Simple Predictive Analytics

Curtis Seare

Copyright: Vault Analytics
July 2010
Contents

Section I: Background Information
Why use Predictive Analytics?
How to use this Guide
Analysis Fundamentals
Method for Creating Predictive Models

Section II: Predictive Models
How to Choose an Appropriate Model
Regression
  Correlation
  Linear Regression
  Multivariate
  Exponential Regression
  Logarithmic Regression
  Polynomial Regression
  Time Series
Logistic Regression
ANOVA
  t-Test
  1 Way ANOVA
  2 Way ANOVA
Chi-Square
Why Use Predictive Analytics?

Strategy Development

Predictive analytics is an indispensible tool for strategy development. In any good strategy, there are three elements: a set of assumptions, a set of actions that need to be taken, and a set of desired outcomes. The outcomes are achieved as a result of the actions being executed and the assumptions being true. In essence, we say: Assuming A is true, if we do B we will achieve C.

Predictive analytics helps out with all three of the steps considered above. Statistical tests tell us if our assumptions are correct and predictive models fit to historic data help us quantify what the result will be with a given set of inputs. The value of having a statistically valid predictive model in strategy development is that you can know what you need to do to produce the best results.

Further, simple predictive models are not above most people’s abilities to produce. It requires some understanding of statistics, as well as various predictive modeling techniques. This guide provides you with both the statistical basics you need to know, as well as a process that can be followed to choose and implement an appropriate predictive modeling technique for your given situation. It stays admittedly away from many technicalities, so that focus can be placed just on the information that is necessary to get a predictive model built and in working order.

Example Business Questions Answered

The reason we do predictive analytics is to solve business problems. Accordingly, here are some example business questions that can be answered with each of the predictive models described in this guide.

Regression

Correlation

- From a large database of demographical data, which factors are associated with response rates?
- From all the metrics on our website, which ones are associated with purchases on the site?
- From all the medical data collected in a patient sample, which factors are associated with blood pressure?

Linear Regression

- If we send out ‘x’ number of mail pieces to a target customer segment, how many sales will we get in return?
- If we sell ‘x’ number of product A, how much can we expect to sell of product B?
- If ‘x’ number of people sign up for this promotional program, how many extra sales can we expect?

Multivariate Linear Regression
• How many subscriptions can we expect to get by spending specific amounts of time in various social media marketing channels?
• What is the optimal word count, number of graphics, and number of topics covered in our weekly newsletter in order to get the most click-throughs?

**Exponential Regression**

• Model what happens in a word of mouth marketing campaign – builds slowly at first, but after a certain point it rises exponentially
• Model a learning curve – at first, it takes a long time to perform a specific task on a new piece of software, but the more you do it, the time lowers to a constant level

**Logarithmic Regression**

• Model the percentage of total expected calls that have come in after a mail campaign is executed - this percentage will grow quickly but taper off as time goes on

**Polynomial Regression**

• What is the appropriate amount of customer contacts per month to maximize sales?
• Where should we put the price in order to sell the maximum amount of a given product?

**Time Series**

• How many sales are we likely to have in the next year?
• How much do the specific months of the year affect our close rates?

**Logistic Regression**

• What is the probability that someone of a given age will respond to our marketing campaign?
• What is the probability that someone will purchase this product given a certain price?

**ANOVA**

**t-Test**

• Does gender affect the sales rate of our products?
• Does this medical treatment affect the blood pressure of our patients?
• Does this training course increase the efficiency of our staff?

**1 Way ANOVA**

• Compare the response rates of various customer segments to a specific marketing campaign, and then allocate more resources to the segment with the higher response
• Compare response rates of different variations of a mailer piece, and use the mailer that has the greatest response rate
• Compare sales conversion rates according to traffic referral information from your web analytics, and then focus on getting more traffic from the highest converting source

**2 Way ANOVA**
• What is the best combination of marketing channel and product offering to get the highest sales rates?

**Chi-Square**

• Out of all of our customer segments, which ones are the most likely to buy our products?
• Do our former assumptions about our customer segments still hold true?
How to use this Guide

This guide is designed to be as simple as possible, giving you the critical information you will need to create, use, and validate various simple predictive models.

