Regression Analysis:

A statistical procedure used to find relationships among a set of variables

In regression analysis, there is a dependent variable, which is the one you are trying to explain, and one or more independent variables that are related to it.

You can express the relationship as a linear equation, such as:

\[ y = a + bx \]

You might recall this as the equation for a line from linear algebra.

- \( y \) is the dependent variable
- \( x \) is the independent variable
- \( a \) is a constant
- \( b \) is the slope of the line
- For every increase of 1 in \( x \), \( y \) changes by an amount equal to \( b \)
- Some relationships are perfectly linear and fit this equation exactly. Your cell phone bill, for instance, may be:

**Total Charges = Base Fee + 30¢ (overage minutes)**

If you know the base fee and the number of overage minutes, you can predict the total charges exactly.

Other relationships may not be so exact. Weight, for instance, is to some degree a function of height, but there are variations that height does not explain. On average, you might have an equation like:

\[ \text{Weight} = -222 + 5.7 \times \text{Height} \]

If you take a sample of actual heights (inches) and weights (pounds), you might see something like the graph to the right.
Regression finds the line that best fits the observations. It does this by finding the line that results in the lowest sum of squared residuals.

Why squared? The sum of the negative residuals (for points below the line) will exactly equal the sum of the positive residuals (for points above the line). Summing just the residuals wouldn’t be useful because the sum is always zero. So, instead, regression methods uses the sum of the squares of the residuals.

An Ordinary Least Squares (OLS) regression finds the line that results in the lowest (least) sum of squared residuals.

Multiple Regression

What if there are several factors affecting the independent variable?

As an example, think of the price of a home as a dependent variable. Several factors contribute to the price of a home… among them are square footage, the number of bedrooms, the number of bathrooms, the age of the home, whether or not it has a garage or a swimming pool, if it has both central heat and air conditioning, how many fireplaces it has, and, of course, location.

The Multiple Regression Equation

Each of these factors has a separate relationship with the price of a home. The equation that describes a multiple regression relationship is:

\[ y = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n + e \]

This equation separates each individual independent variable from the rest, allowing each to have its own coefficient describing its relationship to the dependent variable. If square footage is one of the independent variables, and it has a coefficient of $50, then every additional square foot of space adds $50, on average, to the price of the home.
How Do You Run a Regression?

In a Multiple Regression Analysis of home prices, you take data from actual homes that have sold recently. You include the selling price, as well as the values for the independent variables (square footage, number of bedrooms, etc.). The multiple regression analysis finds the coefficients for each independent variable so that they make the line that has the lowest sum of squared residuals.

How Good is the Model?

One of the measures of how well the model explains the data is the $R^2$ value. Differences between observations that are not explained by the model remain in the error term. The $R^2$ value tells you what percent of those differences is explained by the model. An $R^2$ of .68 means that 68% of the variance in the observed values of the dependent variable is explained by the model, and 32% of those differences remains unexplained in the error (residual) term.

Sometimes There’s No Accounting for Taste

Some of the error is random, and no model will explain it. A prospective homebuyer might value a basement playroom more than other people because it reminds her of her grandmother’s house where she played as a child. This can’t be observed or measured, and these types of effects will vary randomly and unpredictably. Some variance will always remain in the error term. As long as it is random, it is of no concern.

Some of the error isn’t error

Some of the error is best described as unexplained residual—if we added additional variables (such as, for homes in Vancouver, the high school catchment that the home lies within) we might be able to reduce the residual. (See the discussion below on omitted variables.)
“p-values” and Significance Levels

Each independent variable has another number attached to it in the regression results… its “p-value” or significance level.

The p-value is a percentage. It tells you how likely it is that the coefficient for that independent variable emerged by chance and does not describe a real relationship.

A p-value of .05 means that there is a 5% chance that the relationship emerged randomly and a 95% chance that the relationship is real.

It is generally accepted practice to consider variables with a p-value of less than .1 as significant, though the only basis for this cutoff is convention.

Significance Levels of “F”

There is also a significance level for the model as a whole. This is the “Significance F” value in Excel; some other statistical programs call it by other names. This measures the likelihood that the model as a whole describes a relationship that emerged at random, rather than a real relationship.

