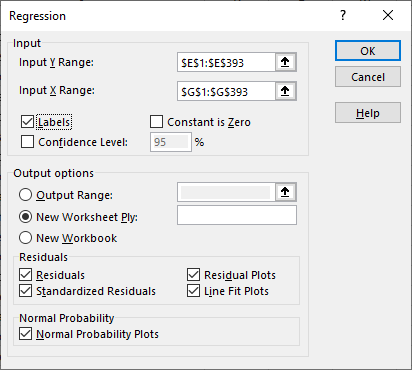
**Regression session: demo notes**

Regress horsepower on acceleration. Check on all the Residuals and Normal Probability options.



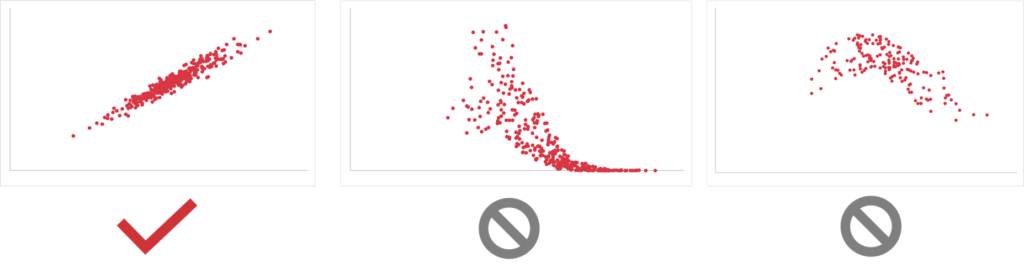
Let’s review the assumptions of regression, then check where we can find all this information:

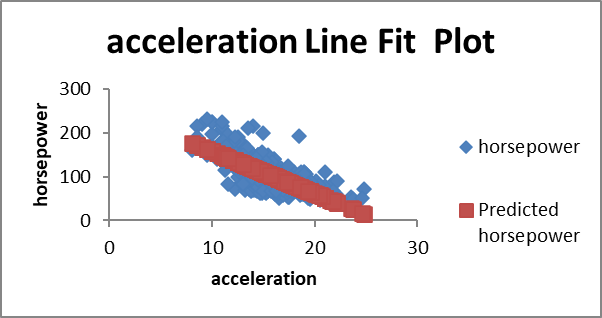
1. **Observations are independent**

This means that the measurement of each datapoint is not influenced by or related to other datapoints. It would be very difficult to test for this in Excel: our best bet is to catch any signs of this during the data collection process.

1. **The relationship between independent and dependent variables is linear**

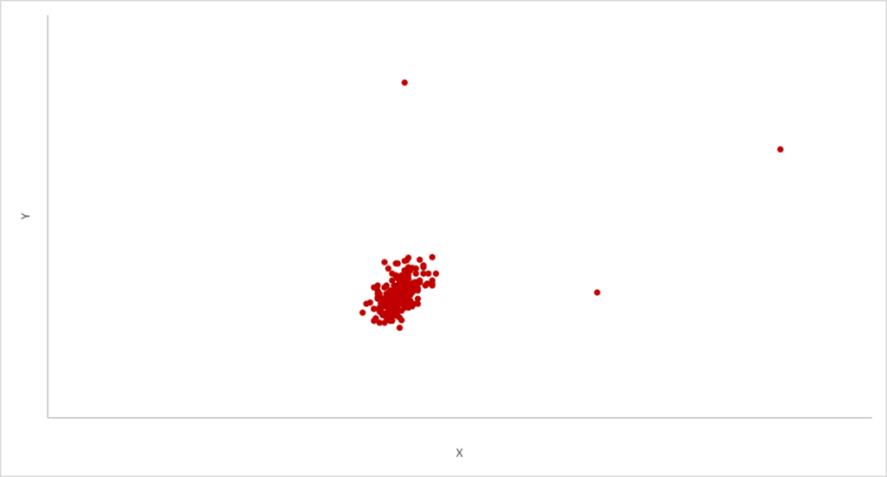
We will test for this graphically with the line fit. Is the relationship truly best describe as linear?



We can check this by adding a trendline to the original scatterplot made in Excel. Additionally, the ToolPak will provide a plot:   
  


1. **No influential cases**

These are datapoints that unduly influence the shape of the line that we use in the regression.

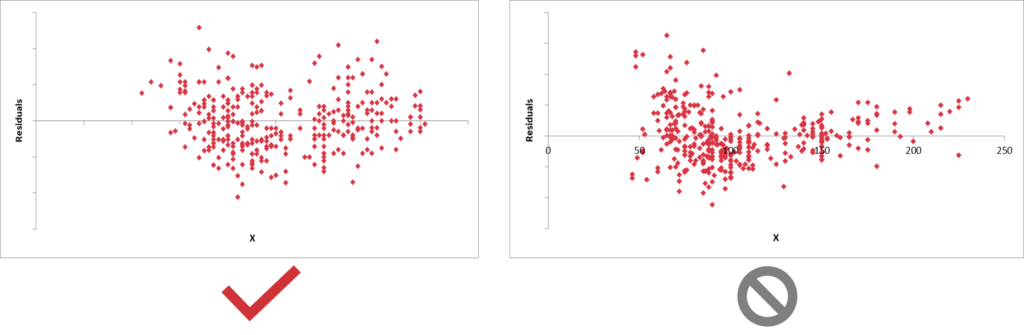


There are more sophisticated ways to check for influential cases, but a visual inspection of your scatterplot suffices for now.

