Demonstrate Regression Teaching Module
Lessons (Developed by Bill Duckworth, Creighton)

In JMP 12 and JMP 12 SE this module is available in the Sample Data Directory
(Help > Sample Data, under Teaching Resources > Teaching Scripts > Interactive
Teaching Modules.)

The module is also available from the JMP Academic Community as part of the
Interactive Teaching Modules add-in (https://community.jmp.com/docs/DOC-6295).
Regression Lesson Zero: Adjusting Population Characteristics

The default choice for the Select Data Source option is “Population Normal.” Unless otherwise directed, leave this at the default. The "My Sample" and "Create Your Own" choices will be demonstrated in later lessons.

This lesson explores the Population Characteristics controls. The sample data plotted on the right side of the script window are sampled from a population with the characteristics specified on the left side of the script window. If you change a population characteristic via slider or direct entry, the script dynamically changes the characteristics of the sampled data as shown in the scatterplot.

- Try adjusting the Intercept slider slowly from the default of 5.00 down to -5 and then up to 10. Notice how the sample data moves in correspondence with the changes in the Intercept value. Now, click in the textbox to the right of the Intercept slider and enter a specific value for the intercept, say 3.14. (To see the effects of the newly entered value either click outside the textbox or hit the Return key on your keyboard.)

- Now, try adjusting the Slope by entering -1 in the Slope textbox. Notice the change in the scatterplot. Now, slowly move the Slope slider to the right to increase the slope to 1. You will likely see that the sample data moves completely out of the scatterplot display.

While adjusting the various Population Characteristics, you will frequently find that the scatterplot will no longer contain all of the sample data points. No worries! Just click the “Rescale” button at the bottom of the Population Characteristics area and the script will attempt to adjust the scatterplot axes to show the sample data again. The “Zoom Out” button can also help you obtain a better view of your data in some cases. The “Revert to Original” button will restore the Population Characteristics to their default values including the scatterplot.

Take some time to experiment with the other Population Characteristics that are adjustable: Correlation, Name of Y, Name of X, Mean of X, Std Dev of X. Notice that adjusting the Mean of X or Std Dev of X produces changes in the scale of the X and Y axes, and that adjusting the Correlation to a value with a different sign produces a corresponding change in the sign of the Slope value because of the mathematical relationship between the slope and the correlation.
Regression Lesson 1: Fit Your Own Line

After adjusting the Population Characteristics as desired, click on the red triangle (hot spot) near the upper-left corner of the scatterplot (beside the scatterplot title “Sample Data”). You will see the following commands:

Let’s begin with the second set of commands: Fit My Line, Fit My Residuals, and Show My Squares. These three commands allow us to explore the results of attempting to find the line of best fit by the “eyeball method.”

- Select Fit My Line. The script draws a blue horizontal line on the scatterplot at the mean of Y. Attached to the ends of the line are two boxes that allow you to move the line. Simply click inside one of the boxes and drag that end of the line to the desired location. By using both boxes, you can drag the line until you believe it best fits the overall data pattern.
- In the far right area of the window, the script also plots the residuals from your line against the X-variable. If you select Fit My Residuals, vertical lines will be drawn from each data point to your line in the scatterplot allowing you to visualize your residuals in another way. Select Fit My Residuals a second time. The script removes the vertical lines representing the residuals for your line. Most of the commands in this menu will toggle on/off the requested option.
- Finally, select Show My Squares. This command visualizes the squared errors of the SSE (Sum of Squared Errors) as blue squares with side lengths equal to the residuals of each data point with respect to your line. The SSE is calculated and displayed visually on the SSE bar graph to the right of the scatterplot. Recall that the line of best fit is the line that minimizes the SSE, so you may want to fine tune your fitted line (by dragging the two boxes) to see if you can decrease your SSE!

The first set of commands are similar to the second set except they apply to the Least-Squares (LS) line rather than the line you manually fit to the data.
To see how well you did fitting your line to the data, select Fit LS Line. The script now draws the Least-Squares (LS) regression line in red (and provides much of the standard JMP output for simple linear regression in the far right area of the window).

- How well did you do with your (blue) line?
- Is the slope of the LS line greater or less than your slope?
- How do the intercepts compare?

The Fit LS Line command also draws a red indicator at the minimal SSE-value in the SSE bar graph.

- Is your line’s SSE noticeably above the minimal SSE-value? You can find the minimal SSE-value in the Analysis of Variance output on the far right --- find the Error row and the SS column.
- Report the value of the minimal SSE.

Now, use the Clear All Options command to remove the effects of all the commands used thus far.

Redo the steps of this lesson with another sample. You can obtain a new sample based on the same Population Characteristics by clicking the “Draw Additional Samples” button found near the bottom-left corner of the script window.
Regression Lesson 2: Comparing Sample to Population

Before beginning this lesson, click the "Restart Demo" button near the bottom-left of the script window (this gives you a fresh start.)

Set the Population Characteristics to the desired values and select Fit LS Line from the hot spot menu near top-center of the script window beside the scatterplot title “Sample Data.”

Remember: the LS line is our estimate of the population line. This means that the LS line’s slope is our estimate of the population slope entered on the left side of the window, and the LS line’s intercept is our estimate of the population intercept we specified on the left side of the window.

- How close does our estimated line come to the true population line?
- Compare the true population slope and intercept to the least-squares estimated slope and intercept.

Select the Show True Line command to have the population line overlayed onto the scatterplot with the LS line. This allows you to visually compare the population line to the LS line for this sample.

The correlation of our data is our estimate of the population correlation entered on the left side of the window. While we could calculate the sample’s correlation by taking the square root of the R Square value provided by the Fit LS Line command, it is simpler to use the Density Ellipse (Correlation) command. This command draws an ellipse oriented along the LS line and containing approximately 95% of the data points as well as providing the sample’s correlation to the far right of the window in a new output section titled “Correlation.”

