Chapter 2 : Section 2

**Multiple Linear Regression**

Introduction to Multiple Linear Regressions

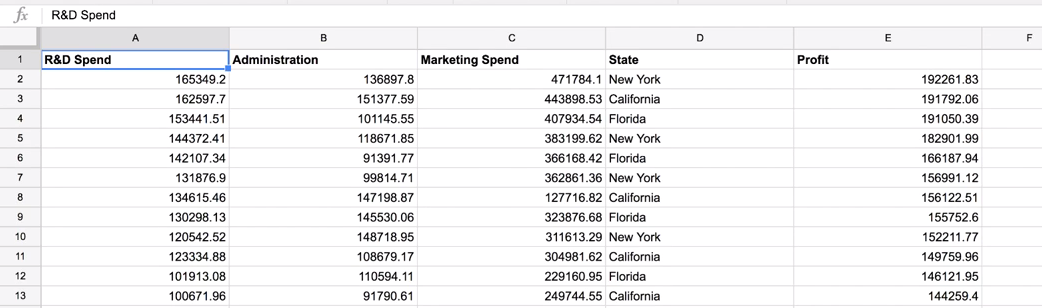
**2.1 Multiple Linear Regression**

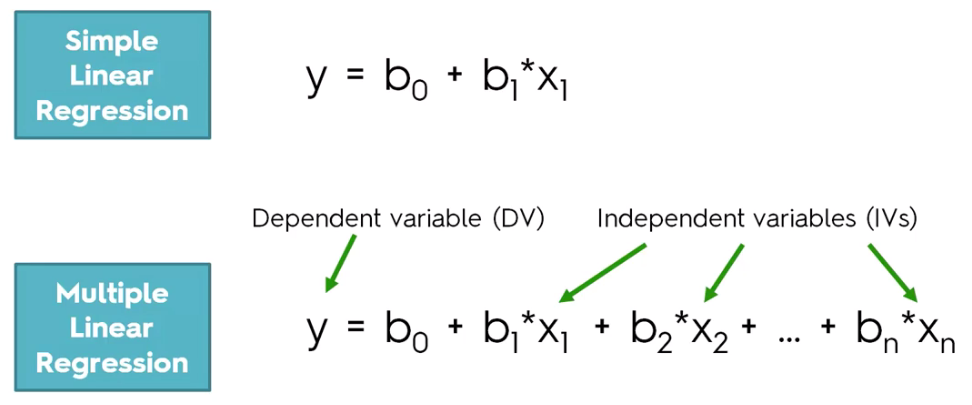
There many cases in which the *response/dependent variable* is affected by more than one *predictor/independent variable*; for such cases, the Multiple Linear Regression algorithm is used.

* Multiple Linear Regression is an extension of Simple Linear regression as it takes *more* than one *predictor variable* to predict the *response variable*.
* Multiple Linear Regression is one of the important regression algorithms which models the *linear relationship* between a *single dependent continuous variable* and *more than one independent variable*.
* Example: Prediction of CO2 emission based on engine size and number of cylinders in a car.
* Some key points about MLR:
* For ***MLR***, the *dependent or target variable* Y must be the *continuous/real*, but the *predictor or independent variable* may be of *continuous* or *categorical* form.
* Each *feature variable* must model the *linear relationship* with the *dependent variable*.
* ***MLR*** tries to ***fit*** a ***regression line*** through a ***multidimensional space of data-points***.

**2.2 Equation of MLR|**

* More than two independent variables: In multi linear regression, we deal with more than one independent variables.
* In terms of a student and his Grades is *depend* *variable*. What grade does a student get? Then in the *independent* *variables* could be how much the student has studied for the exam (***study time***), how much he has slept before the exam (***sleep time***), how many lectures he has attended throughout the course (***attendance***) and the things like that.
* The equation:
* is the dependent variable, is the constant, are coefficients and are independent variables .
* Consider the following dataset.





**2.3 Multiple Linear Regression assumptions**

|  |  |
| --- | --- |
| * ASSUMPTIONS: Before building a linear regression model you need to check that these assumptions are true. * If you ever do need to build a linear regression then make sure that you don't just blindly follow the steps presented here. * First you need to check the assumptions of the linear regression and research them and make sure that they are correct when you're building your regression model. * And then you can only proceed and be sure that you're building a good linear regression model. |  |

1. Linearity: There must be a linear relationship between the dependent variable and the independent variables.

* ***Scatterplots*** can show whether there is a linear or curvilinear relationship.

1. Homoscedasticity: This assumption states that the variance of error terms is similar across the values of the independent variables.

