**Learning Data Science with GRETL**

**What is Data Science?**

* Why Data Science?
* Areas of Data Science – Statistics, Data Visualization, Data Mining, Data Processing, Machine Learning

**Basics of Statistics**

* Introduction
* Types of variables: Qualitative vs Quantitative
* Types of regressions
* Ordinary Least Squares
* R-squared
* Adjusted R-squared

**Simple Linear Regression**

* Introduction to GRETL
* Working with a dataset and analysis
* Import data and run descriptive statistics
* Reading Linear Regression Output – understanding coefficients, P-values and R Squared
* Plotting and analyzing the graph

**Multiple Linear Regression**

* Introduction and assumptions of a Multiple Linear Regression
* Working with a dataset and analysis
* Import data and run descriptive statistics
* Dummy Variables and Dummy variable trap
* Step by Step model building process - BACKWARD, FORWARD, STEPWISE
* Practice Backward Elimination with real-time examples
* Using Adjusted R-squared to create Robust models
* Interpreting coefficients of MLR

**Logistic Regression**

* Introduction to Logistic Regression
* Binary outcome: Yes/No-Type Business Problems
* Logistic regression exercise
* False Positives ,False Negatives, Confusion Matrix
* Interpreting coefficients of a logistic regression

**Geo-demographic Segmentation model**

* What is Geo-demographic segmentation
* Understanding the dataset
* Building the model using Backward elimination method
* Transforming independent variables and creating derived variables
* Checking for Multicollinearity and understanding Correlation matrix

**Model Assessment**

* Need for assessing the model – Accuracy Paradox
* Understanding Cumulative Accuracy Profile (CAP)
* How to build a CAP curve in Excel
* Assessing model using the CAP curve
* Using CAP curve template
* Prevent overfitting
* Applying the model to test data - compare training performance and test performance

**Deriving insights from the model**

* Insights from the CAP
* Coefficients of a Logistic Regression
* Odds ratio
* Odds Ratio vs Coefficients in a Logistic Regression
* Deriving insights from coefficients

**Model maintenance**

* Why do models deteriorate with examples
* Three levels of maintenance of deployed models

**Communication**

* Importance of communication and working with people
* Presentation skills

Statistics

* 1. Introduction
  2. Types of variables: Categorical vs. Numeric
  3. Types of regressions
  4. Ordinary Least Squares
  5. R-squared
  6. Adjusted R-squared

Introduction

* ***Understanding statistics is an important part of Data Science as it is forms the fundamentals based on which models can be created.***
* ***These models when applied to raw data sets are able to transform them into insights by recognizing patterns within the data that can be used to make predictions and forecast behavioral changes.***
* ***Core of Statistics is to be able to recognize the individual data elements of the data set as variables and the inter-relationship between them.***

Variables



***Qualitative*:** Also known as ***Categorical***, describes Quality or characteristic of a data element and its values belong to a collection of known, defined non-overlapping classes.

E.g.: Student letter grading (A, B, C, D), commercial bond rating (AAA, AAB...), consumer clothing shoe sizes (1, 2, 3... 12).

2 Types of Qualitative variables

**Nominal**: Variables that cannot be organized into a particular sequence. E.g.: Colors: Gender

**Ordinal:** Variables that can be ordered or sorted. E.g.: Grades

***Quantitative*:** Variables are numbers which answer the questions like how much or how many.

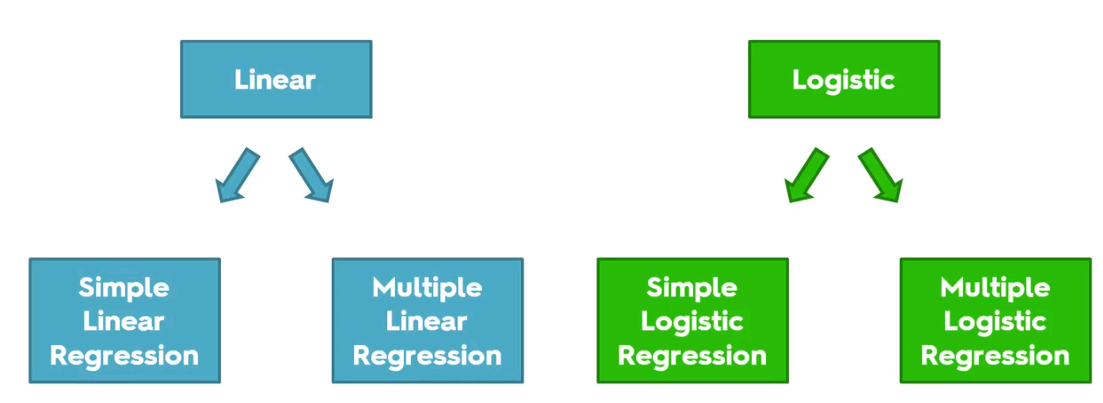
2 types of quantitative variables:

