Select **Nurse Type** (position) as the independent variable.
- Click OK.



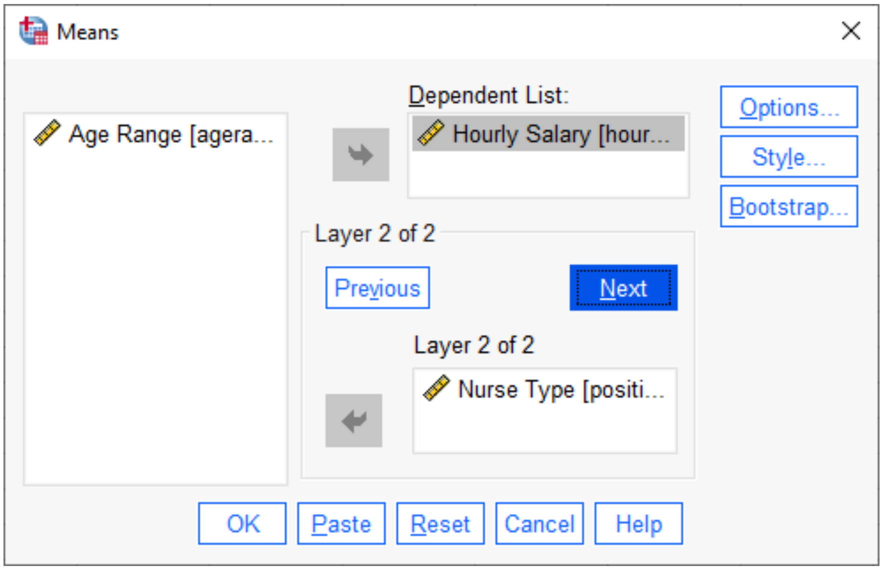
Means

Age Range [agera...

Dependent List: Hourly Salary [hour...

Layer 1 of 2

Layer 1 of 2: Years Experience ...



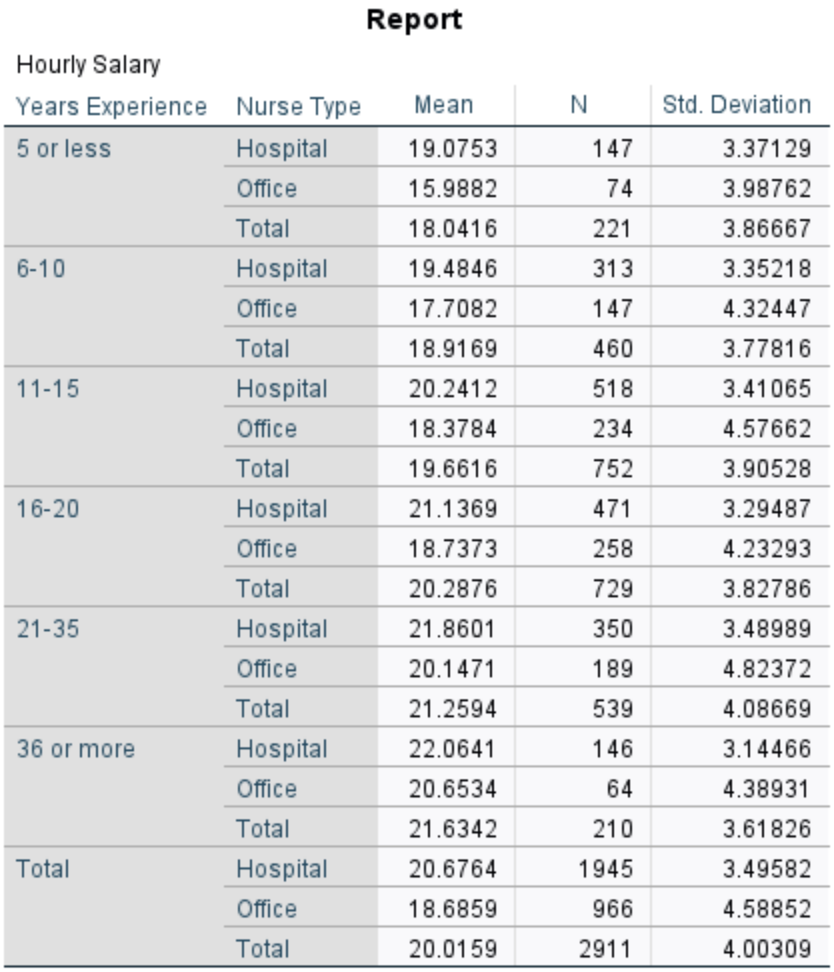
Means

Age Range [agera...

Dependent List: Hourly Salary [hour...

Layer 2 of 2

Layer 2 of 2: Nurse Type [positi...



Report

Hourly Salary

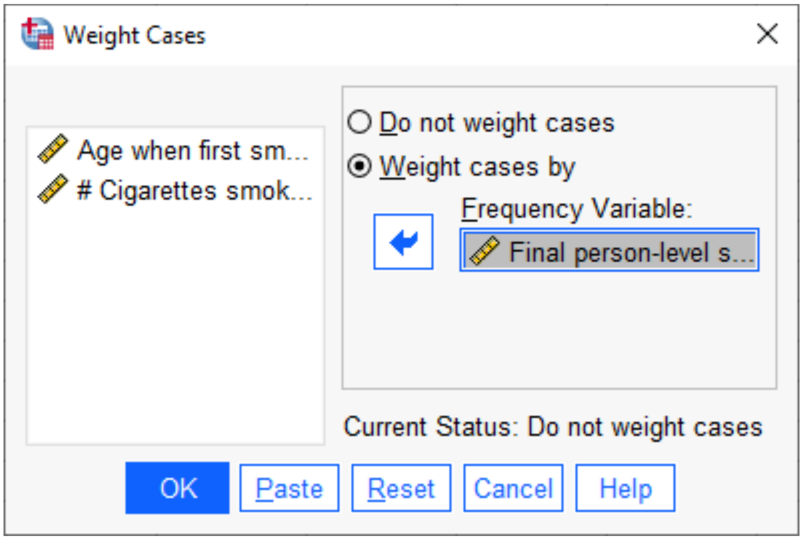
Years Experience	Nurse Type	Mean	N	Std. Deviation
5 or less	Hospital	19.0753	147	3.37129
	Office	15.9882	74	3.98762
	Total	18.0416	221	3.86667
6-10	Hospital	19.4846	313	3.35218
	Office	17.7082	147	4.32447
	Total	18.9169	460	3.77816
11-15	Hospital	20.2412	518	3.41065
	Office	18.3784	234	4.57662
	Total	19.6616	752	3.90528
16-20	Hospital	21.1369	471	3.29487
	Office	18.7373	258	4.23293
	Total	20.2876	729	3.82786
21-35	Hospital	21.8601	350	3.48989
	Office	20.1471	189	4.82372
	Total	21.2594	539	4.08669
36 or more	Hospital	22.0641	146	3.14466
	Office	20.6534	64	4.38931
	Total	21.6342	210	3.61826
Total	Hospital	20.6764	1945	3.49582
	Office	18.6859	966	4.58852
	Total	20.0159	2911	4.00309

Exercise #7 Using an ANOVA table in Means Procedure, to study Linearity between scale and categorial variables

Public health researcher is studying smoking addiction in young people. He believes the data will show that heavier smokers began smoking at a younger age than lighter smokers and is especially interested to know if the association is linear.

- Open the **smoker.sav** sample file
  - C:\ Program Files\IBM\SPSS\Statistics\28\Sample\English\smokers.sav
    Note: first we will "weight" the cases (observations) by **final person-level sample weight** variable, to let some cases in the analysis to have more "weight", because we over-sampled or under sampled from a group, to correct disproportional sample size linearity test and association measure.
- To weight the data to reflect population trends, from the **Data**, menu, choose **Weight Cases…**
- Select the **Weight cases by** radio button.
- Select Final person-level sample weight as the **frequency variable.**
- Click OK. Now the data are weighted and are ready to be analyzed.

