**Running a simple linear regression analysis in SPSS**

This example uses data on housing prices. You can find a [description of where to find the data and what the variables represent](https://github.com/pmean/classes/blob/master/biostats-2/data/woodard-data-dictionary.yaml) on my Github site..

First, it is always a good idea to plot the data before you run the regression analysis. Select Graphs | Scatter/Dot from the SPSS menu. This produces the Scatter/Dot dialog box.

A screenshot of a computer

Description automatically generated

Click on the Simple Scatter and then the Define button. This produces the Simple Scatterplot dialog box.

A screenshot of a computer

Description automatically generated

Add Land $ to the Y Axis field and Acres to the X Axis field.

Click on the OK button to produce your graph.

A graph with lines in the middle

Description automatically generated

In the example shown here, there is a clear and obvious outlier. You should always be cautious about removing outliers, but here is makes sense. You don’t want to draw predictions about the relationship between the size of the lot and the price for lots more than 5 acres, because you have almost no data for such large lots. Restrict your attention to lot sizes where you have more data.

To this point before redrawing the graph, choose Data | Select cases from the menu. This produces the Select Cases dialog box.

A screenshot of a computer

Description automatically generated

Select the If condition is satisfied option and click on the If button. This produces the Select Case If dialog box.

A screenshot of a computer

Description automatically generated

Add the condition Acres < 5 to exclude the one very big lot. Then click on Continue in this dialog box and OK on the previous dialog box.

This is what your data now looks like.

A screenshot of a computer

Description automatically generated

Your data now has a new column (filter\_$) with values of Selected for everything except the fifth row, which has Not Selected. Notice the slash through the row number, further indicating that this value is excluded. This exclusion is temporary and you can undo it at any time.

After redrawing, you should make some adjustments. These might include

1. Placing your name and date in the title
2. Using a comma separator for the labels on the Y-axis
3. Adding tick marks on the X and Y axes.
4. Changing the scale margins to smaller values (2% or 0%).
5. Converting the blue marker interiors to invisible
6. Changing the axes (minimum, maximum, and/or major increment)
7. Adding a trendline

This is the resulting plot.

A graph of a plot

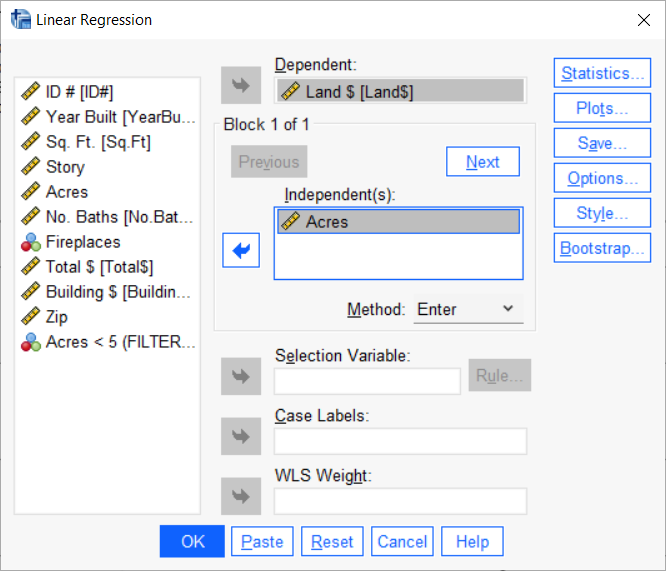
Description automatically generated with medium confidence

To run a regression analysis, you have two options:

1. Select Analyze | Regression | Linear, or
2. Select Analyze | General Linear Model | Univariate

Both have advantages and disadvantages. For simple problems, the former is best and that is the choice illustrated here.

This is the Linear Regression dialog box.



Start by adding the dependent variable to the dependent field and the independent variable to the independent field.

Click on the Statistics button to get the Linear Regression Statistics dialog box.

A screenshot of a computer

Description automatically generated

The options for Estimates and Model fit are selected by default. Add the option for Confidence intervals.

Click on Continue to go back to the Linear Regression dialog box.

Then click on the Save button to bring up the Linear Regression Save dialog box.

A screenshot of a computer

Description automatically generated

Click the boxes for unstandardized residuals and unstandardized predicted values.

Finish up by clicking on Contnue and then OK. This produces five numeric tables. The first table is only relevant in regression models with multiple independent variables, where you might consider adding or removing independent variables.

A screen shot of a screen

Description automatically generated

The second table gives the R Square value. In this model, the size of the lot explains 3% of the variation in lot price.

A screenshot of a graph

Description automatically generated

The third table is the analysis of variance table.

A table with numbers and symbols

Description automatically generated

You have 1 degree of freedom for Regression and 97 degrees of freedom for Residual (or Error). The sum of squares and mean squares are rather large because the lot prices are in the tens and hundreds of thousands. The F-ratio is 3.147. This is close to 1. The p-value, 0.079, is larger than 0.05, so you would accept the null hypothesis and conclude that there is no relationship between the size of the lot in acres and its price in dollars. Some researchers would describe this as a “borderline” finding because the p-value is only slightly larger than 0.05.

The fourth table provides the regression coefficients, statistical tests, and confidence intervals.

A close-up of a graph

Description automatically generated

The intercept is $62,000 and the slope is $40,000. The estimated average lot price increases by 40 thousand dollars for each increase of one acre in the size of the lot. The interpretation of the intercept is a bit difficult--the estimated average price of a lot is 62 thousand dollars for a lot of zero acres.

Notice that the t-statistic for acres is close to zero and the p-value is larger than 0.05. This is testing the hypothesis that the slope is zero. You would accept the null hypothesis and conclude that there is no relationship between the size of the lot in acres and price in dollars. This is consistent with the interpretation provided earlier.

The 95% confidence interval for the slope ranges from -5 thousand dollars to 85 thousand dollars. This confidence interval includes the value of zero, leading you to draw the same conclusion as earlier.

Normally, you do not test any hypotheses or compute any confidence intervals for the intercept.

The fifth table produces simple statistics for the residuals and predicted values.

A table with numbers and a few black text

Description automatically generated

There is nothing of real interest in this table.

To test normality, draw a QQ plot (a histogram would also be fine). Select Analyze | Descriptive Statistics | Q-Q plots from the menu. This produces the Q-Q Plots dialog box.

A screenshot of a computer

Description automatically generated

The default options are all fine.

The Q-Q plot produces three tables and two graphs. The first table provides some technical details about the plot, but none are important for the most data analysts.

A screenshot of a model description

Description automatically generated

The second table gives an accounting of missing values. This is important and you should always track the number of missing values in any data analysis.

A screenshot of a computer

Description automatically generated

The third table gives the estimated mean and standard deviation. The mean of the residuals is always zero. The scale (standard deviation) is 71,000. It can vary from data set to data set, but is usually not of direct interest.

A close-up of a box

Description automatically generated

The first plot is the important one.

A graph with a line

Description automatically generated

For this data set, the QQ plot is not anywhere close to a straight line. The assumption of normality is not met.

The final plot is a detrended plot. It has no value and should be ignored.

A graph with blue dots

Description automatically generated

You should also examine a scatterplot with the independent variable on the X asix and the residuals on the Y axis. This allows you to check for nonlinearity and unequal variances.

The dialog box is similar to the one shown earlier. Here is the graph with some enhancements.

A graph of a plot

Description automatically generated

There is no evidence of nonlinearity and no evidence of unequal variances. There are three outliers where the residual is above 200 thousand. This indicates houses where the observed price was quite a bit larger than the predicted price.