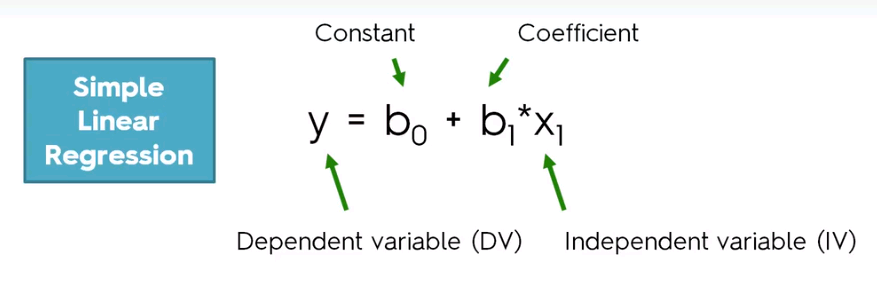
**Discrete**: Whole numbers and a finite range of real values. Eg: 1 to... n number of people, cars

**Continuous:** Unlike discrete variables they can be fractions of infinite nature.

Regression

It is a Statistical process for estimating the relationship between variables. Focus is on relationship between dependent and independent variables.





***Dependent variable (DV)*** is dependent on the result of the function of the IV w.r.t the coefficient and constant. It is something that needs to be explained.

Eg: Employee’s Salary **(y)** is a DV thatchanges with IV years of experience.

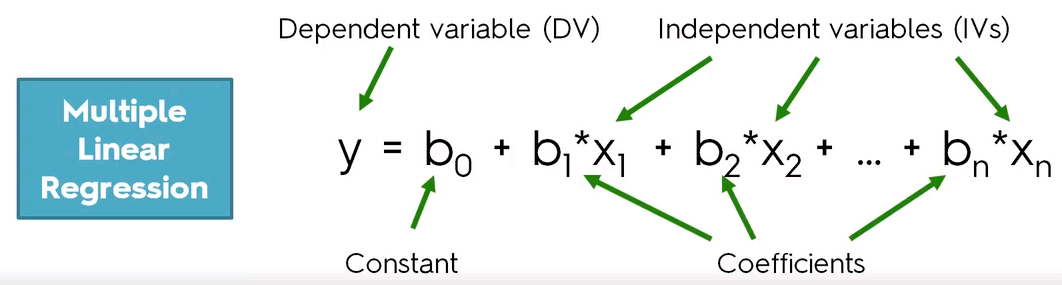
What grade does a Student get depending on the time and effort he has put into his coursework.

***Independent variable (IV)*** is the variable that is assumed to be causing the change in DV.

***Coefficient:*** As per the above formula, the coefficient b1 says how much (or the proportion) of a unit change in X1 has an impact on the DV that is y.

Example: If we think of ‘Y’ as Change in Salary, then we can think of X1 as the HR Appraisal rating.

B0 would be the Current Salary



Dependent variable is dependent on the result of the function of multiple Independent variable w.r.t their individual coefficient and a single constant value.