- To begin the **Means** analysis, from the **Analyze** menu, choose **Compare Means** -> **Means…**
- Select **Age when first smoked a cigarette** as the dependent variable.
- Select **# Cigarettes smoked per day past 30 days** as the independent variable and click **Options**…
- Check the **ANOVA table and eta** and **Test for linearity** boxes and then click **continue**.
- Click OK.

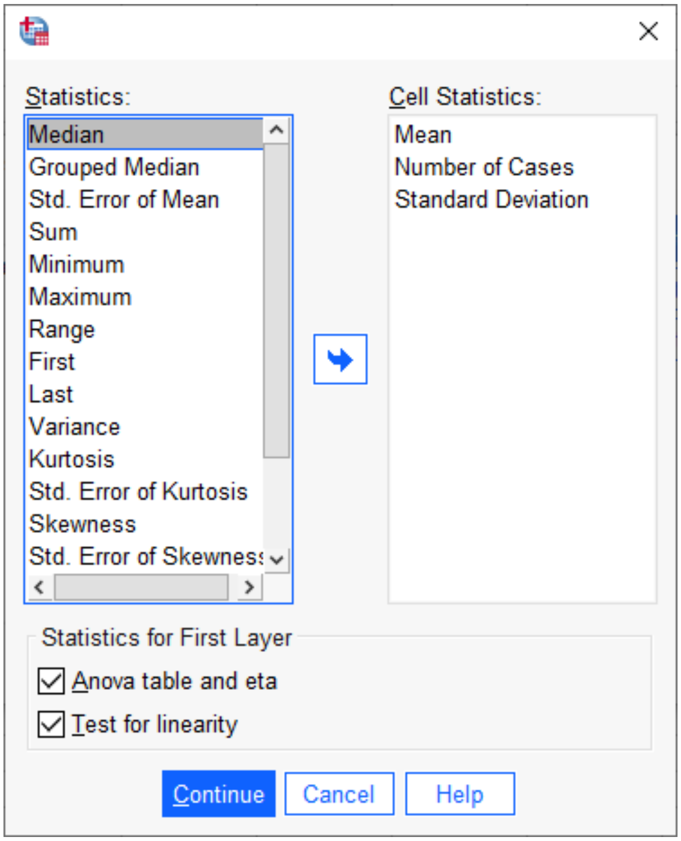

Weight Cases

Age when first sm...
Cigarettes smok...

○ Do not weight cases
● Weight cases by

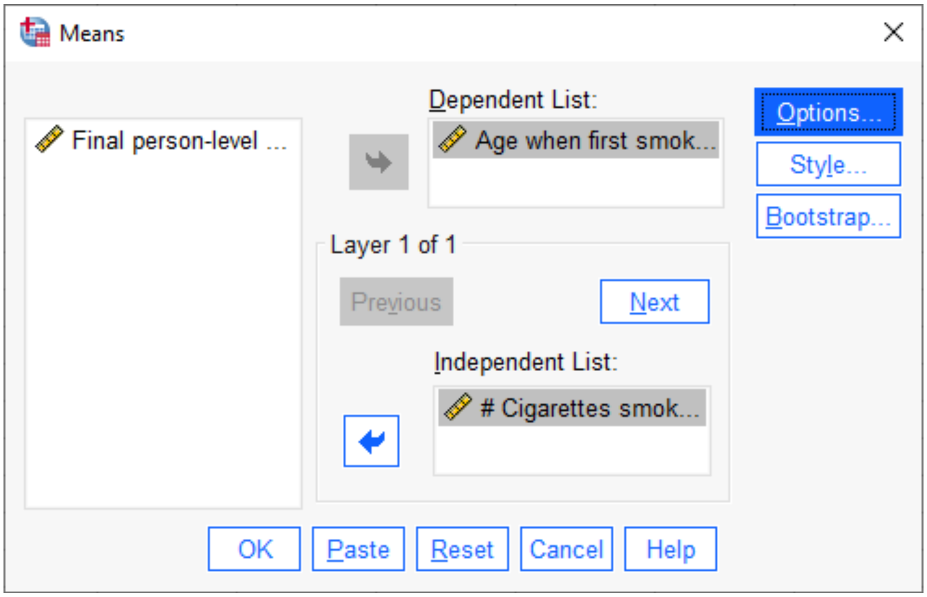
Frequency Variable: Final person-level s...

Current Status: Do not weight cases


Statistics:
Median
Grouped Median
Std. Error of Mean
Sum
Minimum
Maximum
Range
First
Last
Variance
Kurtosis
Std. Error of Kurtosis
Skewness
Std. Error of Skewness

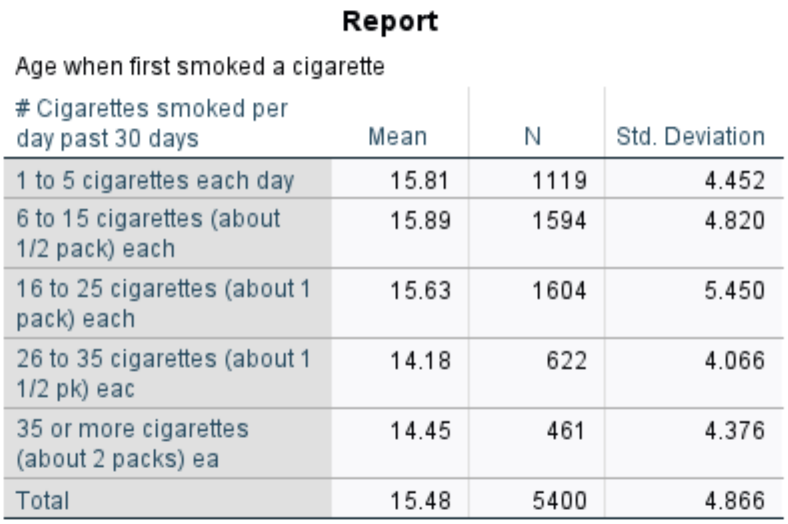
Cell Statistics:
Mean
Number of Cases
Standard Deviation

Statistics for First Layer
☑ Anova table and eta
☑ Test for linearity



Dependent List: Age when first smok...

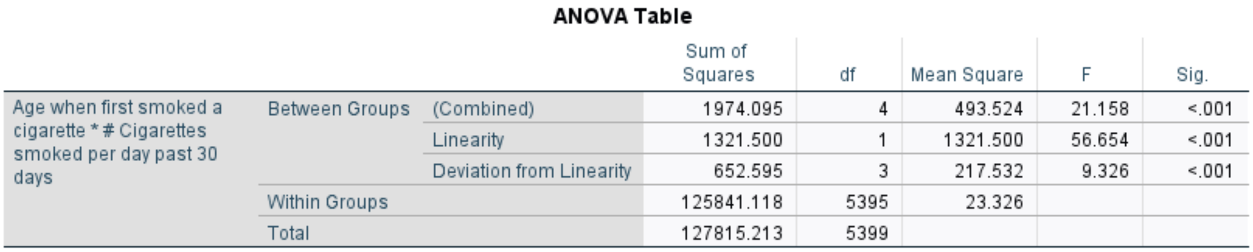
Independent List: # Cigarettes smok...



