* [How to use this tool](http://simple.werf.org/simple/media/RELT/howTo.html)
* [Overview: What is "Remaining Effective Life"?](http://simple.werf.org/simple/media/RELT/index.html)
* [Step 1: Load Data](http://simple.werf.org/simple/media/RELT/stepOne.html)
* [Step 2: Determine Modified Design Life Factors](http://simple.werf.org/simple/media/RELT/stepTwo.html)
* [Step 3: Determine End of Asset Life](http://simple.werf.org/simple/media/RELT/stepThree.html)
* [Step 4: Determine Remaining Effective Life](http://simple.werf.org/simple/media/RELT/stepFour.html)
* [Step 5: Validate and Record](http://simple.werf.org/simple/media/RELT/stepFive.html)
* [Worked Examples](http://simple.werf.org/simple/media/RELT/examples.html)
* [WERF's SAM Project](http://simple.werf.org/simple/media/RELT/SAM.html)
* [Acknowledgements](http://simple.werf.org/simple/media/RELT/acknowledgements.html)

Overview – What is “Remaining Effective Life”?

Asset management can be described as a risk based approach to investment decision-making for tangible assets. Risk based approaches are built around the classic risk metric:

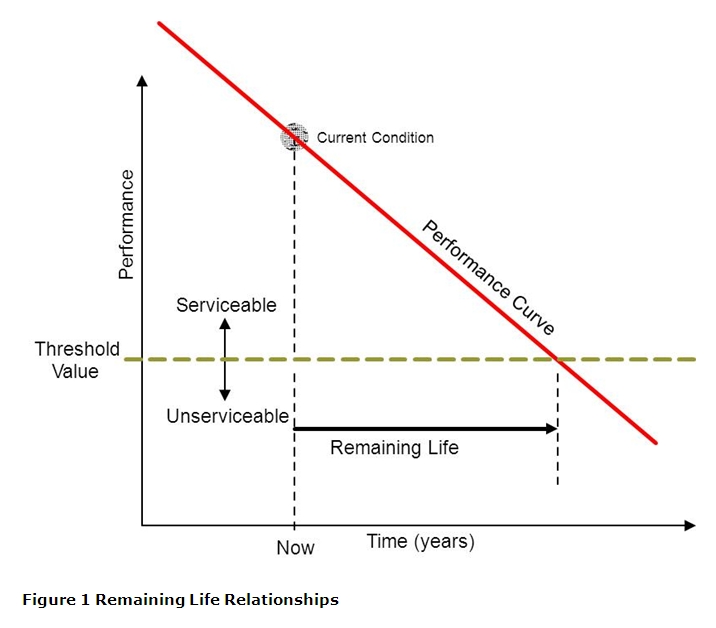
Risk = probability of failure x consequence of failure

To determine the probability of failure for a given asset, *we must determine as accurately as feasible when and how that asset’s useful life can be expected to end.* This determination, then, sets the stage for estimating the remaining useful life of the asset. Unfortunately, determining precisely when an asset’s useful life will end is no simple task. However, relatively simple techniques do exist which allow the asset manager to reasonably assign an “end of life” value to the asset.

This Tool facilitates that systematic determination using relatively simple techniques as to when and how an asset can be reasonably expected to fail. If we know when and how that failure is likely to occur and have a handle on how much lead time we need to effect a reinvestment strategy (major repair, refurbish, replace), we can more effectively plan for and manage the life cycle of the asset – thereby providing sustained performance to the customer/regulator/stakeholder base.