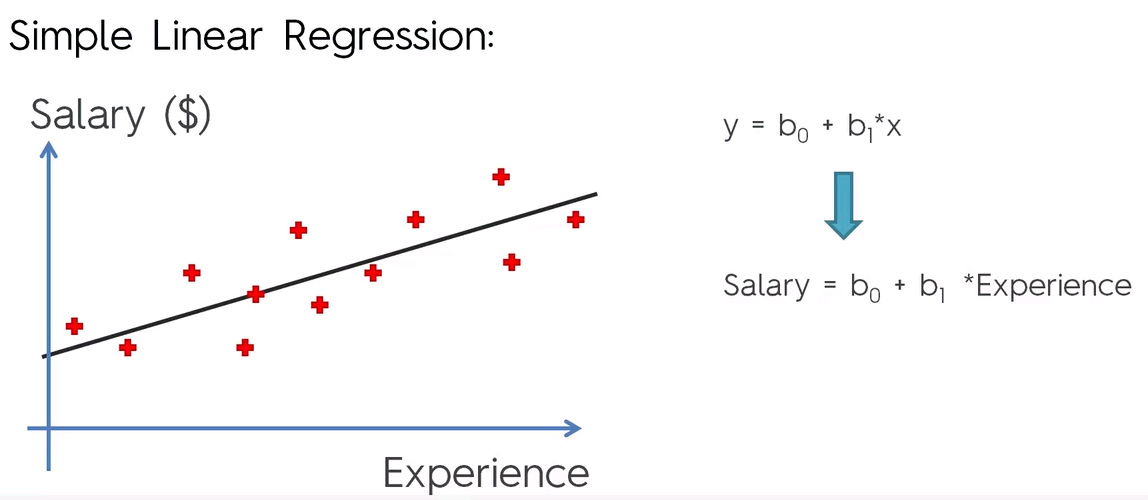
Example: If we think of ‘Y’ as Change in Salary, then we can think of X1 as the HR Appraisal rating,

X2 as Technology/ Domain employee works on and other parameters like Customer feedback, new technology/ skills updated on profile.

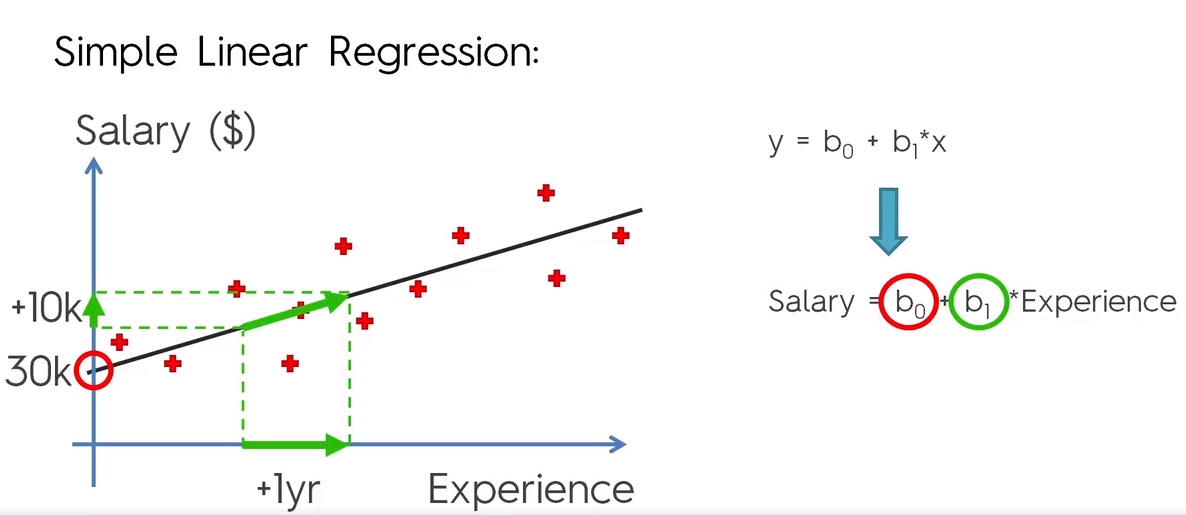
—

Let us take the example by plotting a graph for Experience vs Salary.

Based on sample data available, the Red Cross represents Salaries of people with various years of experience



Linear Regression: **Is the best fitting line in graph as represented above with the black line.**



b0 – Is the constant (or also referred to as the Y intercept) and here it represents the Salary at 0 Yrs of experience, let us consider it as 30K.