First, some analysis fundamentals are provided. These are the basics of doing good and accurate analysis, and it will be important to keep these principles in mind as you create predictive models.

Second, a process is provided that will allow you to follow some easy, predefined steps to creating your own predictive models. This is a ‘big-picture’ process flow meant to give you a basic procedure to follow no matter what type of predictive model you need to create.

Last, the guide gives you an in-depth look into various predictive modeling techniques, organized according to the type of data you have and the type of question you’re trying to answer. This section makes up the bulk of the book, and each of the models is explained by telling you what the predictive model looks like, what it can be used for, the assumptions necessary to use the model, a process to follow to create it (including step by step instructions if you are using Excel 2007), an explanation of some common errors to watch for, a section on analyzing your results, and a section listing a nonparametric counterpart test to use if the conditions for normality are not met in the data.

Preparing to Use Excel 2007

Excel 2007 is used in the specific ‘How To’ sections, as it is the most ubiquitous piece of software with statistical capabilities. Excel is admittedly limited in these statistical capabilities, but affords the basic functionality necessary to create and validate simple predictive models.

Two add-ins are used throughout this guide to aid in doing the necessary statistical tests for each predictive model. They are the Data Analysis Toolpak and the Solver Add-in. These add-ins come with Excel 2007, but you must activate them before you can use them.

To activate both of them, simply go to the home menu and click ‘Excel Options’
Next, click on the ‘Add-Ins’ section on the left side, and click ‘Go. . .’

Lastly, check the ‘Analysis ToolPak’ and ‘Solver Add-in’ check boxes, and click ‘OK’
Analysis Fundamentals

Although an entire book could be written about the fundamentals of good analysis, here we will cover just two fundamentals that are the most critical. These two basics are seeing the data in context and segmentation.

Seeing the Data in Context

Understanding what the data are telling you within the context of the business situation being analyzed is extremely important. This will help you avoid making faulty conclusions and keep your analysis appropriate for the business question being answered. The best way to learn this fundamental is to see it in action, so we will take an example.

We will look at a type of direct mail campaign analysis. We want to know how many calls are expected to come into our call center after we execute the campaign. First, we take some historical data showing us the percentage of total calls coming in according to the number of days after starting a mail campaign, shown below.

<table>
<thead>
<tr>
<th>Days After Mailing</th>
<th>% Total Incoming Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>3</td>
<td>4%</td>
</tr>
<tr>
<td>4</td>
<td>12%</td>
</tr>
<tr>
<td>5</td>
<td>19%</td>
</tr>
<tr>
<td>6</td>
<td>26%</td>
</tr>
<tr>
<td>7</td>
<td>31%</td>
</tr>
<tr>
<td>8</td>
<td>36%</td>
</tr>
<tr>
<td>9</td>
<td>41%</td>
</tr>
</tbody>
</table>

After creating a scatter plot of the data, we try to fit a logarithmic regression line as a model, shown seen below.

\[ y = 0.1607 \ln(x) + 0.1855 \]

\[ R^2 = 0.8365 \]

Even though the \( R^2 \) tells us that the fit is good, the model may not be the best way to explain this data when the context and purpose of this analysis are considered. We want the model to be able to predict what percentage of total calls will come in from a mailing campaign so we can staff the call center efficiently.
center. If I were to use the line above as the model, I would be predicting low values for incoming calls between about day 20 and 100, and high values thereafter. Because of this error, we would not be staffing the call center correctly.

To create a better model, I would consider the fact that, in this context, it is not necessary to fit a trend model to the entire data set. Consider the following model, which can be used to predict the percentage of total calls coming in between days 4 and 35 after the mailing campaign:

![Days 4 - 35](image)

You will notice that this trend model does not contain the same high and low errors as the previous model did. Further, upon doing some calculations on the data in the spreadsheet, we know that anything before day 4 makes up for just 8% of all calls, and anything after day 35 makes up for just 15% of all calls. I have highlighted with a model the time period of the biggest growth to the call percentage, while summarizing the remaining percentages on either side. This will give just the right amount of information needed to staff the call center, while minimizing errors I would have made trying to fit a single trend model to the data.

The point here is to look at the data in the context of the purpose of the analysis. What are you going to use the predictive model for? Is it necessary to fit a model to the entire data set? How exact do you need to be with the prediction? What is the most important part of the data set to model? These and other questions are important to consider when performing analysis.