- How does the sample’s correlation compare to the population correlation specified in the left side of the window?
- Compare the population intercept, slope, and correlation to the LS estimated intercept, slope, and correlation for several different samples (all drawn from the same population) by clicking the “Draw Additional Samples” button several times. Observe the changing scatterplot each time you click the button and record your estimated values below in the table.

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Slope</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population values</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample #1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample #2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample #3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample #4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Regression Lesson 3: How Do Sample Estimates Behave?

Reliable data-based decision making hinges on our ability to know that a statistical model based on a sample will be "reasonably close" to the true population model. Of course, in any real world scenario the true population model is unknown, so the comparison of statistical model to population model is impossible! Our best approach is to study scenarios in which the population model is known. In these special scenarios, we can obtain many samples and calculate many statistical models and see how close our models come to the population model.

How close do our models tend to come to the population model? How far off do our models tend to be from the population model? Some Statisticians focus on comparing various statistical models to true population models in a wide variety of scenarios where the population model is known. (Some Statisticians also focus on mathematically proving that certain statistical models come close to certain population models no matter what the specific population characteristics might be.)

In this lesson, we will take on the role of statistician and investigate how statistical models behave in a scenario where we choose the population characteristics. We will obtain many samples and compare each to the known population model.

Begin by setting the Population Characteristics to the desired values. Record your population model:

<table>
<thead>
<tr>
<th>Population Model</th>
<th>Intercept</th>
<th>Slope</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enter Your Values</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Next, click the grey triangle beneath the scatterplot (to the left of the title “Summary of All Samples”) to reveal the initial output for analyzing the results of repeated sampling. Now, slowly click the Draw Additional Samples button several times. Each click will produce a new sample of 25 data points. You can see the sampled values plotted in the scatterplot at the top of the script window. The script will also plot each sample’s LS line as a red line in the X-Y plot in the center of the script window. The population line is plotted in black for visual comparison to the LS lines.

Rather than considering just a few samples (and thus just a few statistical models), we will adjust the Demo Characteristics to rapidly obtain many samples for comparison to our population model. In the Demo Characteristics area of the script window, you can adjust two items: Sample Size and Number of Samples. Leave Sample Size set to the default value 25. Each sample will consist of 25 data points from the population. For Number of Samples, enter 10000.

There may be a slight pause while the script obtains 10,000 samples (each of 25 data points) and calculates the 10,000 LS lines associated with these samples!
Once the script finishes, the x-y plot will look something like this:

![Summary of All Samples](image)

Below the X-Y plot, the script summarizes the output of the 10,000+ fitted lines by reporting the minimum intercept and slope (of the 10,000+ LS lines), the maximum intercept and slope, the average of the 10,000+ intercepts and slopes, and the standard deviation of the intercepts and slopes. Record the results of your repeated sampling in this table:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Intercept</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- How well does using an LS line to estimate the population line work? Specifically, how close is the average of all our intercept estimates to the population intercept? How close is the average of all our slope estimates to the population slope?

Note: Below the Summary of Fits output, all 10,000+ intercepts and slopes are hidden in the Individual Fits output which can be revealed by clicking the grey triangle next to the “Individual Fits” title. Once revealed, you can right-click on those values to save them as a JMP data table for further study.

Click the red triangle hot spot menu to the left of the “Summary of All Samples” title to reveal the following menu.
The initial Summary of All Samples output is the first item in the menu --- the Fitted Lines output. Select the Intercepts command to view a histogram of the 10,000+ estimated intercept values.

- What do you notice about the shape of this histogram?
- Using the Avg and SD reported by the script, calculate the interval (Avg-2*SD, Avg+2*SD).
- Does this interval appear to contain most of the estimated intercept values?

Next, select the Slopes command to view a histogram of the 10,000+ estimated slope values.

- What do you notice about the shape of this histogram?
- Using the Avg and SD reported by the script, calculate the interval (Avg-2*SD, Avg+2*SD).
- Does this interval appear to contain most of the estimated slope values?

Now, select the Correlations command to view a histogram of the 10,000+ estimated correlation values.

- What do you notice about the shape of this histogram?
- How does the average of our 10,000+ correlation estimates compare to the population correlation we chose?
- How far off was our worst correlation estimate? Was this worst correlation estimate an underestimate or an overestimate of the true population correlation?

Finally, select the R Squares command to view a histogram of the 10,000+ R Square values.

- What do you notice about the shape of this histogram?
- How does the average of our 10,000+ R Square values compare to the square of the population correlation?
- How small was our worst R Square value? How large was our best R Square value?
Potential Script Uses for Additional Lessons

- Use Plot Intercepts by Slopes command to study the joint relationship between this pair of estimates (if we have overestimated the slope, we have probably underestimated the intercept, etc.)

- Focus on hypothesis testing results based on t-Ratios and critical values from the t-distribution for the intercept and the slope, use t-Ratios Intercept and t-Ratios Slope commands

- Focus on hypothesis testing results based on p-Values, use p-Values Intercept and p-Values Slope commands

- Use your own data file (select My Sample under Select Data Source) and then move one point around to create an outlier and watch the effects on the model fit as the point is moved. Next, move the point around to create a leverage point. Finally, move the point around to create an influence point. Save the altered data set by using the Save Sample to Data Table command for further analysis.

- Create your own data by clicking or click-dragging (select Create Your Own under Select Data Source). The possibilities are endless! For example, create a curved relationship between X and Y. How does the model fit? What do the residuals look like? Save your custom data set by using the Save Sample to Data Table command for further analysis.