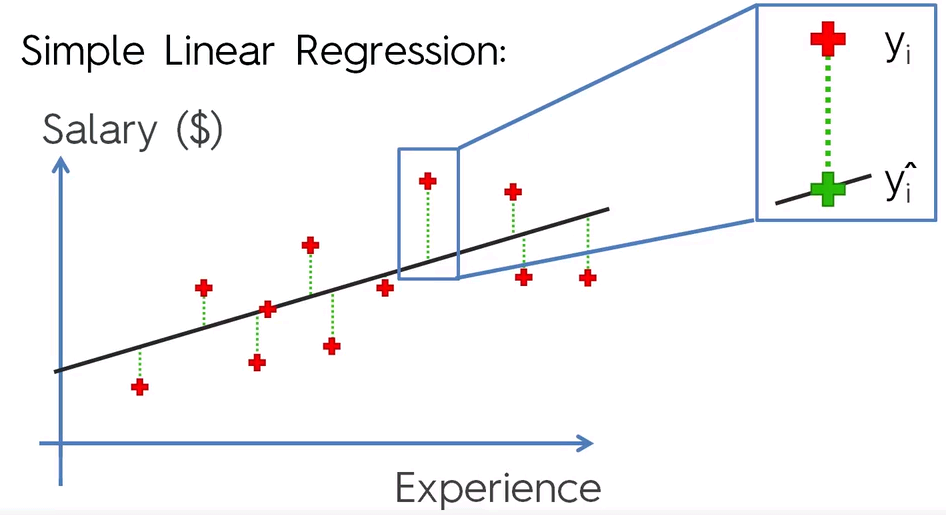
b1 – represents the slope or the angle created by the regression line.

It is a function of the change in the Salary (Y axis) over a unit change in Experience (X Axis).

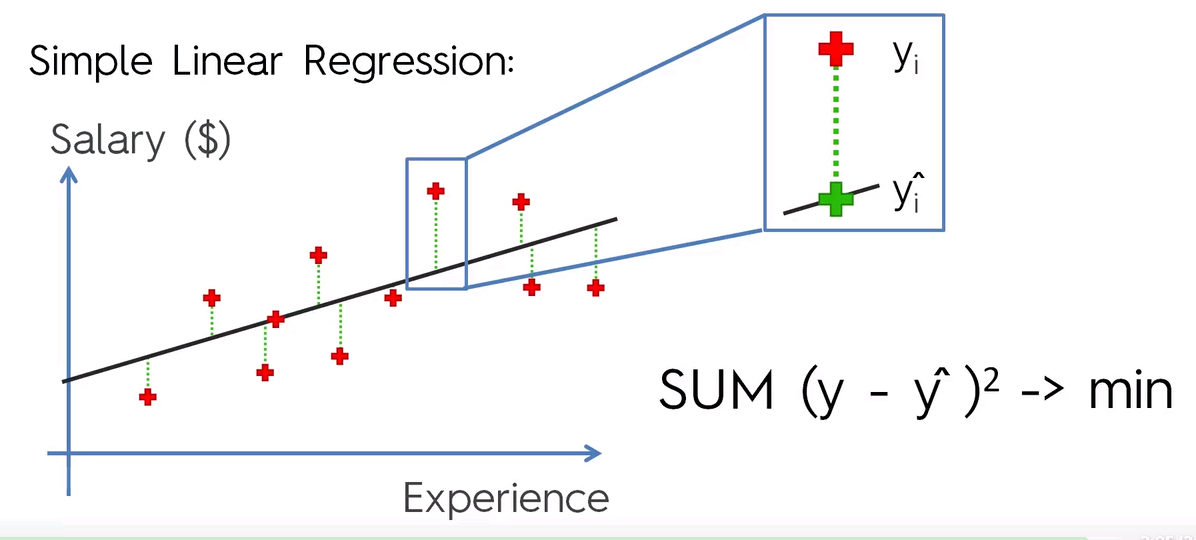
Ordinary Least Square Method

Ordinary least square is **the method that is used to derive linear regression line.**

As shown below, Yi is the actual Salary value and Yi ^ is the Salary value predicted by the linear regression model.