As with the p-value, the lower the significance F value, the greater the chance that the relationships in the model are real.

Some Things to Watch Out For

• Multicollinearity
• Omitted Variables
• Endogeneity
• Other

Multicollinearity

Multicollinearity occurs when one or more of your independent variables are related to one another. The coefficient for each independent variable shows how much an increase of one in its value will change the dependent variable, holding all other independent variables constant. But what if you cannot hold them constant? If you have two houses that are exactly the same, and you add a bedroom to one of them, the value of the house may go up by, say, $10,000. But you have also added to its square footage. How much of that $10,000 is a result of the extra bedroom and how much is a result of the extra square footage?
**Multicollinearity**

If the variables are very closely related, and/or if you have only a small number of observations, it can be difficult to separate these effects. Your regression gives you the coefficients that best describe your set of data, but the independent variables may not have a good p-value if multicollinearity is present. Sometimes it may be appropriate to remove a variable that is related to others, but it may not always be appropriate. In the home value example, both the number of bedrooms and the square footage are important on their own, in addition to whatever combined effects they may have. Removing one of them may be worse than leaving both in. This does not necessarily mean that the model as a whole is problematic, but it may mean that the model should not be used to draw conclusions about the relationship of individual independent variables with the dependent variable.

---

**Omitted Variables**

If independent variables that have significant relationships with the dependent variable are left out of the model, the results will not be as good as if they are included. In the home value example, any real estate agent will tell you that location is the most important variable of all. But location is hard to measure. Locations are more or less desirable based on a number of factors. Some of them, like population density or crime rate, may be measurable factors that can be included. Others, like perceived quality of the local schools, may be more difficult. You must also decide what level of specificity to use. Do you use the crime rate for the whole city, a quadrant of the city, the zip code, the street? Is the data even available at the level of specificity you want to use? These factors can lead to omitted variable bias: variance in the error term that is not random and that could be explained by an independent variable that is not in the model. Such bias can distort the coefficients on the other independent variables, as well as decreasing the $R^2$ and increasing the Significance F. Sometimes data just isn’t available, and some variables aren’t measurable. There are methods for reducing the bias from omitted variables, but it can’t always be completely corrected.

---

**Endogeneity**

Regression measures the effect of changes in the independent variable on the dependent variable. Endogeneity occurs when that relationship is either backwards or circular, meaning that changes in the dependent variable cause changes in the independent variable.
Endogeneity

In the home value example, as mentioned above, the perceived quality of the local school might affect home values. But the perceived quality is likely also related to the actual quality, and the actual quality is at least partially a result of funding levels. Funding levels are often related to the property tax base, or the value of local homes. So... good schools increase home values, but high home values also improve schools. This circular relationship, if it is strong, can bias the results of the regression. There are strategies for reducing the bias if removing the endogenous variable is not an option.

Others

There are several other types of biases or sources of distortion that can exist in a model for a variety of reasons. Spatial autocorrelation is one significant bias that can greatly affect aspatial regression. There are tests to measure the levels of bias, and there are strategies that can be used to reduce it. Eventually, though, one may have to accept a certain amount of bias in the final model, especially when there are data limitations. In that case, the best that can be done is to describe the problem and the effects it might have when presenting the model.

Regression models

• There are many different types of regression models (linear, nonlinear, multiple, GWR, logistic, multi-level, to name but a few) since there are many different types of variables used (nominal, ordinal, ratio) and relations are not always linear nor occurring at the same level of geography (household effects [parental income] as well as neighbourhood effects [the quality of the school]).