**Segmentation**

The second fundamental of analysis is the practice of segmenting the data. As with seeing the data in context, this is best described with an example. Consider the analysis presented below, which shows a linear regression model to predict how much someone will likely donate to your cause according to their age.
The fit of the model is extremely weak, and there seems to be no relationship between donation and age. However, this data was taken and aggregated from two different cities, Boston and New York. If we separate out the data according to those two cities (otherwise known as segmenting by them), we get the following when we run a regression analysis:

By segmenting the data first, we notice that there is, in fact, a relationship between donation and age, but that relationship differs depending on what city you are in.
Method for Creating Predictive Models

Process Outlined
1. Choose a predictive model according to the business question
2. Check to see if all the conditions for the model are met
3. Carry out the analysis
4. Check for statistical significance and fit
5. Validate the predictive model
6. Refine the predictive model

1. Choose a Predictive Model

Once you have a business question in mind, you must look at what kinds of variables you will be working with in order to answer this question. The models in this guide will help you with the following four situations, which are fairly comprehensive:

- Predict a quantitative outcome with quantitative explanatory variables
- Predict a 'yes' or 'no' outcome with quantitative explanatory variables
- Predict a quantitative outcome according to categorical explanatory variables
- Predict a categorical outcome according to categorical explanatory variables

The difference between quantitative and categorical variables is described below:

Quantitative Variable

Anything that can be measured or counted is a quantitative variable. This includes things such as heights, distances, weights, # of items, etc. Because these variables are numerical, you can perform mathematical operations on them, such as addition, multiplication, division, etc. Below is an example of what quantitative variables look like in a spreadsheet:

<table>
<thead>
<tr>
<th>Sales</th>
<th>Costs</th>
<th>Distance (mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>457</td>
<td>$24.00</td>
<td>5.4</td>
</tr>
<tr>
<td>514</td>
<td>$57.00</td>
<td>2.4</td>
</tr>
<tr>
<td>234</td>
<td>$84.00</td>
<td>2.1</td>
</tr>
<tr>
<td>541</td>
<td>$54.00</td>
<td>4.6</td>
</tr>
<tr>
<td>465</td>
<td>$74.00</td>
<td>6.2</td>
</tr>
<tr>
<td>451</td>
<td>$71.00</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Above, the variables are ‘Sales’, ‘Costs’, and ‘Distance’, each followed by the data points that make them up.

Categorical (or Qualitative) Variable
These variables use categories to classify individuals. This can include things such as gender (male, female), industry (chemical, finance, construction, etc.), zip code, and others. Below is an example of what categorical variables look like in a spreadsheet:

<table>
<thead>
<tr>
<th>Channel</th>
<th>Product</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>Product A</td>
<td>Blue</td>
</tr>
<tr>
<td>Direct Mail</td>
<td>Product B</td>
<td>Red</td>
</tr>
<tr>
<td>Tele Sales</td>
<td>Product C</td>
<td>Green</td>
</tr>
</tbody>
</table>

Above, the categorical variables are ‘Channel’, ‘Product’, and ‘Color’, and each are followed by the values that make them up.

2. Check Conditions

Each statistical test has a set of conditions that must be met in order for it to be valid. These are listed in each of the sections describing the different statistical predictive models. Make sure all of the conditions hold before and during the analysis, or your conclusions may be faulty.

3. Carry Out the Analysis

Each section describing a predictive model has a process to follow with step by step instructions in Excel. Be sure to look to the warnings (also included in each section) as you are performing the analysis to make sure everything turns out right.

4. Check for Significance and Fit

The two statistical values described below are useful for almost all of the predictive models described in this guide. They will help you know the validity and fit of your model. Any other statistics that will help validate the predictive model are included in the individual sections.

\textit{p-value}

The p-value is used to know if your model is statistically significant (meaning it can be applied to the entire population, and not just your data sample). It represents the probability that you will be incorrect by saying your predictive model is valid. The lower that value is, therefore, the greater probability you have that the predictive model is statistically valid. Values are between 0 and 1, and in general a model that has a p-value less than .05 is considered statistically significant. This is the level of significance used throughout this guide.