Report

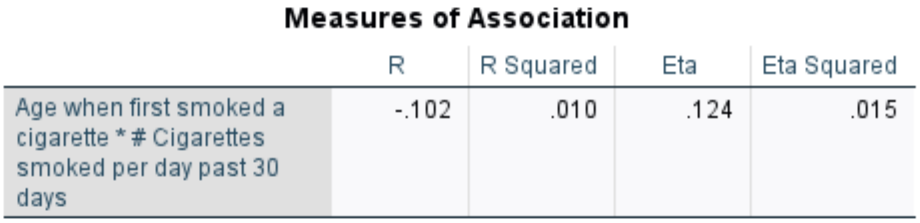
Age when first smoked a cigarette

# Cigarettes smoked per day past 30 days	Mean	N	Std. Deviation
1 to 5 cigarettes each day	15.81	1119	4.452
6 to 15 cigarettes (about 1/2 pack) each	15.89	1594	4.820
16 to 25 cigarettes (about 1 pack) each	15.63	1604	5.450
26 to 35 cigarettes (about 1 1/2 pk) eac	14.18	622	4.066
35 or more cigarettes (about 2 packs) ea	14.45	461	4.376
Total	15.48	5400	4.866



ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Age when first smoked a cigarette * # Cigarettes smoked per day past 30 days	Between Groups	(Combined)	1974.095	4	493.524	21.158	<.001
		Linearity	1321.500	1	1321.500	56.654	<.001
		Deviation from Linearity	652.595	3	217.532	9.326	<.001
	Within Groups		125841.118	5395	23.326		
	Total		127815.213	5399			



Measures of Association

	R	R Squared	Eta	Eta Squared
Age when first smoked a cigarette * # Cigarettes smoked per day past 30 days	-.102	.010	.124	.015

Results:

- Reports table indicates that teens who report smoking about one pack began doing so at almost 16 years of age.
- Comparatively, teens who report smoking over one pack per day began smoking around the age of 14.
- The **ANOVA** table contains tests for the linear, nonlinear, and combined relationship between Age when first smoked a cigarette, and # cigarettes smoked per day past 30 days.
- The test for **Linearity** has a significance value smaller then 0.05, indicating that there is a linear relationship between age and smoking level (F=56.654).
- The test for **Deviation from linearity** also has a small significance value, which means that there is a nonlinear relationship in addition to the linear component (F=9.326).
- In these data, the **squared association measure** is R^2= 0.010. The amount of variation in the age at which a person began smoking that is explained by current smoking level is statistically significant, but relatively small.

Exercise #8 Testing the Equality of Group Variances

A sales manager wishes to determine the optimal number of product training days needed for new employees. He has performance scores for the three groups: employees with one, two, or three days of training. The data are in the file **salesperformance.sav** sample file.

- Open the **salesperformace.sav** sample file
  - C:\ Program Files\IBM\SPSS\Statistics\28\Sample\English\salesperformace.sav
- To create an error bar chart, from the **Graphs** menus, choose **Chart Builder…**
- On the **Gallery** tab, select **Bar and** drag and drop the **Simple Error Bar** icon onto the canvas area.
- Drag and drop **Score on training exam** onto the y axis.
- Right-click **Sales training** group and select **Nominal** for the measurement level.
- Drag and drop sales training group onto the x axis. Under the **Element properties** tab on the right select the **Standard Error** radio button under the **Error bars represent.**
- Click OK.

  Note: that the average performance clearly increases with added training days, but the variation in performance decreases at the same time. Therefore the ANOVA assumption of equality of variance across groups may not hold for these data.

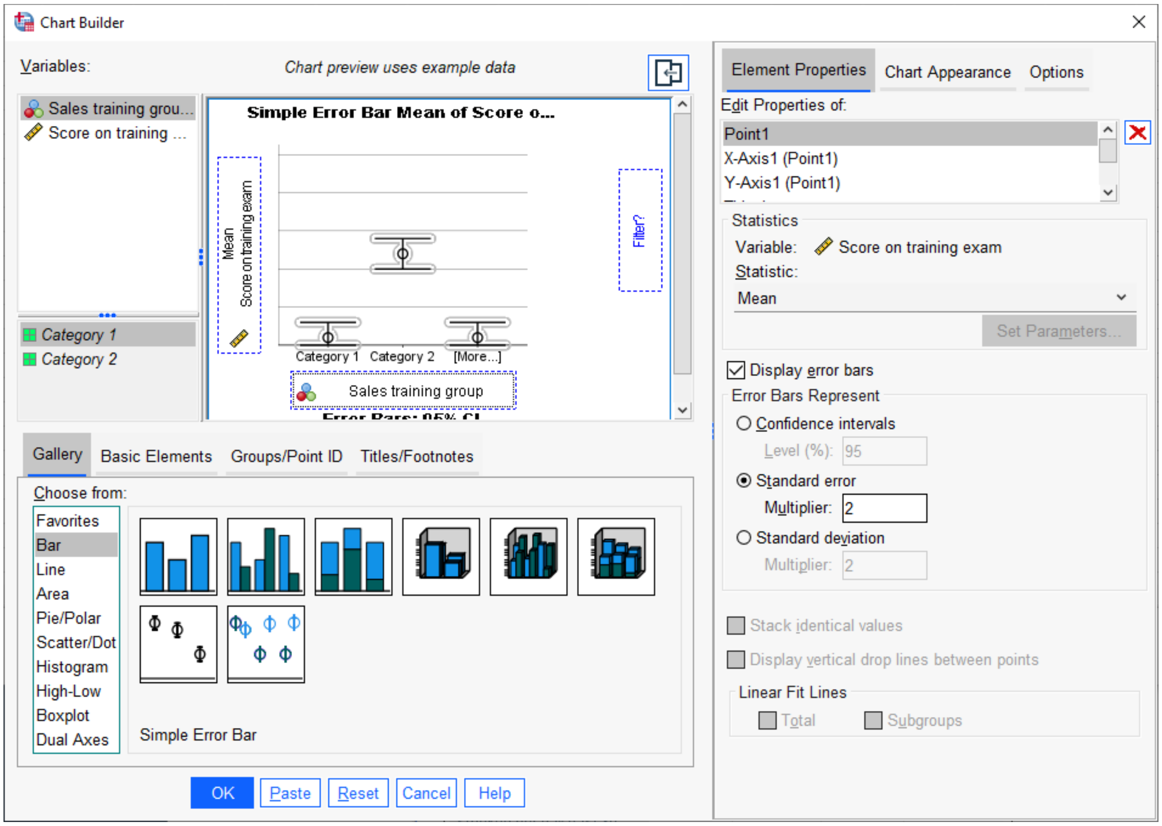


Chart Builder

Variables:

Sales training grou...
Score on training ...

Category 1
Category 2

Chart preview uses example data

Gallery | Basic Elements | Groups/Point ID | Titles/Footnotes

Choose from:

Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes

Simple Error Bar

OK | Paste | Reset | Cancel | Help

Element Properties | Chart Appearance | Options

Edit Properties of:

Point1
X-Axis1 (Point1)
Y-Axis1 (Point1)

Statistics

Variable: Score on training exam

Statistic:

Mean

Set Parameters...