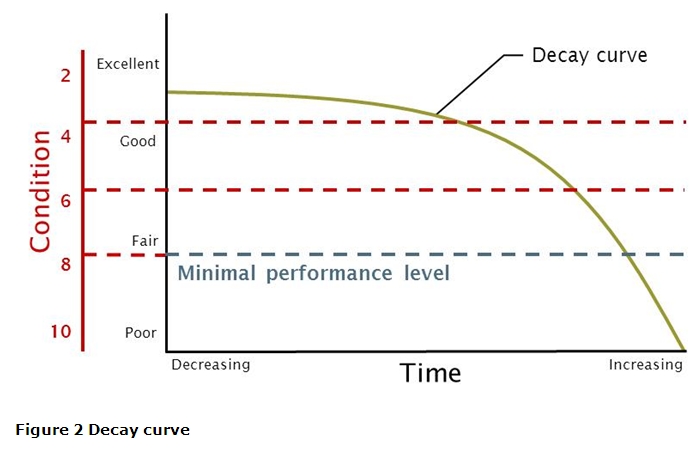
**What Is Meant By “Remaining Life” And Why Is It Important?**

The following graphic illustrates the basic concepts at work in determining the remaining life of an asset.



The graphic depicts the general recognition that the performance of an asset declines with usage over time. It also demonstrates the concept of service life, the threshold point below which the asset is not doing what is required (at which, from the user’s standpoint, the asset is not providing acceptable service – it is unserviceable), *even though it may still be technically functioning.*

This decline is depicted as a straight line in the graphic; in reality, most infrastructure and plant assets fail in a curved line rather than a straight line, with the rate of decline accelerating as the end of physical life is approached (see Figure 2). This curved line is often envisioned as the “deterioration curve” of the asset class. Ideally, a condition score would be linked to the curve such that the condition score would map the rate of deterioration. Unfortunately, while simple in concept, determining the exact nature of the performance curve for a given class of assets (much less a specific asset in its operating environment) is a substantial challenge, to say nothing of *precisely* relating the curve to condition scores. At best, current practice generally attempts to roughly relate condition scores to the state of deterioration.



This Tool incorporates several moderately simple techniques to establish the relationship among performance, time and optionally, condition. These simpler approaches are generally sufficient for – and critical to - strategic level asset planning. Unfortunately, the step between a simple, generalized approach and a more specific, sophisticated approach is substantial. However, a Tool, the End of Asset Life Reinvestment Tool is available in the WERF SIMPLE toolkit to assist those who require a more advanced approach.

**Different End of Asset Life Concepts**

Most of us tend to think of the physical collapse of the asset as its “end of life”. While this is certainly true, the end of physical life by itself only occasionally determines the optimal time for reinvestment in infrastructure assets. At least four other “end-of-life” determinations are recognized, three being equally if not more compelling for asset management purposes: service level, capacity and economic life. These other end-of-asset-life definitions are:

* **End of financial life** – when an asset is fully financially depreciated on the “books”
* **End of physical life** – when an asset is physically non-functioning (e.g., failed, collapsed, stopped working)
* **End of service level life** - when an asset can no longer do functionally what we/our customers/stakeholders require it to do because what is now required exceeds the designed functionality of the asset (does not include volumetric capacity, see next)o
* **End of capacity life** – when the volume of demand placed on an asset exceeds its design capability
* **End of economic life** – when an asset ceases to be the lowest cost alternative to satisfy a specified level of performance or service level. The end of economic life embraces such terms as “financial efficiency”, “business efficiency”, and “efficiency.

While the first (end of financial life) is highly relevant to financial accounting, it is not directly relevant to our purposes here; the reader is directed to financial accounting texts for further discussion. **The remaining four are termed Primary Failure Modes.**

The Remaining Effective Life Tool incorporates in a simple manner these distinctions among end of asset life forces to estimate remaining effective life by assisting the asset **manager to identify which of the four Primary Failure Modes will likely lead to failure first.**

**The Meaning and Key Role of “Remaining Effective Life”**

Because each Primary Failure Mode has a different perspective on when failure occurs for a given asset, *each of the Primary Failure Modes usually has a different end of asset life date associated with it.* While all four of the Primary Failure Modes are at work on an asset at all times, *only one is the most imminent in time.* That one is termed the *imminent Primary Failure Mode.* This is the mode that will most likely lead to actual failure; its associated estimated date of failure – and therefore its associated remaining life – is the timeframe within which the asset manager must work. This remaining life – the life associated with the imminent Primary failure Mode is termed the “remaining effective life”. **“Remaining effective life”, then, is the lowest expected life for a selected asset given its operating environment where that life is derived from a determination of the most imminent trigger among the four asset life triggers (service level life, capacity life, physical life, and economic life).**

It follows then, that the time between the point of time the analysis is being conducted and the end of effective life is the residual or “remaining useful life”. *This is the life period that the asset manager has left in the asset to wring performance out of and for which management strategies must be derived.* It is this residual life, of course, when combined with the lead time to take action that guides the timing of the investment decision.