Multiple Regression: Example

• Suppose a gas utility company is trying to estimate revenue. They may have determined that heating cost is a function of the temperature, the amount of insulation in an attic, and the age of a furnace. They decided to look at 20 customer sites, and quantify the data as shown.
• Microsoft Excel has a regression tool for relatively small problems. You will find it under the tools -> data analysis tab.  
• Once you select the tool, an interactive dialog will come up steering you through the regression wizard.  
• Here is where you will enter the range for the Y value (a single column), and the X values (multiple columns) as shown to the right. 
• You should type the numbers into Excel, and attempt to perform the regression yourself. Check your answers against the ones presented here. 
• What this tells us is that our R-square value is quite high (.80) representing a good fit, and we have a standard error of 51 (in dollars) for our 20 observations. 
• The next chart tells us our coefficient values (intercept, temperature, insulation, age of furnace). It also tells us our P-value, or a measure of significance. All the values except age of furnace are very low, meaning that they are all significant at the 95% level.  
• So, what we now have is a formula 
  – Therefore, if a person with no attic insulation decided to add 12 inches, what would they save when the average temperature if 12 degrees? 
• The intercept is 427.194. This the cost of heating when all the independent variables are equal to zero.  
• The regression coefficients for the mean temperature and the amount of attic insulation are both negative. This is logical: as the outside temperature increases, the cost of heating the house will go down.  
• For each degree the mean temperature increases, we expect the heating cost to decrease $4.583 per month.  
• P-value for all the coefficients are significant for \( \alpha =0.05 \) except for the coefficient of the variable “age of furnace”\( \beta _3 \). Hence, we can conclude that for the other variables they are significantly different from zero.  
• However, if we examine the p-value for the variable “age of furnace”, we see that it is not significant at \( \alpha =0.05 \). Hence, we cannot conclude that it is significantly different from zero.  
• In that case, we can drop this variable from the model. Let’s see what happens if we drop the “age of furnace variable from the model 

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>427.1938033</td>
<td>59.60142931</td>
<td>7.167509374</td>
<td>2.23764E-06</td>
</tr>
<tr>
<td>Temperature (oF)</td>
<td>-4.582662626</td>
<td>0.772319353</td>
<td>-5.933636915</td>
<td>2.10035E-05</td>
</tr>
<tr>
<td>Attic Insulation</td>
<td>-14.83086269</td>
<td>4.754412281</td>
<td>-3.119389277</td>
<td>0.006605963</td>
</tr>
<tr>
<td>Age of Furnace</td>
<td>6.101032061</td>
<td>4.012120166</td>
<td>1.520650381</td>
<td>0.147862484</td>
</tr>
</tbody>
</table>

What it means 

• The intercept is 427.194. This the cost of heating when all the independent variables are equal to zero.  
• The regression coefficients for the mean temperature and the amount of attic insulation are both negative. This is logical: as the outside temperature increases, the cost of heating the house will go down.  
• For each degree the mean temperature increases, we expect the heating cost to decrease $4.583 per month.  
• P-value for all the coefficients are significant for \( \alpha =0.05 \) except for the coefficient of the variable “age of furnace”\( \beta _3 \). Hence, we can conclude that for the other variables they are significantly different from zero.  
• However, if we examine the p-value for the variable “age of furnace”, we see that it is not significant at \( \alpha =0.05 \). Hence, we cannot conclude that it is significantly different from zero.  
• In that case, we can drop this variable from the model. Let’s see what happens if we drop the “age of furnace variable from the model.
Multiple Regression

- Calculating the savings due to increased insulation:
  - Cost to Heat Home = 427 + -4.58(12) + -14.83(0) + 6.101(6).
  - $408 (with no insulation)
  - $230 (with the additional 12 inches of insulation)

- A utility company could then use this information to determine how much revenue they would generate if they provided service to a neighborhood.

Using Geography in Multiple Regression

- GIS is a great tool for obtaining the explanatory variables. For example, consider the following problem to solve.
  - Assume that an environmental remediation company wants to know how much phosphorous is being dumped in a lake.
  - If they had all the data together, they could develop a regression model to predict the amount, and then prescribe different land use options for reducing the load.

- Let's assume that the company determined that the following information is a pretty good predictor of phosphorous loading:
  - The land use – developed land, and dairy farms will have more phosphorous than forest
  - Distance to a stream – areas near a stream will be more likely to load into the lake
  - The soil type – certain soil types and their erosion factors may play a role in the amount of loading.
  - Slope – areas uphill from the water will be more likely to load into the lake.

- You could extract all of the data from the various layers and run a regression model to see what impact each factor has on the observations.

- A major problem with map-based regression models is that the variables are likely spatially autocorrelated, which means that the results can be biased.

- GWR attempts to overcome that bias.