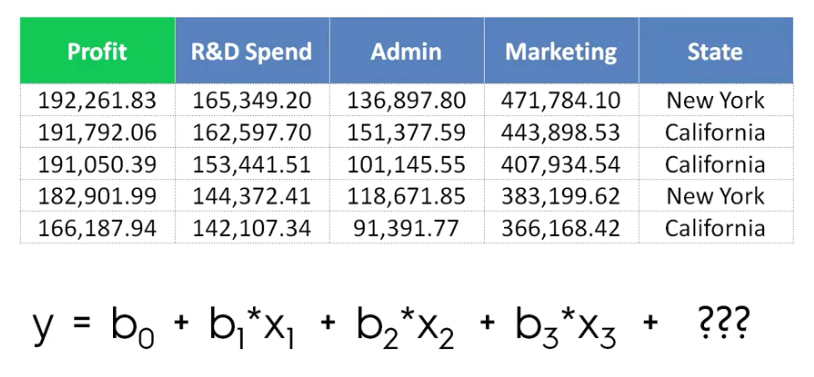
* A ***plot*** of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

1. Multivariate Normality: Multiple Linear Regression assumes that the residuals (the **differences** **between** the observed value of the dependent variable and the predicted value )are ***normally*** ***distributed***.
2. Independence of errors: Multiple Linear Regression assumes that the residuals (the **differences** **between** the observed value of the dependent variable and the predicted value ) are ***independent***.
3. Lack of multi-collinearity: Multiple Linear Regression assumes that the independent *variables* are not *highly* *correlated* with each other. This assumption is tested using Variance Inflation Factor (VIF) values.

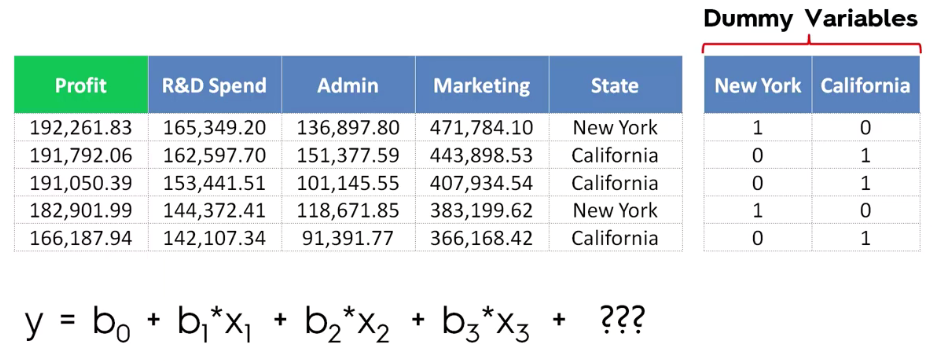
**2.4 Dummy Variables**

In following table profit is our dependent variable and the rest the blue ones are all independent variables.

* What should we place in our equation for the ***State*** column because we don't actually have numerical values here, the ***State*** is actually a ***categorical*** ***variable*** (there is categorical variables and there's numeric variables).
* First you need to go through your column and find all the different categories you have. For every single category that you find, you need to create a new column. Then in each category-column put **1** where corresponding matching category appears and **0** for unmatched.



* The approach that you need to take when you face ***categorical*** ***variables*** in ***regression*** ***models*** is you need to create dummy variables.
* So in this case we have two categories. Hence, we need to create a new column for New York and one for California. So we're kind of expanding our data set and adding some additional columns into it.
* Then we find all of rows where the state actually says New York and put a **1**.



|  |  |
| --- | --- |
|  |  |

* Do not use all the Dummy variables: This can leads us to collinearity and we eventually fall into dummy variable trap.
* All you have to do is use the ***New York*** column instead of ***States***.
* Don't use the California column (leads to dummy variable trap).
* Here all the information in our data is preserved: If we just stick to the one ***New York***, **1** means ***New York*** and **0** means ***California***. So we didn't lose any information by including only the New York. work as switches.
* Dummy variables act as switches: Actually all of the dummy variables they work as switches.
* Omit one dummy?: When you look at this approach it might seem biased. There is a coefficient for ***New York*** but for ***California*** there's no coefficient.
* In reality that's not the case because the way *regression* *models* work is that the coefficient of the dummy variable that you have not included will become the *default* *situation* for this *regression* *model*.

What that means is that the coefficient for California is going to be included in the constant by default. When is equal to zero this whole equation will turn into an equation for California.