☑ Display error bars

Error Bars Represent

○ Confidence intervals

Level (%): 95

● Standard error

Multiplier: 2

○ Standard deviation

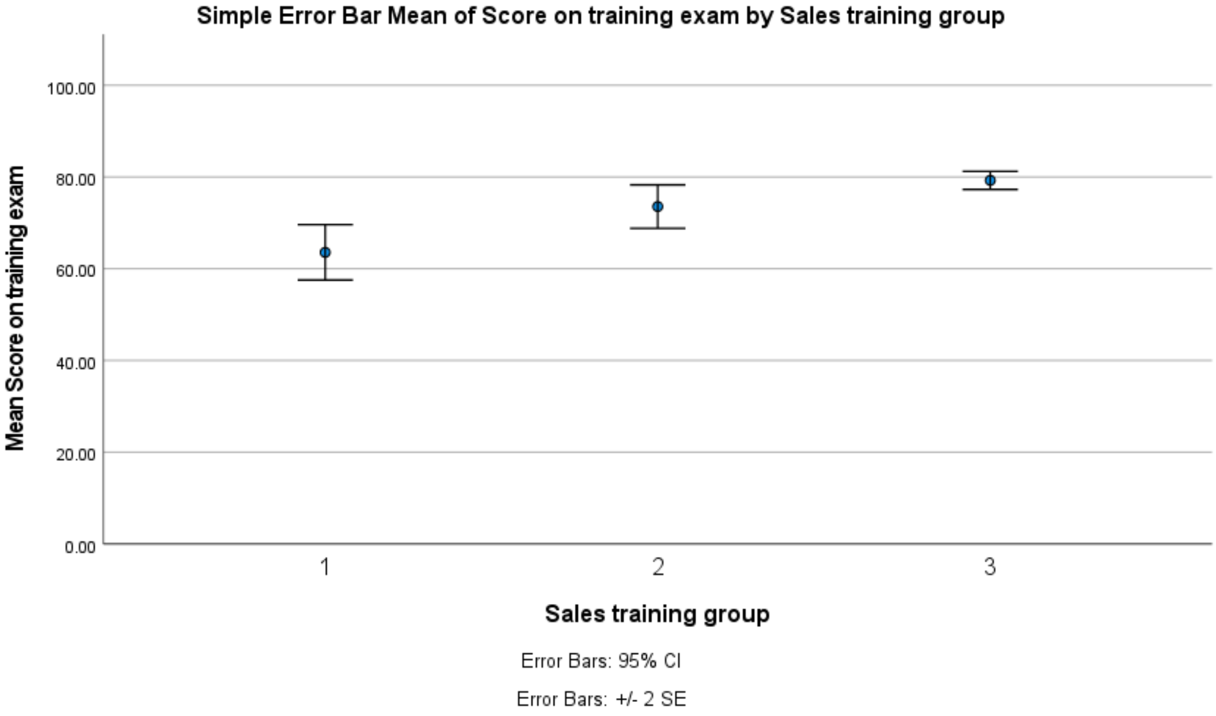
Multiplier: 2

□ Stack identical values

□ Display vertical drop lines between points

Linear Fit Lines

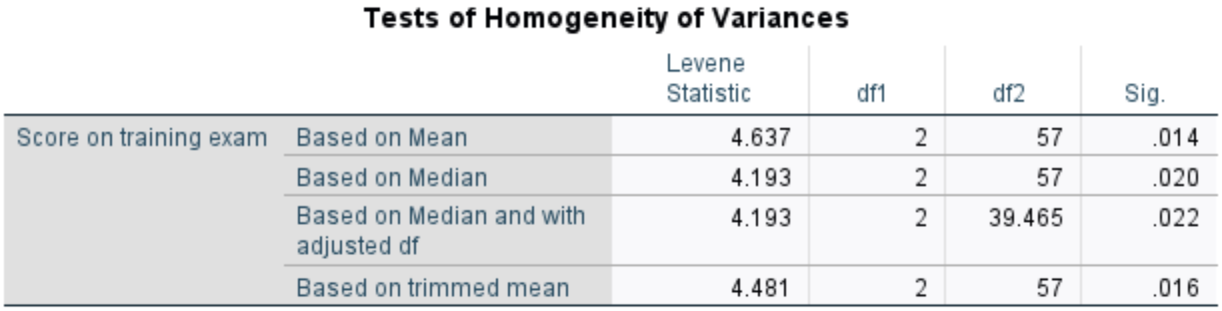
□ Total □ Subgroups



Exercise #9 Testing the equality of variance assumption in exercise #8

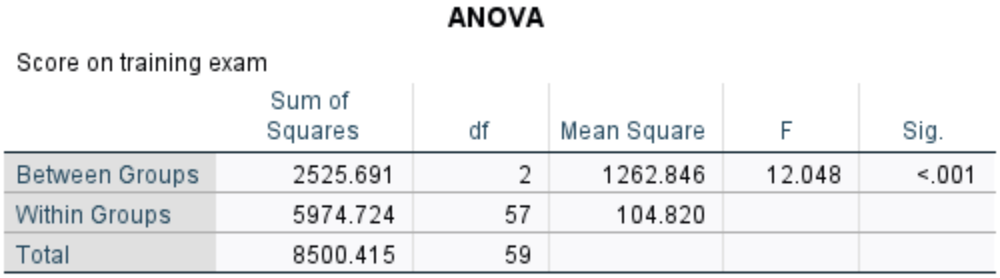
- From the Analyze menu, choose **Compare Means -> one way ANOVA…**

- Select **Score on Training exam** as the dependent variable.
- Select **Sales training group** as the factor variable and click **Options…**Check the Homogeneity of variance test box and click Continue.
- Click OK.



Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Score on training exam	Based on Mean	4.637	2	57	.014
	Based on Median	4.193	2	57	.020
	Based on Median and with adjusted df	4.193	2	39.465	.022
	Based on trimmed mean	4.481	2	57	.016



ANOVA

Score on training exam

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2525.691	2	1262.846	12.048	<.001
Within Groups	5974.724	57	104.820		
Total	8500.415	59			

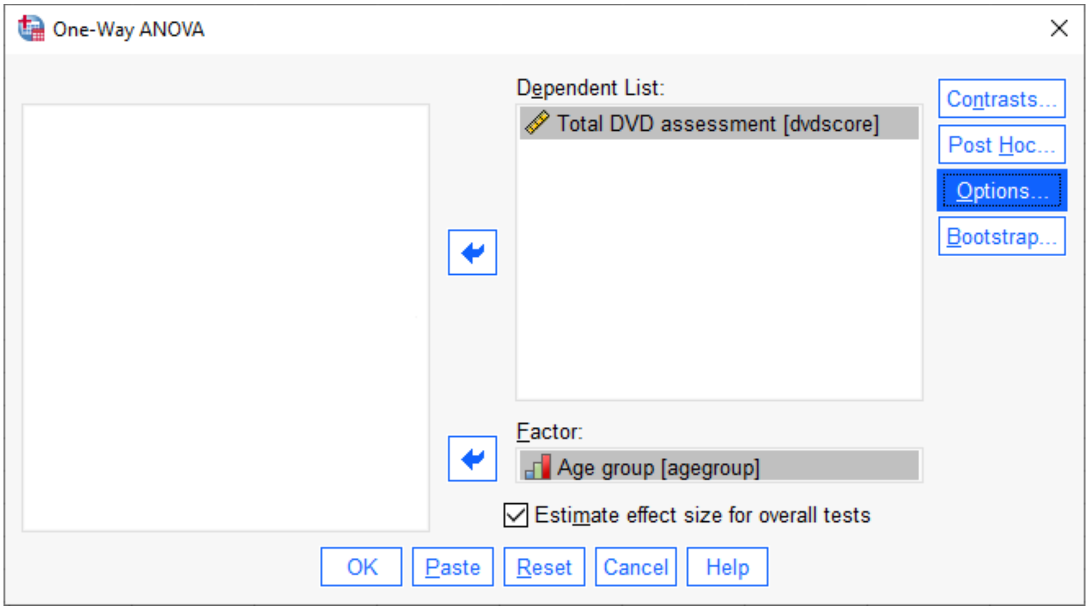
Results:

The **Levene** statistic (for equality of variances) rejects the null hypothesis that the group variances are equal (see Test of Homogeneity of variance table).

Exercise #10 Performing a One-way ANOVA

In response to customers requests, and electronics firm is developing a new DVD player. Using a prototype, the marketing team has collected focus group data. ANOVA is being used to discover is consumers of various ages rated the design differently. This exercise uses the file **dvdplayer.sav**.

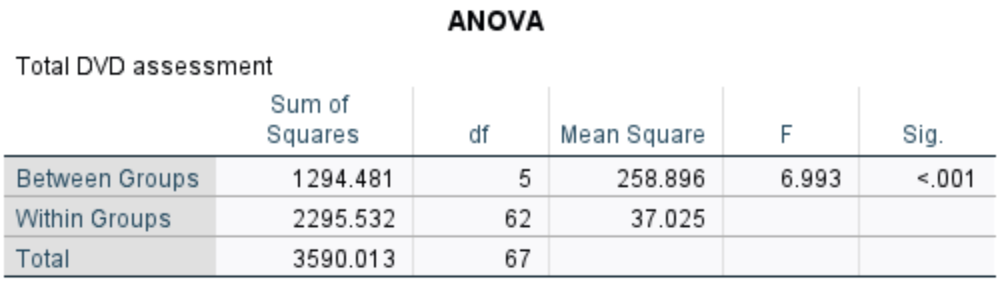
- Open the **dvdplayer.sav** sample file
    - C:\ Program Files\IBM\SPSS\Statistics\28\Sample\English\dvdplayer.sav
- From the **analyze** menu, choose **Compare Means** -> **One-Way ANOVA…**
- Select Total DVD assessments as the dependent variable.
- Select **Age Group** as the Factor Variable, and click **Options…**
- Check the **Means plot** box (the means plot is a useful way to visualize the group differences), and click **continue.**
- Click Ok.