**Age Based versus Condition Based Options**

Two conceptual alternatives for determining end of asset life are available to the practitioner, one based on age, the other on condition. Of the two, condition is generally far superior to an age based approach and should be used where reliable and timely condition based information is available (a good AM program embeds periodic and quality checked condition data as a core business process; see the Condition/Performance Scoring Tool for more elaboration).

The systematic but moderately simple approximation of the remaining effective life is the focus of the Remaining Effective Life Tool.

© 2011 WERF, Water Research Foundation, GWRC, GHD Consulting Inc.

1. The Acme Chain Company claims that their chains have an average breaking strength of 20,000 pounds, with a standard deviation of 1750 pounds. Suppose a customer tests 14 randomly-selected chains. What is the probability that the average breaking strength in the test will be no more than 19,800 pounds?   
     
   **Solution:**  
     
   One strategy would be a two-step approach:
   * Compute a t score, assuming that the mean of the sample test is 19,800 pounds.
   * Determine the cumulative probability for that t score.

We will follow that strategy here. First, we compute the t score:

t = [ x - μ ] / [ s / sqrt( n ) ]   
t = (19,800 - 20,000) / [ 1750 / sqrt(14) ]   
t = ( -200 ) / [ (1750) / (3.74166) ] = ( -200 ) / (467.707) = -0.4276

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, n is the sample size, and t is the t score.   
  
Now, we can determine the cumulative probability for the t score. We know the following:

* + The t score is equal to -0.4276.
  + The number of degrees of freedom is equal to 13. (In situations like this, the number of degrees of freedom is equal to number of observations minus 1. Hence, the number of degrees of freedom is equal to 14 - 1 or 13.)

Now, we are ready to use the [T Distribution Calculator](http://stattrek.com/online-calculator/t-distribution.aspx#TopPage). Since we have already computed the t score, we select "t score" from the drop-down box. Then, we enter the t score (-0.4276) and the degrees of freedom (13) into the calculator, and hit the Calculate button. The calculator reports that the cumulative probability is 0.338. Therefore, there is a 33.8% chance that the average breaking strength in the test will be no more than 19,800 pounds.   
  
Note: The strategy that we used required us to first compute a t score, and then use the T Distribution Calculator to find the cumulative probability. An alternative strategy, which does not require us to compute a t score, would be to use the calculator in the "Sample mean" mode. That strategy may be a little bit easier. It is illustrated in the next example.

1. Let's look again at the problem that we addressed above in Example 1. This time, we will illustrate a different, easier strategy to solve the problem.   
     
   Here, once again, is the problem: The Acme Chain Company claims that their chains have an average breaking strength of 20,000 pounds, with a standard deviation of 1750 pounds. Suppose a customer tests 14 randomly-selected chains. What is the probability that the average breaking strength in the test will be no more than 19,800 pounds?   
     
   **Solution:**  
     
   We know the following:
   * The population mean is 20,000.
   * The standard deviation is 1750.
   * The sample mean, for which we want to find a cumulative probability, is 19,800.
   * The number of degrees of freedom is 13. (In situations like this, the number of degrees of freedom is equal to number of observations minus 1. Hence, the number of degrees of freedom is equal to 14 - 1 or 13.)