\textit{R}^2

The $R^2$ value is used to show how well your predictive model fits the data. In other words, it tells you how well the model can predict outcomes. Values are between 0 and 1. Values between .8 and 1 show a strong fit, .6-.8 a good fit, .4-.6 show a questionable fit, and under .4 shows there that the model is not a good fit for the data. If the fit of your model is small, you can try three things:

- Try a different predictive model to fit the data
- Add more or different explanatory variables to your predictive model
• Break the data you are testing into smaller categories (a practice known as segmentation), and then build separate models for each of the categories.

If you are using a model containing more than one explanatory variable, you will want to use the ‘R^2 adjusted’ to give you the correct value for this statistic.

5. Validate the Predictive Model

After you have a statistically significant, well fit model, it’s important to do one last test to check its performance. You do this by applying the predictive model to data that you’ve collected, but that was not used in the creation of the model. One method for doing this is to, before creating the predictive model, split the data into two sections. Use one section of data to build the model, and then test the model on the other section to see how accurate its predictions are.

6. Refine the Predictive Model

Now you are ready to apply the predictive model. As you continue to use it, it is important to check its performance and look at ways to improve it. This way your model will stay current with the changing environment.
How to Choose an Appropriate Model

Predict a quantitative outcome with quantitative explanatory variables:

**Regression**
- Correlation
- Linear Regression
- Multivariate
- Exponential Regression
- Logarithmic Regression
- Polynomial Regression
- Time Series

Predict a ‘yes’ or ‘no’ outcome with quantitative explanatory variables:

**Logistic Regression**

Predict a quantitative outcome according to categorical explanatory variables:

**ANOVA**
- t-Test
- 1 Way ANOVA
- 2 Way ANOVA

Predict a categorical outcome according to categorical explanatory variables:

**Chi-Square**
Regression Overview

Predictive Model
Regression models always take the form of an equation, with ‘x’ representing the input, or explanatory variable, and ‘y’ representing the output, or response variable. It is the most common type of statistical test run to create predictive models. It allows the practitioner to predict the outcome of a quantitative variable according to one or more quantitative inputs.

Choosing the Appropriate Model
The type of regression you choose to use will be according to the type of relationship the variables exhibit. (You can see this relationship visually by making a scatter plot)

- Correlation, Linear Regression, and Multivariate Linear Regression all describe a linear relationship between variables:

![Linear Relationship](image1)

- Exponential Regression describes an exponential relationship:

![Positive Exponential](image2)
• **Logarithmic Regression** describes a logarithmic relationship:

![Logarithmic Regression Graph]

• **Polynomial Regression** describes a polynomial relationship:

![2nd Order Polynomial]  
![3rd Order Polynomial]

• **Time Series** describes a trend and/or seasonal relationship:

![Time Series Graph]
General Regression Conditions
The following conditions must hold for all the statistical tests described in this section

- x and y must be quantitative
- y values must have a normal distribution
  - On a standardized residual plot, this is true if you see more values close to zero and less further away, (This is true in the plot below)
  - If your sample size is over 50, it's less important that this condition is met

- y values must have the same variance around each x
  - Looking at the best fit line on a scatter plot - this condition is not met if y output values within specific ranges of x tend to be further or closer from the best fit line than all other y value. (below you can see that the y values spread out further from the best fit line as x gets larger, so the criteria is not met)
- The data must be homogeneous. You can look at the scatter plot to make sure of this (you can’t have large spans of x values with no data).

![Non Homogeneous Group](image)

- The residuals must be independent. You can tell this from the Standardized Residual plot – if there is no pattern in the data (for instance, an upward trend), then they are independent. (The only pattern that is alright to see in the data is that more values are closer to zero than further away. This does not negate independence). The plot below shows independence, as there is no pattern. If this condition is not met, you may need to run a time series analysis (the only test in this section that does not require this condition to be met)

![Standardized Residuals](image)

**How to Check Conditions in Excel 2007**
- To make a scatter plot – highlight the two data sets of interest, the hit ‘insert’ – ‘scatter’ – choose upper left option
• To get a **best fit line** – right click a data point on the scatter plot, click ‘Add Trendline’, and then choose the line that best represents the data


### Regression General Warnings

- Watch for outliers. You can find them on the standardized residual plot – generally points higher than 3 or lower than -3 are considered outliers (these numbers represent standard deviations from the mean). You can see an example of one below:

If you have outliers in the data, do the following:

- Check and make sure there wasn’t a mistake in the collection of the data. If there was, then throw out the outlier from the data set.
- If there was no mistake, it may not be ethical or accurate to throw it out, because it is a real point of data. It’s best to run the regression with and without the outliers, present both results, and give an interpretation of what the outliers may signify.
Correlation

Uses
Correlation is used to find which quantitative variables are associated with each other. It is a preparatory analysis before creating predictive linear regression models. Correlation analysis is especially useful in crunching a lot of data quickly in order to find where relationships exist. Once these relationships are found, it is easier to know what variables to use for a linear predictive model.

Example Questions Answered
- From a large database of demographical data, which factors are associated with response rates?
- From all the metrics on our website, which ones are associated with purchases on the site?
- From all the medical data collected in a patient sample, which factors are associated with blood pressure?

Conditions
- X and Y must be quantitative
- X and Y must have a linear relationship (on a scatter plot, it looks like you could draw a line through them, as seen below)

- Check for normality (if normality is not met, you must use a nonparametric test)
  - Y values must have a normal distribution and the same variance around each X
    - On a standardized residual plot, this is true if you see more values close to zero and less further away (see ‘General Regression Conditions’ for graphic)
    - If your sample size is over 50, it’s less important that this condition is met
  - Y values must have the same variance around each X
    - Look at the best fit line on the scatter plot - this condition is not met if higher or lower values of X tend to be further from the best fit line than other values (see ‘General Regression Conditions’ for graphic)
- The data must be a homogeneous group (see ‘General Regression Conditions’ for graphics)
- The residuals must be independent. You can tell this from the Standardized Residual plot – if there is no pattern in the data (for instance, an upward trend), then they are independent.
‘General Regression Conditions’ for graphic) If this condition is not met, you may need to run a time series analysis

‘How to’ in Excel 2007

1. First, run the correlation - hit ‘Data’ – ‘Data Analysis’ – ‘Correlation’, check the box for ‘Labels in First Row’, and highlight all the data you want to test for correlation (including column labels), hit ‘OK’. You will get out a matrix of r values. Below, the values that have a high correlation are highlighted, showing which data sets are associated with each other.

<table>
<thead>
<tr>
<th></th>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column 1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 2</td>
<td>-0.05351</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 3</td>
<td>0.892018</td>
<td>-0.09966</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Column 4</td>
<td>-0.28392</td>
<td>0.756836</td>
<td>-0.24352</td>
<td>1</td>
</tr>
</tbody>
</table>

2. Next, make scatter plots of the correlated data sets – highlight the two data sets of interest, then hit ‘insert’ – ‘scatter’ – choose upper left option

3. To fit a line on the scatter plot – right click a data point on the plot and click ‘add trendline’ – choose ‘linear’ – ‘close’. This will give you a better idea on how well the data are correlated.

Analysis

**Correlation Coefficient, r (Pearson’s coefficient)**

This measure shows the strength and direction of the correlation between two quantitative variables.

- Positive values mean that as x increases, so does y
- Negative values mean that as x increases, y decreases
- Values are always between -1 and 1
  - .8 – 1 means a very strong correlation
  - .6-.8 means strong correlation
  - .4-.6 means some correlation
  - Less than .4 means little or no correlation

**Scatterplot**

This shows graphically the strength, direction, and consistency of the association

**Warnings**

- You cannot predict anything or say one variable is dependent upon another in this analysis; you can merely state how strongly two variables are or are not correlated to each other
- Watch for Outliers
Nonparametric Counterpart

If the normality condition is not met, you will have to use the nonparametric counterpart to this test, which is known as Spearman’s rank. Spearman’s Rank test can be used with ordinal data (explained below), there is no need for a normal distribution, and the data don’t even need to have a linear relationship.