Dependent List: Total DVD assessment [dvdscore]

Factor: Age group [agegroup]

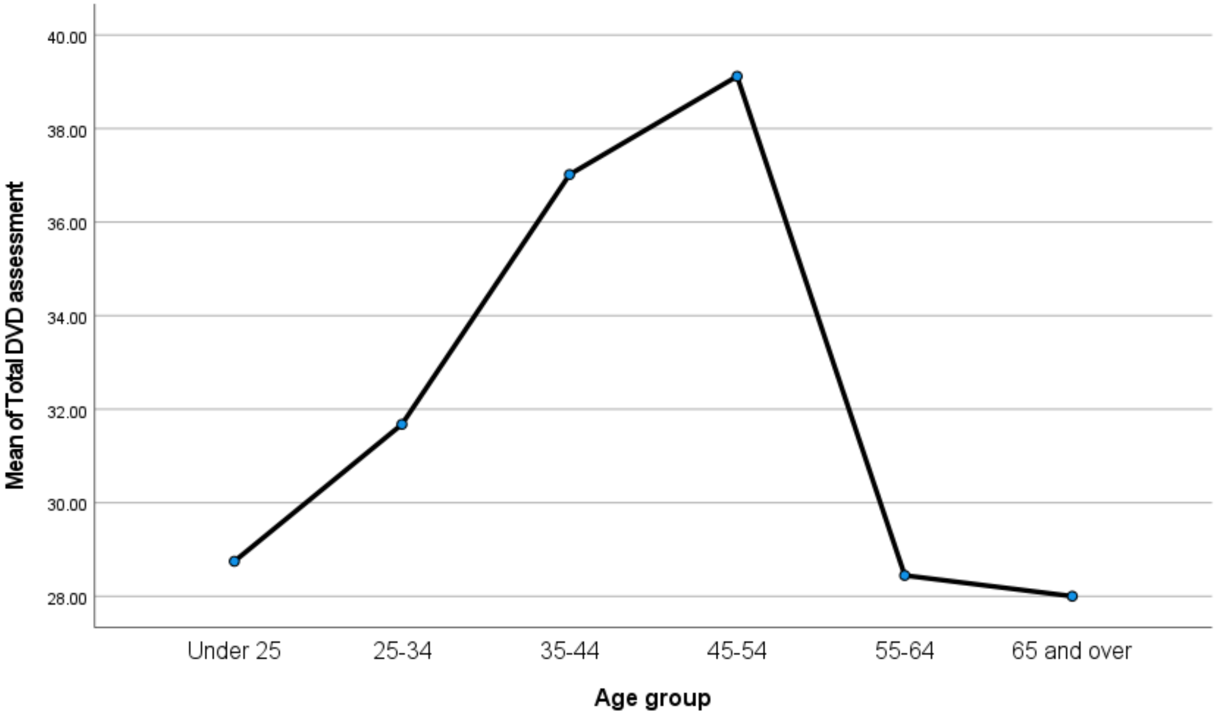
☑ Estimate effect size for overall tests



ANOVA

Total DVD assessment

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1294.481	5	258.896	6.993	<.001
Within Groups	2295.532	62	37.025		
Total	3590.013	67			



Results:

The significance value of the F test in the ANOVA table is 0.001. Thus, you must reject the hypothesis that average assessment scores are equal across age groups. Now that you know the groups differ in some way, you need to learn more about the structure of the differences. The means plot helps you to “see” this structure. Participants between the ages of 35 and 54 rated the DVD player more highly than their counterparts. If more detailed analyses are desired, then the team can use the range tests, pairwise comparisons, or contrast features in One-Way ANOVA.

Linear Regression

- Linear regression (sometimes called Ordinary Least Squares, or OLS), is used to model the value of a dependent scale variable based on its linear relationship to one or more predictors (independent variables).
- The linear regression model assumes that there is a linear relationship between the dependent variable and each predictor.
- Regression analysis is used to predict the value of one variable (the dependent variable) on the basis of other variables (the independent variables).
- Linear regression estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.
    - **Example 1:** you can try to predict a salesperson’s total yearly sales (the dependent variable) from independent variables such as age, education, and years of experience.
    - **Example 2:** Is the number of games won by a basketball team in a season related to the average number of points the team scores per game.

Data Considerations and Assumptions

- The dependent and independent variables should be quantitative. Categorical variables, such as religion, major field of study, or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables.
- For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable.
- The relationship between the dependent variable and each independent variable should be linear, and all observations should be independent.

Correlation analysis and Regression Analysis

In studies involving two or more variables, a common requirement is to investigate the existence of possible relationships between the variables. For example, an exercise physiologist might be interested in determining which physiological measurements are the best predictors of athletic performance. Or a medical researcher might be interested in relating the rate of recovery from an illness to different treatment regimes.

Two methods commonly used to discover relationships between variables are:
Correlation analysis and Regression analysis

- **Correlation analysis** provides a means of identifying **pairs of variables** that may be related, but id limited to looking at variables two at a time. Also, it provides a measure of linear association only.
- **Regression analysis** allows the effect of multiple variables upon a response variable to be explored.

Exercise #11 Using Linear Regression to Predict Polishing Times

An automotive industry group keeps track of the sales for a variety of personal motor vehicles. In an effort to be able to identify over- and underperforming models, you want to establish a relationship between vehicle sales and vehicle characteristics.

Information concerning different makes and models of cars is contained in car_sales.sav. Use linear regression to identify models that are not selling well.

- Open the **car_sales.sav** sample file
  - C:\ Program Files\IBM\SPSS\Statistics\28\Sample\English\car_sales.sav
- To produce a scatter plot of the variables, from the **Graphs** menu, choose **Chart Builder…**
- Click the scatter/dot gallery, choose Simple Scatter, and drag it to the chart canvas area.
- Select **Logtransformed sales (lnsales)** variable, drag it to **y-axis** area and **price in thousands** to **x-axis area.**
  - *Note: The distribution of Log-transformed sales is closer to normal than Sales in thousands, and the linear regression model works better with normal variables.*
- Click OK.
- To see a best fit line overlaid on the points in the scatter plot, activate the graph by double clicking it.
- Click the **add fit line tool,** then close the chart editor.
- To a run a linear regression analysis, from the **analyze** menu, choose -> **regression** -> **Linear…**
- Select **logtransformed sales** as the dependent variable.
- Select **price in thousands** as the independent variable.
- Select **model** as the case labeling variable.
- Click **Plots…**
- Check the **histogram** and **normal probability plot** boxes and click **continue**.
- Click **continue.**
- Click ok.
- Next select **analyze** -> **descriptive statistics** -> **descriptives…**
- Move logtransformed sales and price in thousands over to the variables box and click OK.