Ordinary Least Square method



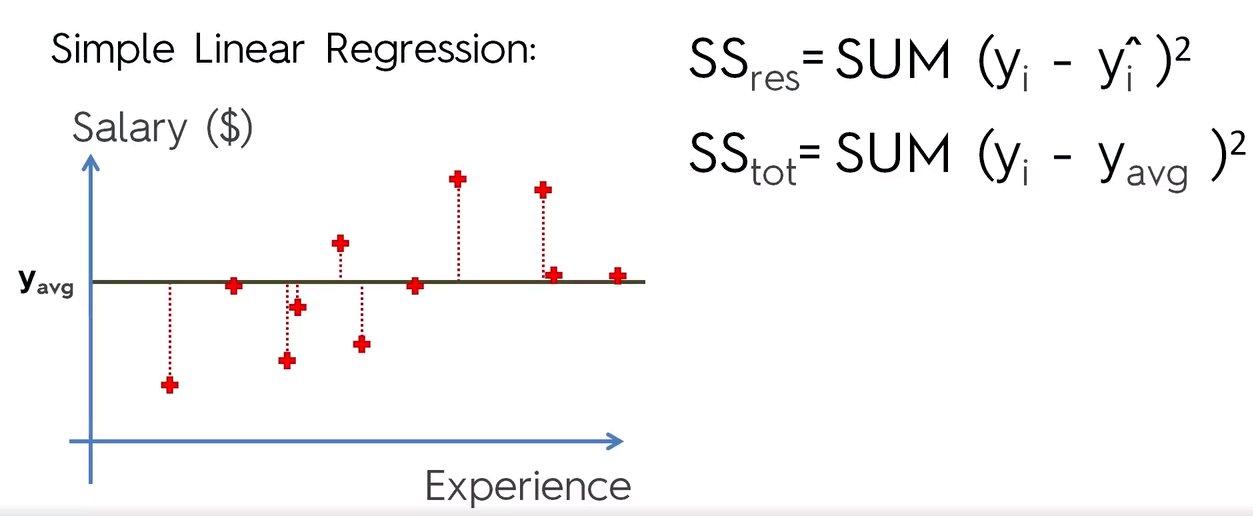
Find the sum of the square of the difference between the actual value shown as the Red Cross and the value derived by the linear regression formula represented as the green cross which is derived by drawing a vertical line along the y axis.

This value is known as the sum of square of residuals represented by **SS**res.

In linear regression, the idea is to draw a line that results in the minimum residual value.

Now let us replace the regression line with the average line that represents the average salary for all the participants in this data set.

That would look similar to the line **y**avg shown below.



Now similar to the sum of residual calculation, calculate the sum of square of difference between the actual value represented by the Red Cross and the average value **y**avg.

This value is known as the total of the sum of squares represented by **SS**tot.

R Squared



**The Value R Square is indicating how good the Regression line is in comparison to the average value.**

**R Square tending to 1 is considered as an ideal value.**

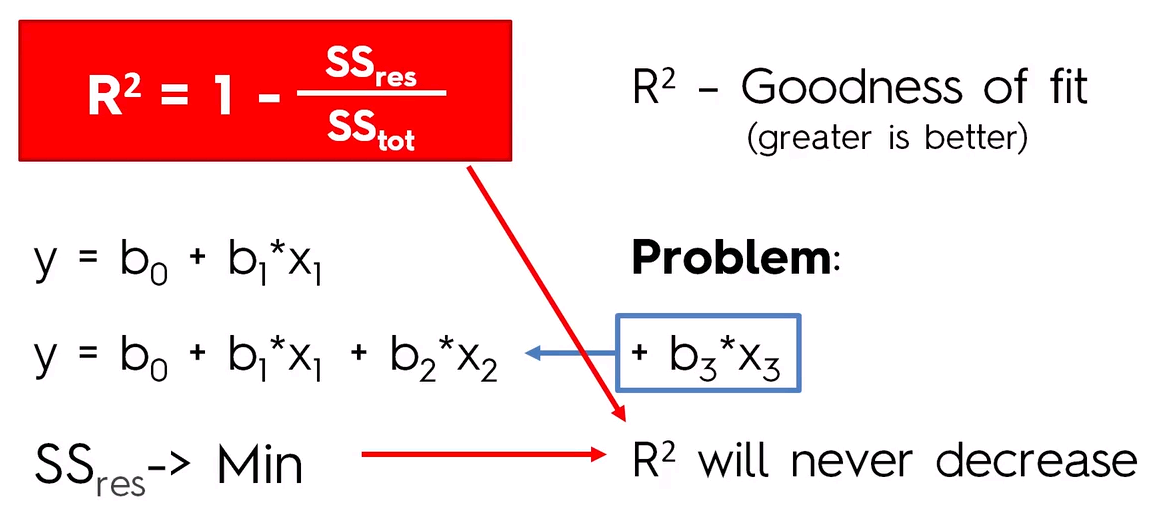
This means that the closeness of the value of Sum of Square of residual and the total of the sum of Square being closest to each other will result in the best R Square result.