First, we select "Sample mean" from the dropdown box, in the [T Distribution Calculator](http://stattrek.com/online-calculator/t-distribution.aspx#TopPage). Then, we plug our known input (degrees of freedom, sample mean, standard deviation, and population mean) into the [T Distribution Calculator](http://stattrek.com/online-calculator/t-distribution.aspx#TopPage) and hit the Calculate button. The calculator reports that the cumulative probability is 0.338. Thus, there is a 33.8% probability that an Acme chain will snap under 19,800 pounds of stress.   
  
Note: This is the same answer that we found in Example 1. However, the approach that we followed in this example may be a little bit easier than the approach that we used in the previous example, since this approach does not require us to compute a t score.

1. The school board administered an IQ test to 25 randomly selected teachers. They found that the average IQ score was 115 with a standard deviation of 11. Assume that the cumulative probability is 0.90. What population mean would have produced this sample result?   
     
   Note: In this situation, a cumulative probability of 0.90 suggests that 90% of the random samples drawn from the teacher population will have an average IQ of 115 or less. This problem asks you to find the true population IQ for which this would be true.   
     
   **Solution:**  
     
   We know the following:
   * The cumulative probability is 0.90.
   * The standard deviation is 11.
   * The sample mean is 115.
   * The number of degrees of freedom is 24. (In situations like this, the number of degrees of freedom is equal to number of observations minus 1. Hence, the number of degrees of freedom is equal to 25 - 1 or 24.)

First, we select "Sample mean" from the dropdown box, in the [T Distribution Calculator](http://stattrek.com/online-calculator/t-distribution.aspx#TopPage). Then, we plug the known inputs (cumulative probability, standard deviation, sample mean, and degrees of freedom) into the calculator and hit the Calculate button. The calculator reports that the population mean is 112.1.   
  
Here is what this means. Suppose we randomly sampled every possible combination of 25 teachers. If the true population mean were 112.1, we would expect 90% of our samples to have a sample mean of 115 or less.

---------------------------------------------------------------------------------------------------------------------------------------------

The t distribution calculator accepts two kinds of [random variables](http://stattrek.com/Help/Glossary.aspx?Target=Random_variable) as input: a [t score](http://stattrek.com/Help/Glossary.aspx?Target=T_score) or a sample mean. Choose the option that is easiest. Here are some things to consider.

* If you choose to work with t scores, you may need to transform your raw data into a t score. You can accomplish this transformation by using the following equation:

t = [ x - μ ] / [ s / sqrt( n ) ]

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, n is the sample size, and t is the t score.

In regression analysis, you'd like your regression model to have significant variables and to produce a high R-squared value. This low P value / high R2 combination indicates that changes in the predictors are related to changes in the response variable and that your model explains a lot of the response variability.

This combination seems to go together naturally. But what if your regression model has significant variables but explains little of the variability? It has low P values and a low R-squared.

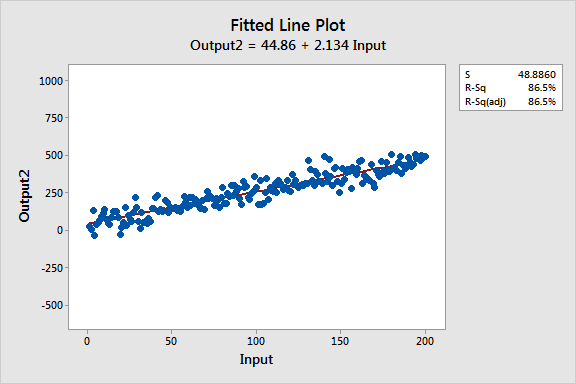
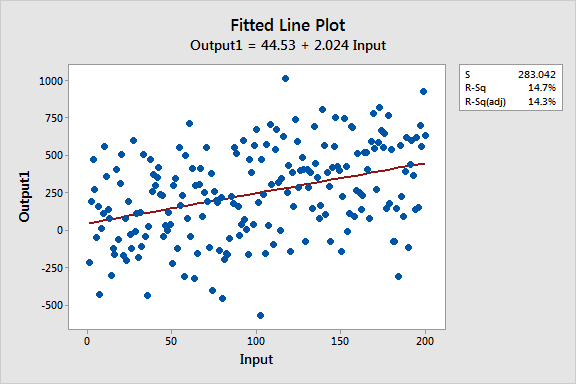
At first glance, this combination doesn’t make sense. Are the significant predictors still meaningful? Let’s look into this!