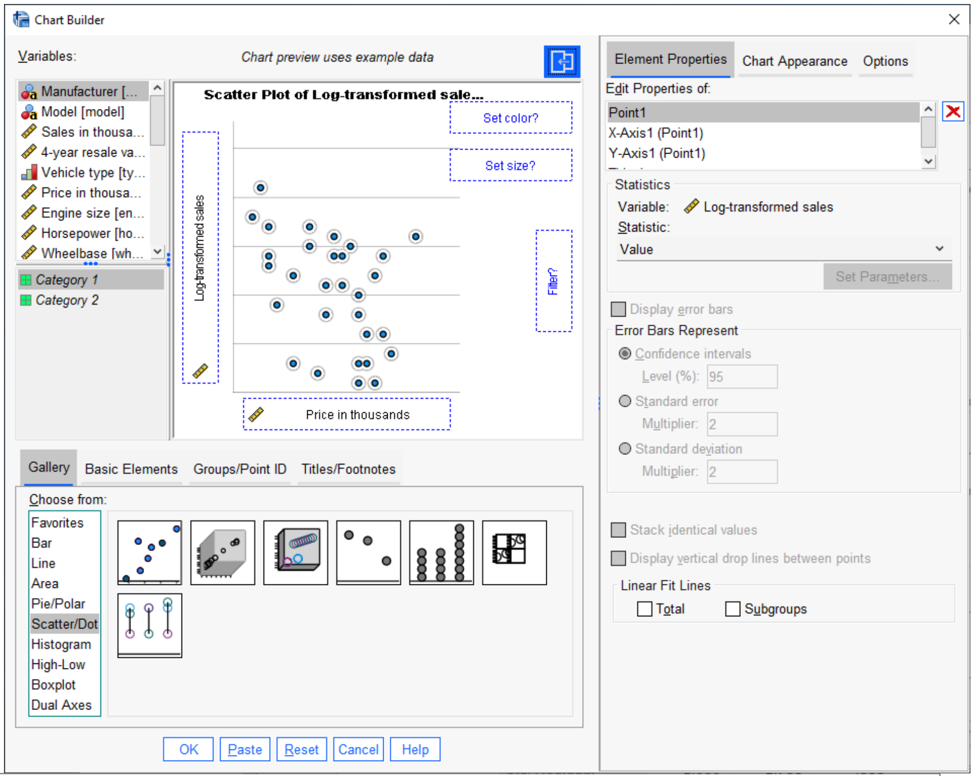


Chart Builder
Variables:
Chart preview uses example data
Manufacturer [...
Model [model]
Sales in thousa...
4-year resale va...
Vehicle type [ty...
Price in thousa...
Engine size [en...
Horsepower [ho...
Wheelbase [wh...
Category 1
Category 2
Scatter Plot of Log-transformed sale...
Set color?
Set size?
Log-transformed sales
Filter?
Price in thousands
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes
OK
Paste
Reset
Cancel
Help
Element Properties
Chart Appearance
Options
Edit Properties of:
Point1
X-Axis1 (Point1)
Y-Axis1 (Point1)
Statistics
Variable: Log-transformed sales
Statistic:
Value
Set Parameters...
Display error bars
Error Bars Represent
Confidence intervals
Level (%): 95
Standard error
Multiplier: 2
Standard deviation
Multiplier: 2
Stack identical values
Display vertical drop lines between points
Linear Fit Lines
Total
Subgroups

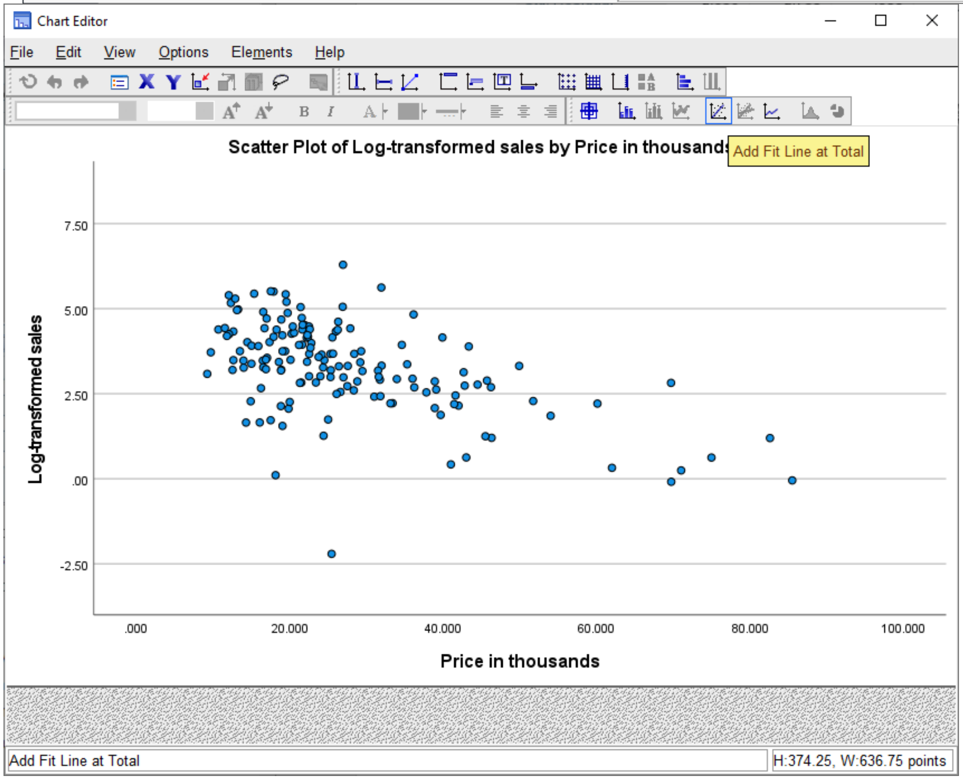
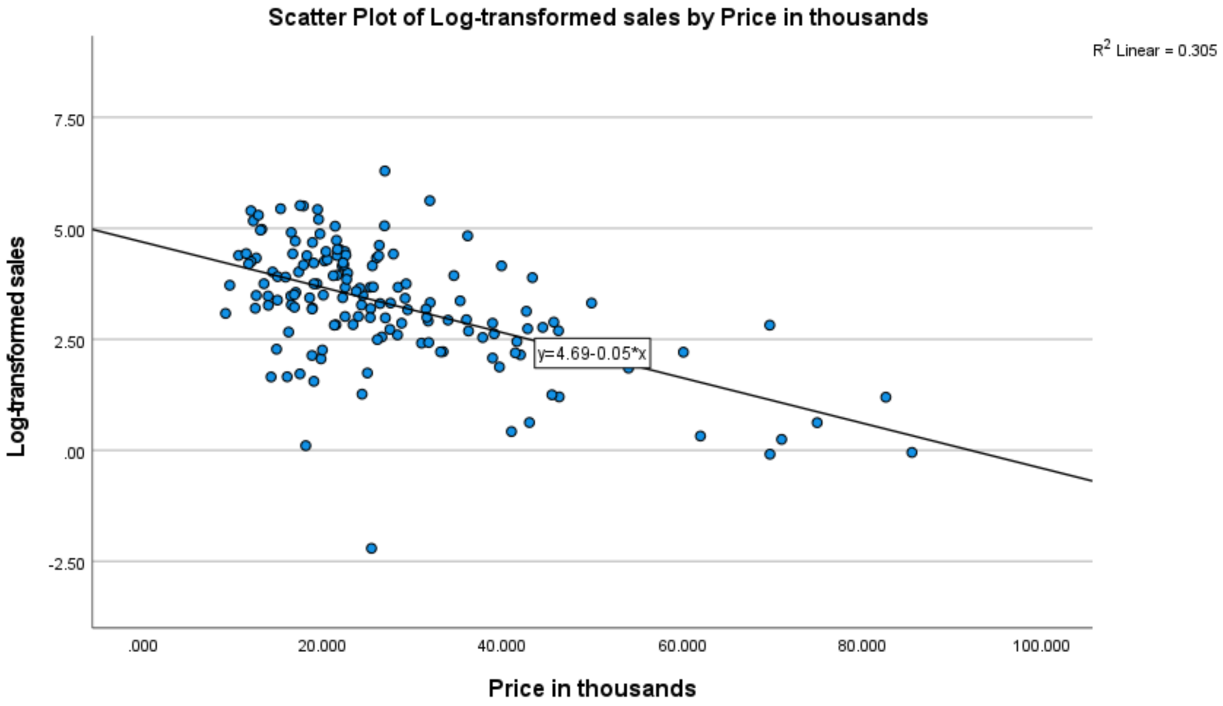
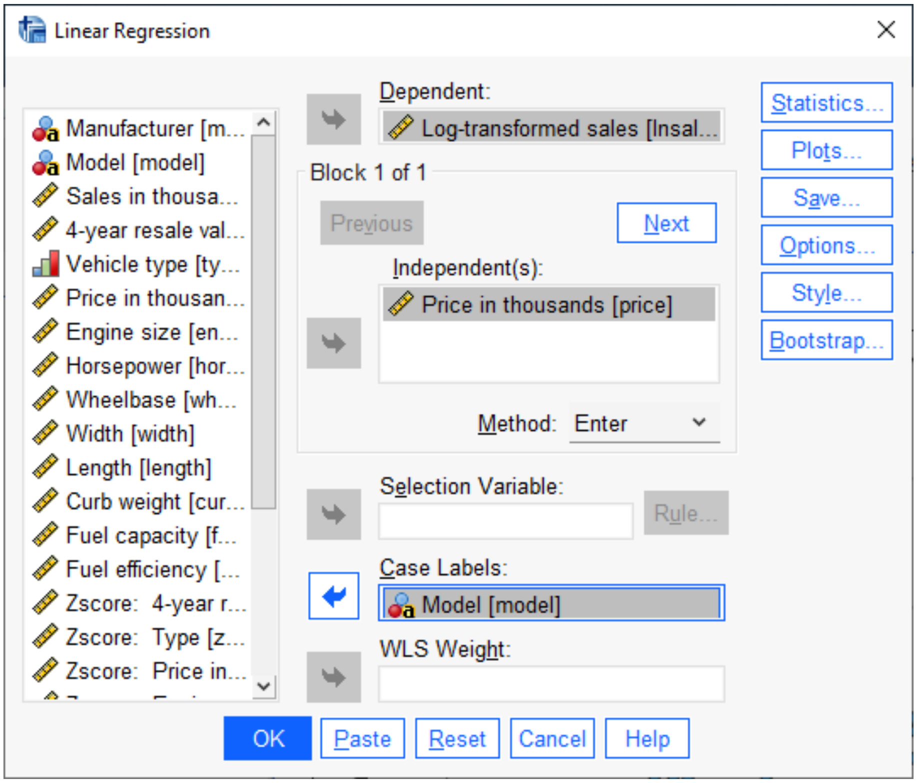