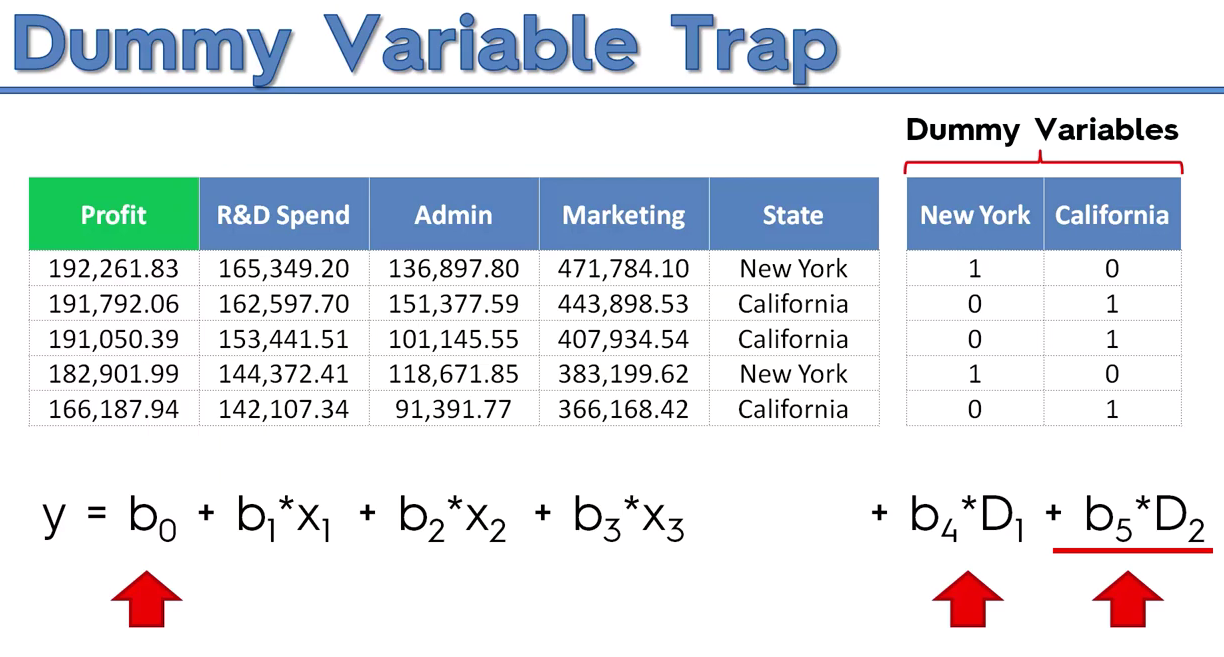
When becomes you're adding which is the difference between *New York* and *California* *coefficient*.

So basically you're altering from California to New York by flipping this light switch if it's on off. And the default state or the whole equation is working for California.

**2.5 Dummy variable trap**

**Why you should never include all of your dummy variable columns**

* Multicollinearity: Intuition here is that you're basically ***duplicating a variable***. This is because .
* The phenomenon where one or several independent variables in a linear regression predict another is called MULTICOLLINEARITY. As a result of this effect the model cannot distinguish between the effects of from the effects from of. Therefore it won't work properly. And this is the Dummy Variable Trap.
* If you do the math behind this scenario you will see that the real problem is that you cannot have these three elements in your model at the same time the constant and both the dummy variables , .

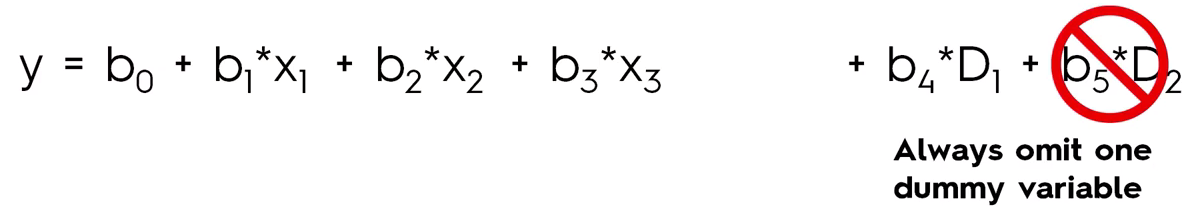


* How is the coefficient related to the dummy variable trap?

Since then if you include both and you get:

Where and

Therefore the information of the redundant dummy variable is going into the constant .



* Whenever you're building a model always omit one dummy variable and this applies to the number of dummy variables in that specific dummy set. If you have **9** then you should only include **8**, if you have **100** then you should only include **99** of them.
* Also note that if you have two sets of dummy variables then you need to apply the same rule to each set.