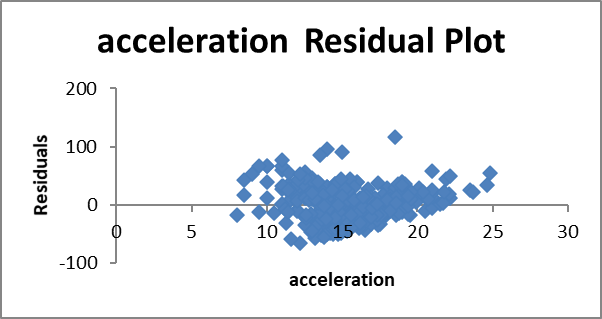
1. **Errors have equal variance.**

We can check for this by plotting the residuals across our X. We should see consistent values over time.

If our scatterplot of residuals resembles an “exploding cigar” and has a fanning effect in either direction, we have problems with this, what is called *heteroskedasticity.*



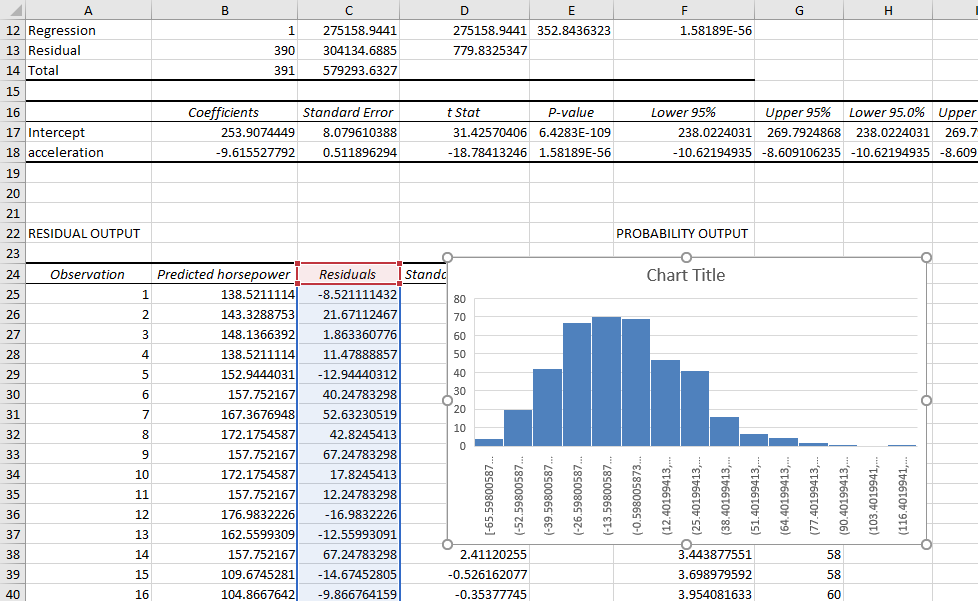
This will be found as part of the ToolPak output under “Residual plot”



1. **Errors are normally distributed**

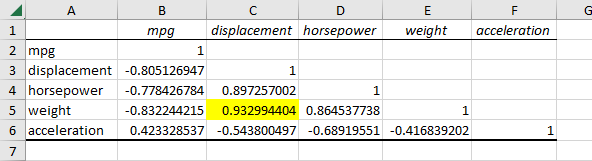
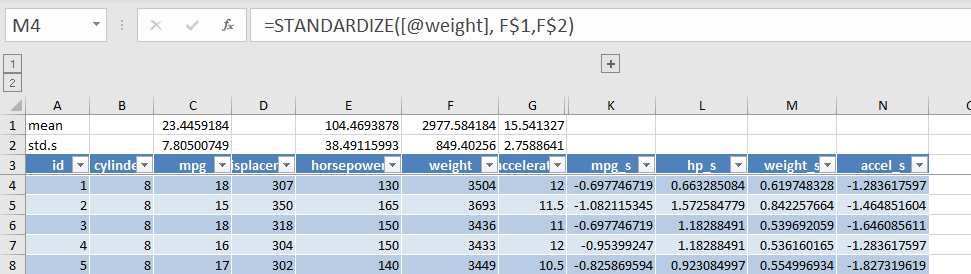
If we have extreme values in the residuals it may mean that our regression is more sensitive to some parts of our independent variable versus others.

For this one, we will plot the residual output as a histogram. We should see a “bell curve”-like shape in the histogram to meet the assumption.

We can test this one by plotting the Residuals as part of the output:   


**Multiple regression**

Continuing on with the dataset, we’ll analyze the influence of multiple independent variables on a vehicle’s mileage.

1. Put together a correlation matrix of the relevant variables. Which two are most highly correlated?   
   
   1. It’s weight and displacement – *why?*
2. Due to the relationship of weight and displacement, we are going to exclude it from our *multiple* regression.
3. We’re going to do one more thing before getting started too: *standardize* the variables. This will make interpreting our results much easier: all variables will have a mean of zero and standard deviation of one.
   1. Use the STANDARDIZE() function to do this, passing in each value along with that particular variable’s mean and standard deviation:   
      
4. Any resulting value greater than +/-3 tells us it’s an outlier as (as 95% of values should fall within 3 standard deviations of the mean). We can check that with conditional formatting.
5. Run the regression where the standardized mpg is the dependent variable and standardized horsepower, weight and acceleration are the independent. Keep the Residuals and Normal Probability options turned on.
6. Analyze the results. Do we meet our assumptions?
7. With all our variables standardized, the way to interpret the coefficients is that a one-unit increase in the standard deviation of the independent variable leads to an increase in Y of the dependent variable.
8. Are all of our independent variables significantly influencing the relationship?
9. Drop the one that is not and rerun the regression.

Drill

* Check for unsually high correlations
* Standardize the variables and check for potential outliers
* Run the regression, check assumptions and interpret the model.
* If any aren’t significant, drop the least-significant one first, then re-run the regression.