R Square also tells how well the model fits to the data.

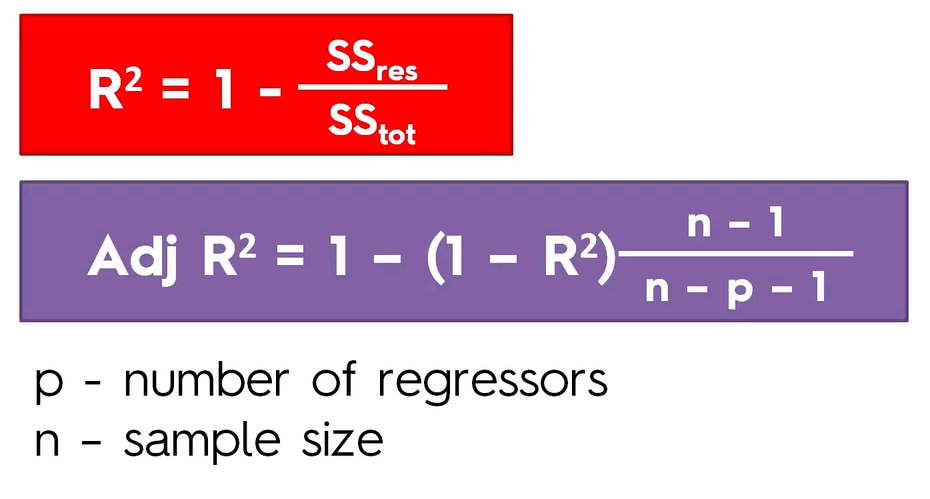
R Square and multiple linear regression.

Applying R Square calculation to multiple linear regression, where we are dealing with more than one variable causes the value of RSquare to increase further as per the calculations, which will simply mean that adding more IV’s will improve the R Square and hence can be misleading.

This shows that adjusted R2 doesn’t work when it comes to multiple linear regression.



Adjusted R Square



As shown here, adding more regressors (IV) will cause a penalization in the form of increase in value of P, thus causing the second part of formula after **“1-“** to have a lower value.

This will result in the value of Adjusted R square to tend towards 1.

This process of adjusting the R2 value based on the number of IV’s and sample size results in balancing the R2 and hence the formula is effective for handling multiple linear regression.

Introduction to GRETL

GRETL stands for GNU Regression, Econometrics and Time Series Library.

GNU Stands for General Public License which makes GRETL free to use and distribute.

Let us run a Simple Linear Regression

1. Open the sample data file Salary.csv in excel
2. Find the co-relation between salary and experience.
3. The company is a newly formed entity and the HR department wants you to come up with a proper model that can be used to set salaries for future employees and also based on the model, make necessary Salary corrections.
4. Run the Ordinary Least Square (Simple Linear Regression)
5. Observe the constant coefficients and p-value

Model 2: OLS, using observations 1-30

Dependent variable: Salary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *t-ratio* | *p-value* |  |
| **const** | **25792.2** | 2273.05 | 11.3469 | <0.0001 | \*\*\* |
| **YearsExperience** | **9449.96** | 378.755 | 24.9501 | **<0.0001** | **\*\*\*** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 76003.00 |  | S.D. dependent var | 27414.43 |
| Sum squared resid | 9.38e+08 |  | S.E. of regression | 5788.315 |
| R-squared | 0.956957 |  | Adjusted R-squared | 0.955419 |
| F(1, 28) | 622.5072 |  | P-value(F) | 1.14e-20 |
| Log-likelihood | −301.4412 |  | Akaike criterion | 606.8823 |
| Schwarz criterion | 609.6847 |  | Hannan-Quinn | 607.7788 |