Chart Editor
File
Edit
View
Options
Elements
Help
Scatter Plot of Log-transformed sales by Price in thousands
Add Fit Line at Total
7.50
5.00
2.50
.00
-2.50
Log-transformed sales
.000
20.000
40.000
60.000
80.000
100.000
Price in thousands
Add Fit Line at Total
H:374.25, W:636.75 points





Dependent: Log-transformed sales [lnsal...

Block 1 of 1

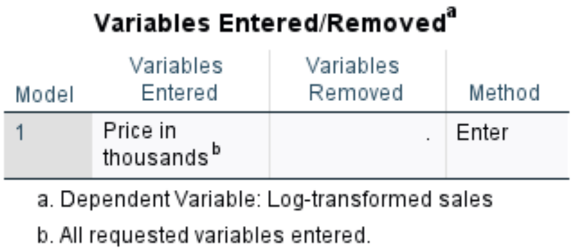
Independent(s): Price in thousands [price]

Method: Enter

Selection Variable:

Case Labels: Model [model]

WLS Weight:

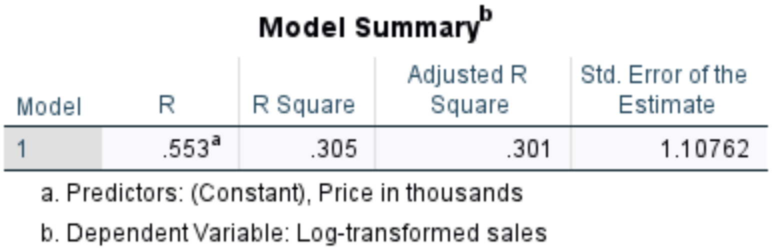


Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Price in thousands[b]	.	Enter

a. Dependent Variable: Log-transformed sales

b. All requested variables entered.

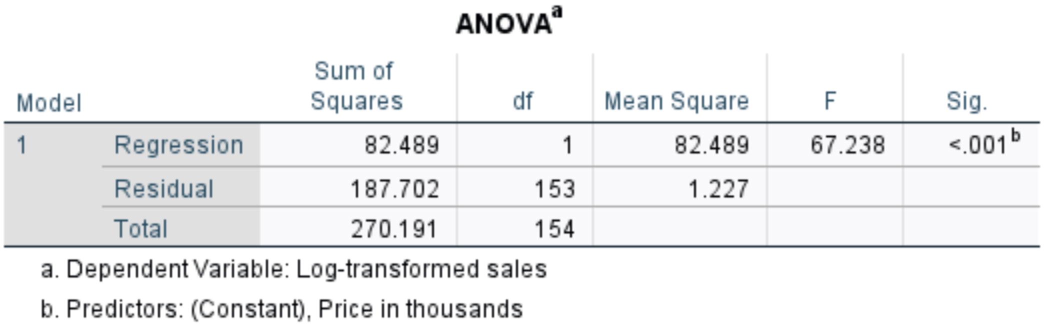


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.553[a]	.305	.301	1.10762

a. Predictors: (Constant), Price in thousands

b. Dependent Variable: Log-transformed sales

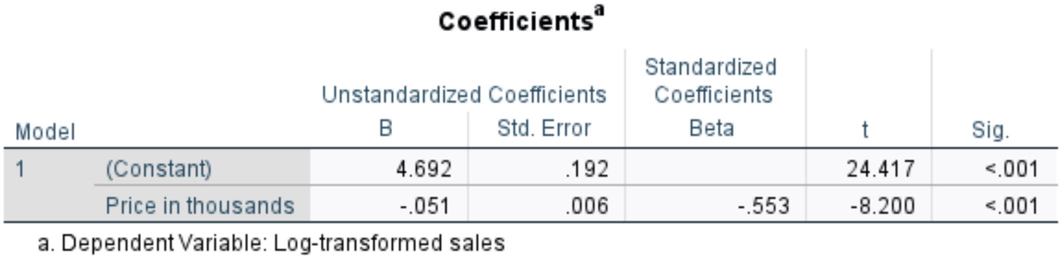


ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	82.489	1	82.489	67.238	<.001[b]
	Residual	187.702	153	1.227		
	Total	270.191	154			

a. Dependent Variable: Log-transformed sales

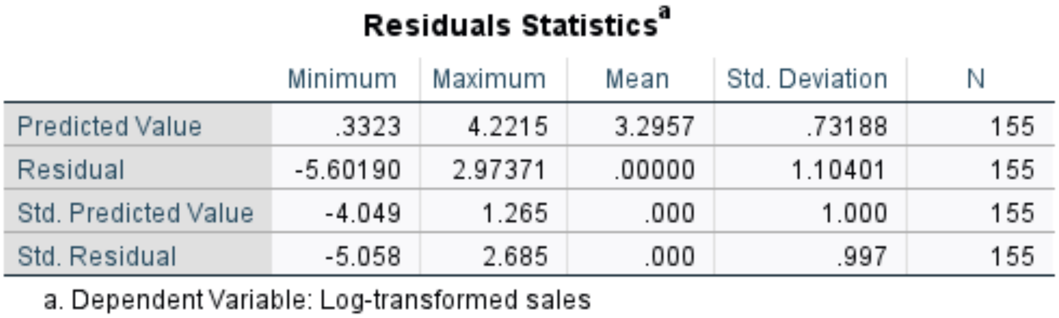
b. Predictors: (Constant), Price in thousands



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.692	.192		24.417	<.001
	Price in thousands	-.051	.006	-.553	-8.200	<.001

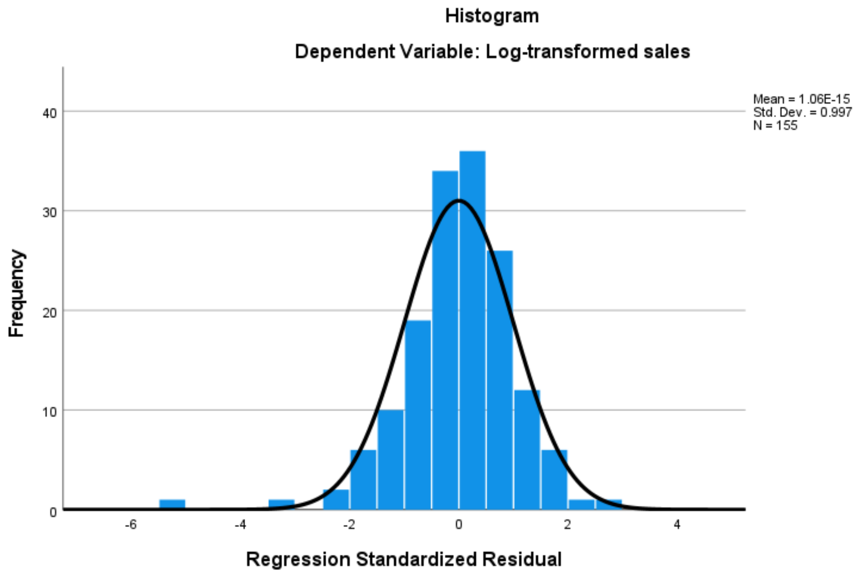
a. Dependent Variable: Log-transformed sales