**2.6 p-value**

Before we get into Backward Elimination, make sure to be introduced to the ***p-value*** and have a basic understanding of how it works. By looking at almost all the explanations of the p-value on the internet

* What is the p-value?
* Null Hypothesis: To understand the P-value, we need to start by understanding the null hypothesis: the null hypothesis is the assumption that the parameters associated to your independent variables are equal to zero. Therefore under this hypothesis, your ***observations are totally random***, and don’t follow a certain pattern. For example:
* ***Does the size of a state affect population density?*** The null hypothesis is "all states have the same population density."
* ***Do cats prefer fish or milk?*** The null hypothesis is "cats have no preference; they like them the same."
* P-value: The *P-value* is the probability that the parameters associated to your independent variables have *certain nonzero values*, given that the *null hypothesis is True*. The most important thing to keep in mind about the P-Value is that it is a statistical metric: the ***lower is the P-Value, the more statistically significant is an independent variable,*** that is the better predictor it will be.
* The ***smaller*** the ***p-value***, the stronger the evidence that you should ***reject*** the ***null hypothesis***. A ***p-value*** less than ***0.05*** (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the *null is correct (and the results are random)*.
* The first step in backward elimination is pretty simple, you just select a *significance level*, or select the *P-value*. Usually, in most cases, a *5%* significance *level is selected*. This means the *P-value* will be *0.05*.

|  |  |
| --- | --- |
| * *P value* is a *statistical* *measure* that helps scientists determine whether or not their *hypotheses* *are correct*. *P values* are used to *determine* whether the results of their experiment are within the *normal* *range* *of values* for the events being *observed*. Usually, if the P value of a data set is below a certain pre-determined amount (like, for instance, 0.05), scientists will *reject* the *"null hypothesis"* of their experiment. * The p-value is actually the probability of getting a sample like ours, or *more* ***extreme*** *than ours IF the* ***null hypothesis*** *is true*. So, we *assume* the *null hypothesis is true* and then determine how “strange” our sample really is. If it is *not that strange* (a *large p-value*) then we *don’t change our mind about the* ***null******hypothesis***. As the **p-valu**e gets **smaller**, we start wondering if the null really is true and well maybe we should change our minds (and reject the null hypothesis). |  |

* A little more detail: A small p-value indicates that by pure luck alone, it would be unlikely to get a sample like the one we have if the null hypothesis is true. If this is small enough we start thinking that maybe we aren’t super lucky and instead our assumption about the null being true is wrong. Thats why we reject with a small p-value.
* A large p-value indicates that it would be pretty normal to get a sample like ours if the null hypothesis is true. So you can see, there is no reason here to change our minds like we did with a small p-value.
* In inferential statistics, the null hypothesis (often denoted H0)[1] is that two possibilities are the same. The null hypothesis is that the observed difference is due to chance alone. Using statistical tests, it is possible to calculate the likelihood that the null hypothesis is true.

**2.7 Feature Selection methods**

Different methods to select significant features

Implement MLR

Implementation of Multiple Linear Regression (MLR) model using Python. To implement MLR using Python, we have below problem:

* Problem Description:

We have a dataset of 50 start-up companies. This dataset contains five main information: *R&D* Spend (Research and development), *Administration* Spend, *Marketing* Spend, *State*, and *Profit* for a financial year.

* Goal: To create a model that can easily determine which company has a maximum profit, and which is the most affecting factor for the profit of a company.

Since we need to find the Profit, so it is the dependent/target variable, and the other four variables are independent variables. Below are the main steps of deploying the MLR model:

1. *Data Pre-processing* Steps
2. *Fitting the MLR* model to the training set
3. *Predicting* the result of the test set

|  |  |
| --- | --- |
| * We analyze these 50 companies over this data set and create a model that will tell us which types of companies we should invest and our main criteria is the profit. * What we are looking for is: we want to understand for instance where companies perform better in *New York* or *California* all other things held equal or which companies perform better if you hold this "State" column equal. * We want to know: * Will a company that *spends more* on *marketing* perform *better* or a company spends less on marketing. * Also we want to understand how a company spend more on *R&D* spend or to spend more on *marketing*. | D:\100_days_comer_soln\multiple-linear-regression-in-machine-learning.png |

* It will help us to set up a set of guidelines for our own venture capitalist fund . For example, we are more interested in companies that work in *New York* and that have a very *low* *administration* spend and a very *high* *R&D* spend which is much higher than *Administration* or *Marketing* spend.
* So basically we are creating a model based off of this sample that will allow us to assess where and in which into which companies we want to invest to achieve their goal of maximizing profit.