Ordinal Data

These are qualitative variables with a special feature: they can be ordered and given a numerical value. For example, if a survey asks you to rate your customer experience on a scale from 1=very poor to 5=excellent, it is collecting ordinal data. The numbers 1-5 are categories related to how good the experience was, and their order holds meaning. In simple categorical variables, this is not true (you cannot order male and female in any way to give added meaning).
Linear Regression

Predictive Model

The predictive model you will create in a linear regression analysis is an equation, shown below:

\[ y = Ax + B \]

Above, ‘y’ is the outcome variable, and ‘x’ is the explanatory variable. A is known as the coefficient of x, and B is a constant. The regression test will allow you to find the appropriate coefficient and constant values to define the relationship between y and x, allowing you to predict outcomes according to specific ‘x’ values.

Uses

Linear regression predicts the value of one response variable, y, given the value of one explanatory variable, x. It is often done directly after a correlation analysis.

Example Questions Answered

- If we send out x number of mail pieces to a target customer segment, how many sales will we get in return?
- If we sell x number of product A, how much can we expect to sell of product B?
- If x number of people sign up for this promotional program, how many extra donations can we expect?

Conditions

- x and y must be quantitative
- x and y must have a linear relationship
- Check for normality (if normality is not met, you must use a nonparametric test)
  - y values must have a normal distribution and the same variance around each x
    - On a standardized residual plot, this is true if you see more values close to zero and less further away (see ‘General Regression Conditions’ for graphic)
    - If your sample size is over 50, it’s less important that this condition is met
  - y values must have the same variance around each x
    - Look at the best fit line on the scatter plot - this condition is not met if higher or lower values of x tend to be further from the best fit line than other values (see ‘General Regression Conditions’ for graphic)
- Must have a homogeneous group (see ‘General Regression Conditions’ for graphics)
- The residuals must be independent. You can tell this from the Standardized Residual plot – if there is no pattern in the data (for instance, an upward trend), then they are independent. (see ‘General Regression Conditions’ for graphic) If this condition is not met, you may need to run a time series analysis
‘How to’ in Excel 2007

1. To make a scatter plot – highlight the two data sets of interest, then hit ‘insert’ – ‘scatter’ – choose upper left option. If you notice a linear relationship, continue with this method.


3. The p-values, $R^2$, and coefficients of interest are highlighted below in the regression output

### SUMMARY OUTPUT

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Squ</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

### ANOVA

<table>
<thead>
<tr>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>2629223.328</td>
<td>2629223</td>
<td>111.0869</td>
</tr>
<tr>
<td>Residual</td>
<td>23</td>
<td>544366.0924</td>
<td>23668.18</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>3173591.36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-Value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
<th>Lower 95.0%</th>
<th>Upper 95.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>185.5072973</td>
<td>63.84813574</td>
<td>2.938808</td>
<td>0.0068827</td>
<td>57.42671345</td>
<td>321.5878511</td>
<td></td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.050958119</td>
<td>0.004813839</td>
<td>10.3978</td>
<td>2.815-10</td>
<td>0.04095649</td>
<td>0.060959746</td>
<td></td>
</tr>
</tbody>
</table>

(The predictive model for the above output would be $y = 0.51x + 189.518$

**Analysis**

**p-value**

There will be a p-value for the coefficient and the intercept, and both must be below .05 for the predictive model to be completely validated.

**$R^2$**

This shows the percentage of variability in ‘$y$’ that is explained by ‘$x$’ in the predictive equation. If the value is equal to 1, that means all the variability in ‘$y$’ is explained perfectly (don’t ever expect this to happen).

**Confidence Interval**

Excel gives you ‘Lower 95%’ and ‘Upper 95%’ values for both the intercept and the coefficients. These values provide a range in which, statistically, you are 95% sure that the true value of either the intercept or the coefficient lies within. These values are highlighted below:
With the above data, I can say that I am 95% sure that the true value of the intercept of my equation lies between 57 and 321. I can also say that I am 95% sure that the true value of the x coefficient lies in between .04 and .06.

**Warnings**

- Don’t predict responses in ‘y’ for an ‘x’ that is outside the range of the data you built the model on; you can’t be sure the relationship will hold outside this range.
- Watch for [Outliers](#)

**Nonparametric Counterpart**

If the normality condition is not met, you will have to use the nonparametric counterpart to this test. One popular nonparametric test that can replace any parametric regression test is MARSplines (Multivariate Adaptive Regression Splines). It doesn’t impose the condition that any specific type of relationship exists between the variables (such as linear or exponential), can be used with multiple explanatory variables, and can even predict multiple outcome variables.