## Comparing Regression Models with Low and High R-squared Values

It’s difficult to understand this situation using numbers alone. Research shows that [graphs are essential](http://blog.minitab.com/blog/adventures-in-statistics/applied-regression-analysis-how-to-present-and-use-the-results-to-avoid-costly-mistakes-part-1) to correctly interpret regression analysis results. Comprehension is easier when you can see what is happening!

With this in mind, I'll use fitted line plots. However, a 2D fitted line plot can only display the results from simple regression, which has one [predictor variable](http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/regression-models/what-are-response-and-predictor-variables/) and the [response](http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/regression-models/what-are-response-and-predictor-variables/). The concepts hold true for multiple linear regression, but I can’t graph the higher dimensions that are required.

These fitted line plots display two regression models that have nearly identical regression equations, but the top model has a low [R-squared](http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit) value while the other one is high. I’ve kept the graph scales constant for easier comparison. Here are the [data](https://app.compendium.com/api/post_attachments/cca92f43-1d94-48e4-aac7-c33e0779b7e1/view) for these examples.



# Regression Analysis

The data used here is from the 2004 Olympic Games. We are going to see if there is a correlation between the weights that a competitive lifter can lift in the snatch event and what that same competitor can lift in the clean and jerk event.

We will use a response variable of "clean" and a predictor variable of "snatch".

## Data

The heaviest weights (in kg) that men who weigh more than 105 kg were able to lift are given in the table.

### Data Dictionary

Age

The age the competitor will be on their birthday in 2004.

Body

The weight (kg) of the competitor

Snatch

The maximum weight (kg) lifted during the three attempts at a snatch lift

Clean

The maximum weight (kg) lifted during the three attempts at a clean and jerk lift

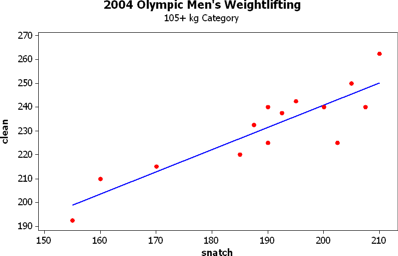
Total

The total weight (kg) lifted by the competitor

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age** | **Body** | **Snatch** | **Clean** | **Total** |
| 26 | 163.0 | 210.0 | 262.5 | 472.5 |
| 30 | 140.7 | 205.0 | 250.0 | 455.0 |
| 22 | 161.3 | 207.5 | 240.0 | 447.5 |
| 27 | 118.4 | 200.0 | 240.0 | 440.0 |
| 23 | 125.1 | 195.0 | 242.5 | 437.5 |
| 31 | 140.4 | 190.0 | 240.0 | 430.0 |
| 32 | 158.9 | 192.5 | 237.5 | 430.0 |
| 22 | 136.9 | 202.5 | 225.0 | 427.5 |
| 32 | 145.3 | 187.5 | 232.5 | 420.0 |
| 27 | 124.3 | 190.0 | 225.0 | 415.0 |
| 20 | 142.7 | 185.0 | 220.0 | 405.0 |
| 29 | 127.7 | 170.0 | 215.0 | 385.0 |
| 23 | 134.3 | 160.0 | 210.0 | 370.0 |
| 18 | 137.7 | 155.0 | 192.5 | 347.5 |

## Correlation

The first rule in data analysis is to make a picture. In this case, a scatter plot is appropriate.



You can see from the data that there appears to be a linear correlation between the clean & jerk and the snatch weights for the competitors, so let's move on to finding the correlation coefficient.

Here is the Minitab output.

Pearson correlation of snatch and clean = 0.888  
P-Value = 0.000

The Pearson's correlation coefficient is r = 0.888. Remember that number, we'll come back to it in a moment.