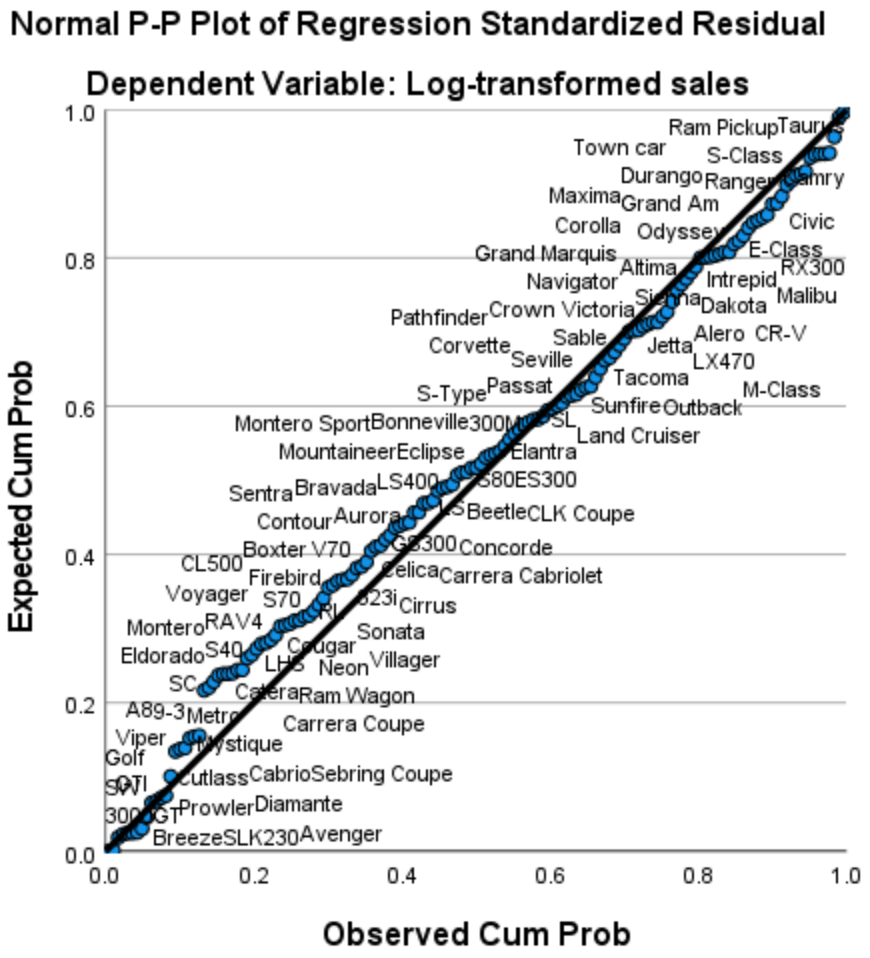


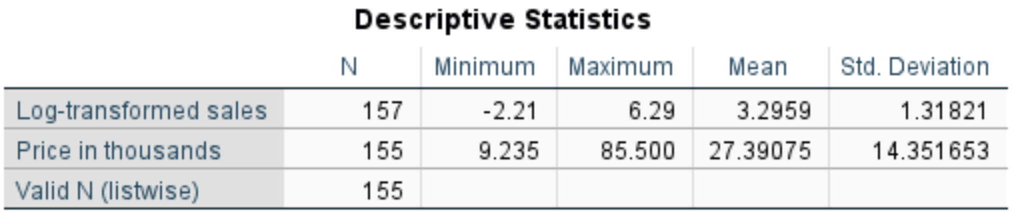
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.3323	4.2215	3.2957	.73188	155
Residual	-5.60190	2.97371	.00000	1.10401	155
Std. Predicted Value	-4.049	1.265	.000	1.000	155
Std. Residual	-5.058	2.685	.000	.997	155

a. Dependent Variable: Log-transformed sales


Histogram
Dependent Variable: Log-transformed sales
Mean = 1.06E-15
Std. Dev. = 0.997
N = 155
Frequency
40
30
20
10
0
-6
-4
-2
0
2
4
Regression Standardized Residual


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: Log-transformed sales
Expected Cum Prob
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Observed Cum Prob
Ram Pickup
Taurus
Town car
S-Class
Durango
Ranger
Camry
Maxima
Grand Am
Corolla
Odyssey
Civic
Grand Marquis
E-Class
Altima
RX300
Navigator
Intrepid
Malibu
Sienna
Dakota
Pathfinder
Crown Victoria
Corvette
Sable
Jetta
Alero
CR-V
Seville
LX470
Tacoma
S-Type
Passat
M-Class
Sunfire
Outback
Montero Sport
Bonneville
300M
SL
Land Cruiser
Mountaineer
Eclipse
Elantra
Sentra
Bravada
LS400
S80
ES300
Beetle
CLK Coupe
Contour
Aurora
Boxter
V70
GS300
Concorde
CL500
Firebird
Celica
Carrera Cabriolet
Voyager
S70
RL
323i
Cirrus
Montero
RAV4
Sonata
Eldorado
S40
Cougar
LHS
Neon
Villager
SC
Catera
Ram Wagon
A89-3
Metro
Carrera Coupe
Viper
Mystique
Golf
GTI
Cutlass
Cabrio
Sebring Coupe
SW
3000GT
Prowler
Diamante
Breeze
SLK230
Avenger



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Log-transformed sales	157	-2.21	6.29	3.2959	1.31821
Price in thousands	155	9.235	85.500	27.39075	14.351653
Valid N (listwise)	155				

Results:

- The coefficients table (shows the coefficients of the regression line) show that the expected log transformed sales is equal to -.051 * sales price + 4.692. So, according to this model if a car typically sells for $25,300 you should expect the log transformed sales to be -.051* 25.300 + 4.692 or around 3.4017.
- The ANOVA table indicates that the regression model predicts the log transformed sales significantly well. Look at the "Regression" row and go to the "Sig." column. This indicates the statistical significance of the regression model that was run. Here, p < 0.0005, which is less than 0.05, and indicates that, overall, the regression model statistically significantly predicts post test scores (i.e., it is a good fit for the data).
- The significance value of the F statistic (F= 67.238) is less than 0.05, which means that the variation explained by the model is not due to chance.
- R, the multiple correlation coefficient, is the linear correlation between the observed and model predicted values of the dependent variable (log transformed sales). This value is medium sized indicating a moderately strong relationship.
- R square, the coefficient of determination, is the squared value of the multiple correlation coefficient. It shows that about 30% of the of the variance in log transformed sales is explained by this model.
- As a further measure of the strength of the model fit, compare the standard error of the estimate in the model summary table (1.10762) to the standard deviation of log transformed sales reported in the descriptive statistics table (1.31821).
- Without prior knowledge of the purchase price of a vehicle, your best guess for the log transformed sales would be 3.2959 with a standard deviation of 1.31821.
- With this linear regression model, the error of your estimate is slightly lower at 1.10762.
- A residual is the difference between the observed and model predicted values of the dependent variable. The residual for a given model is the observed error term for that model.
- The shape of the histogram should approximately follow the shape of a normal curve. This histogram does.
- The P-P plotted residuals should follow the 45 degree line. Neither the histogram nor the P-P plot indicate that the normality assumption is violated.