For now, the p-value is 0.000. Every time you have a p-value, you have a hypothesis test, and every time you have a hypothesis test, you have a null hypothesis. The null hypothesis here is H0: ρ = 0, that is, that there is no significant linear correlation.

The p-value is the chance of obtaining the results we obtained if the null hypothesis is true and so in this case we'll reject our null hypothesis of no linear correlation and say that there is significant positive linear correlation between the variables.

## Regression Analysis

Let's start off with the descriptive statistics for the two variables. Remember, our predictor (x) variable is snatch and our response variable (y) is clean.

Variable N Mean SE Mean StDev Minimum Q1 Median Q3 Maximum  
snatch 14 189.29 4.55 17.02 155.00 181.25 191.25 203.13 210.00  
clean 14 230.89 4.77 17.86 192.50 218.75 235.00 240.63 262.50

The following explanation assumes the regression equation is y = b0 + b1x.

The formula for the slope is b1 = r (sy / sx).

For our data, that would be b1 = 0.888 ( 17.86 / 17.02 ) = 0.932.

Since the best fit line always passes through the centroid of the data, the y-intercept, b0, is found by substituting the slope just found and the means of the two variables into the regression equation and solving for b0.

230.89 = b0 + 0.932 ( 189.29 )

Solving for b0 gives the constant of 54.47.

So we can write the regression equation as clean = 54.47 + 0.932 snatch.

Here is the regression analysis from Minitab.

The regression equation is  
clean = 54.6 + 0.931 snatch

Predictor Coef SE Coef T P  
Constant 54.61 26.47 2.06 0.061  
snatch 0.9313 0.1393 6.69 0.000

S = 8.55032 R-Sq = 78.8% R-Sq(adj) = 77.1%

Analysis of Variance

Source DF SS MS F P  
Regression 1 3267.8 3267.8 44.70 0.000  
Residual Error 12 877.3 73.1  
Total 13 4145.1

Notice that the regression equation we came up with is pretty close to what Minitab calculated. Ours is off a little because we used rounded values in calculations, so we'll go with Minitab's output from here on, but that's the method you would go through to find the equation of the regression equation by hand.

Let's go through and look at this information and how it ties into the ANOVA table.

### Table of Coefficients

A quick note about the table of coefficients, even though that's not what we're really interested in here.

Predictor Coef SE Coef T P  
Constant 54.61 26.47 2.06 0.061  
snatch 0.9313 0.1393 6.69 0.000

The "Coef" column contains the coefficients from the regression equation. The 54.61 is the constant (displayed as 54.6 in the previous output) and the coefficient on snatch of 0.9313 is the slope of the line.

The "SE Coef" stands for the standard error of the coefficient and we don't really need to concern ourselves with formulas for it, but it is useful in constructing confidence intervals and performing hypothesis tests.

Speaking of hypothesis tests, the T is a test statistic with a student's t distribution and the P is the p-value associated with that test statistic.

Every hypothesis test has a null hypothesis and there are two of them here since we have two hypothesis tests.

The model for the regression equation is y = β0 + β1 x + ε where β0 is the population parameter for the constant and the β1 is the population parameter for the slope and ε is the residual or error term. The b0 and b1 are just estimates for β0 and β1.

The null hypothesis for the constant row is that the constant is zero, that is H0: β0 = 0 and the null hypothesis for the snatch row is that the coefficient is zero, that is H0: β1 = 0. If the coefficient is zero, then the variable (or constant) doesn't appear in the model since it is multiplied by zero. If it doesn't appear in the model, then you get a horizontal line at the mean of the y variable. That's the case of no significant linear correlation.

So, another way of writing the null hypothesis is that there is no significant linear correlation. Notice that's the same thing we tested when we looked at the p-value from the correlation section. For that reason, the p-value from the correlation coefficient results and the p-value from the predictor variable row of the table of coefficients will be the same -- they test the same hypothesis.

The t test statistic is t = ( observed - expected ) / (standard error ). Since the expected value for the coefficient is 0 (remember that all hypothesis testing is done under the assumption that the null hypothesis is true and the null hypothesis is that the β is 0), the test statistic is simple found by dividing the coefficient by the standard error of the coefficient.

Go ahead, test it. 54.61 / 26.47 = 2.06 and 0.9313 / 0.1393 = 6.69. Pretty cool, huh?

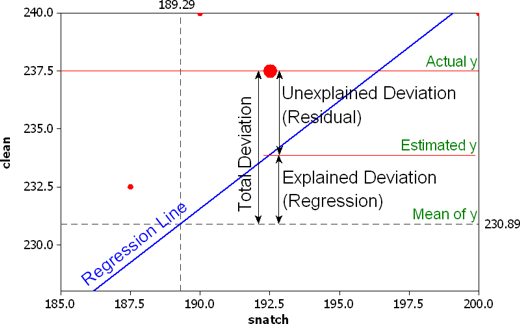
There is one kicker here, though. The t distribution has df = n-2. That's because there are two parameters we're estimating, the slope and the y-intercept.

One further note, even though the constant may not be significantly different from 0 (as in this case with a p-value of 0.061, we marginally retain the null hypothesis that β0 = 0), we usually don't throw it out in elementary statistics because it messes up all the really cool formulas we have if we do.

Okay, I'm done with the quick note about the table of coefficients. On to the good stuff, the ANOVA.

### Sources of Variation

The sources of variation when performing regression are usually called Regression and Residual. Regression is the part that can be explained by the regression equation and the Residual is the part that is left unexplained by the regression equation. This is what we've been calling the Error throughout this discussion on ANOVA, so keep that in mind. Notice that Minitab even calls it "Residual Error" just to get the best of both worlds in there.

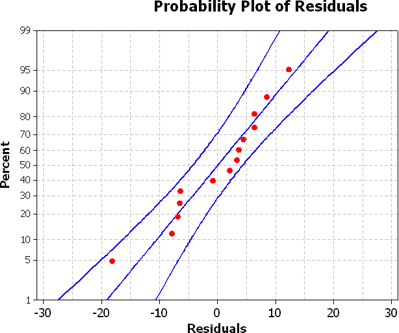
The picture to right may help explain all this. It takes one data point, for Shane Hamman of the United States who snatched 192.5 kg and lifted 237.5 kg in the clean and jerk.

The centroid (center of the data) is the intersection of the two dashed lines.

The blue line is the regression line, which gives us predicted values for y.

The total deviation from the mean is the difference between the actual y value and the mean y value. In this case, that difference is 237.5 - 230.89 = 6.61. Part of that 6.61 can be explained by the regression equation. The estimated value for y (found by substituting 192.5 for the snatch variable into the regression equation) is 233.89. So the amount of the deviation that can be explained is the estimated value of 233.89 minus the mean of 230.89 or 3. The residual is the difference that remains, the difference between the actual y value of 237.5 and the estimated y value of 233.89; that difference is 3.61.

### Residuals

Now is as good of time as any to talk about the residuals. Here are the residuals for all 14 weight lifters.

12.3164, 4.4727, -7.8555, -0.8709, 6.2855, 8.4419, 3.6137, -18.1991, 3.2701, -6.5581, -6.9017, 2.0675, 6.3803, -6.4633

The residuals are supposed to be normally distributed. As you can see from the normal probability plot, the residuals do appear to have a normal distribution.

If you simply add the residuals together, then you get 0 (possibly with roundoff error). That's not a coincidence, it always happens. That's why the sum of the residuals is absolutely useless as anything except for a check to make sure we're doing things correctly.

So, what do we do? We square each value and then add them up.

Wait a minute, what are we doing? We're finding the sum of the squares of the deviations ... Hey! That's a variation.

## Analysis of Variance

Yep, that's right, we're finding variations, which is what goes in the SS column of the ANOVA table. There are two sources of variation, that part that can be explained by the regression equation and the part that can't be explained by the regression equation.

Here's how the breakdown works for the ANOVA table.

|  |  |  |
| --- | --- | --- |
| **Source** | **SS** | **df** |
| Regression (Explained) | SS(Regression)Sum the squares of the explained deviations | # of parameters - 1 always 1 for simple regression |
| Residual / Error (Unexplained) | SS(Residual)Sum the squares of the unexplained deviations | sample size - # of parameters n - 2 for simple regression |
| Total | SS(Total)Sum the squares of the deviations from the mean | sample size - 1 n - 1 |

Some notes on the degrees of freedom.

The df(Reg) is one less than the number of parameters being estimated. For simple regression, there are two parameters, the constant β0 and the slope β1, so there are always 2-1 = 1 df for the regression source.

The df(Res) is the sample size minus the number of parameters being estimated. For simple regression, there are two parameters so there are n-2 df for the residual (error) source.

The df(Total) is one less than the sample size, so there are n-1 df for the total df.

The F test statistic has df(Regression) = 1 numerator degrees of freedom and df(Residual) = n - 2 denominator degrees of freedom. The p-value is the area to the right of the test statistic.

Since there is a test statistic and p-value, there must be a hypothesis test. The null hypothesis is that the slope is zero, H0: β1 = 0. If that's true, then there is no linear correlation. That's the same thing we tested with the correlation coefficient and also with the table of coefficients, so it's not surprising that once again, we get the same p-value. In this particular problem, that's not so obvious since the p-value is 0.000 for all of them, just take my word for it :)

Do you remember when we said that the MS(Total) was the value of s2, the sample variance for the response variable? For our data, the MS(Total), which doesn't appear in the ANOVA table, is SS(Total) / df(Total) = 4145.1 / 13 = 318.85.

When you take the standard deviation of the response variable (clean) and square it, you get s2 = 17.862 = 318.98. The slight difference is again due to rounding errors.

## Regression Revisited

The Coefficient of Determination is the percent of variation that can be explained by the regression equation. It's abbreviated r2 and is the explained variation divided by the total variation.

The variations are sum of squares, so the explained variation is SS(Regression) and the total variation is SS(Total).

Coefficient of Determination = r2 = SS(Regression) / SS(Total)

There is another formula that returns the same results and it may be confusing for now (until we visit multiple regression), but it's

Coefficient of Determination = r2 = ( SS(Total) - SS(Residual) ) / SS(Total)

For our data, the coefficient of determination is 3267.8 / 4145.1 = 0.788. Notice this is the value for R-Sq given in the Minitab output between the table of coefficients and the Analysis of Variance table.

S = 8.55032 R-Sq = 78.8% R-Sq(adj) = 77.1%

But why is it called r2? Well, our value for the correlation coefficient was r = 0.888 and 0.8882 is 0.788544 = 78.8%. Does the coolness ever end?

Remember how I mentioned the multiple regression coming up? The formula for the Adjusted R2 is the same as the second one for r2 except you use the variances (MS) instead of the variations (SS).

Adjusted R2 = ( MS(Total) - MS(Residual) ) / MS(Total)

Adjusted R2 = ( 318.85 - 73.1 ) / 318.85 = 0.771 = 77.1%

One caveat, though. The S = 8.55032 is not the same as the sample standard deviation of the response variable. It is more appropriately called se, known as the standard error of the estimate or residual standard error. That value of se = 8.55032 is the square root of the MS(error). The square root of 73.1 is 8.55.

## Summary of ANOVA

Wow! There is a lot of good information there, but the only real difference in how the ANOVA table works in how the sum of squares and degrees of freedom are computed. We'll leave the sum of squares to technology, so all we really need to worry about is how to find the degrees of freedom.

df(Regression) = # of parameters being estimated - 1 = 2 - 1 = 1   
df(Residual) = sample size - number of parameters = n - 2

Last modified June 6, 2006 2:21 PM

Return to [ICTCM 2004 Short Course](https://people.richland.edu/james/ictcm/2004/) page

Return to [James Jones homepage](https://people.richland.edu/james/)