***Coefficient constant=* 25792.2 -** This represents the starting salary at 0 years of experience.

***Coefficient* YearsExperience = 9449.96**

Average Increase in salary year after year as per model assumption, not as per real data

**P-Value** –

**It helps us to understand the statistical significance of individual variables**.

**It will help us filter out variables that are not predicting anything.**

Since the p value generated above is very small for **Years of Experience,** it means that

**NULL hypothesis of Years of Experience has no impact on the Salary does not hold good.**

**Rather the inverse or Alternate Hypothesis (HA) that unit change in Years of experience really impacts Salary has to be accepted !!!**

Analyzing the graphs

Below is the actual vs Predicted graph

Here we can observe the actual values in red and the predicted values represented by the linear regression with the blue line.



Forecasting

For 95% confidence intervals, t(28, 0.025) = 2.048

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Obs | Salary | prediction | std. error | 95% interval |
| 1 | 39343.0 | 36187.2 | 6096.56 | (23698.9, 48675.4) |
| 2 | 46205.0 | 38077.2 | 6077.17 | (25628.6, 50525.7) |
| 3 | 37731.0 | 39967.1 | 6058.67 | (27556.5, 52377.8) |
| 4 | 43525.0 | 44692.1 | 6016.33 | (32368.2, 57016.0) |
| 5 | 39891.0 | 46582.1 | 6000.99 | (34289.6, 58874.6) |
| 6 | 56642.0 | 53197.1 | 5954.57 | (40999.7, 65394.5) |
| 7 | 60150.0 | 54142.1 | 5948.87 | (41956.4, 66327.8) |
| 8 | 54445.0 | 56032.1 | 5938.19 | (43868.2, 68195.9) |
| 9 | 64445.0 | 56032.1 | 5938.19 | (43868.2, 68195.9) |
| 10 | 57189.0 | 60757.1 | 5915.64 | (48639.4, 72874.7) |
| 11 | 63218.0 | 62647.1 | 5908.30 | (50544.5, 74749.6) |
| 12 | 55794.0 | 63592.0 | 5904.99 | (51496.2, 75687.9) |
| 13 | 56957.0 | 63592.0 | 5904.99 | (51496.2, 75687.9) |
| 14 | 57081.0 | 64537.0 | 5901.92 | (52447.5, 76626.6) |
| 15 | 61111.0 | 68317.0 | 5892.05 | (56247.7, 80386.4) |
| 16 | 67938.0 | 72097.0 | 5886.08 | (60039.9, 84154.1) |
| 17 | 66029.0 | 73987.0 | 5884.55 | (61933.1, 86041.0) |
| 18 | 83088.0 | 75877.0 | 5884.00 | (63824.2, 87929.8) |
| 19 | 81363.0 | 81547.0 | 5888.19 | (69485.6, 93608.4) |
| 20 | 93940.0 | 82492.0 | 5889.74 | (70427.4, 94556.6) |
| 21 | 91738.0 | 90051.9 | 5910.88 | (77944.1, 102160.) |
| 22 | 98273.0 | 92886.9 | 5922.78 | (80754.7, 105019.) |
| 23 | 101302. | 100447. | 5965.00 | (88228.2, 112666.) |
| 24 | 113812. | 103282. | 5984.71 | (91022.8, 115541.) |
| 25 | 109431. | 108007. | 6022.19 | (95671.0, 120343.) |
| 26 | 105582. | 110842. | 6047.41 | (98454.3, 123229.) |
| 27 | 116969. | 115567. | 6093.92 | (103084., 128050.) |
| 28 | 112635. | 116512. | 6103.89 | (104009., 129015.) |
| 29 | 122391. | 123127. | 6179.70 | (110468., 135785.) |
| 30 | 121872. | 125017. | 6203.27 | (112310., 137724.) |

Graphical representation of Forecast interval shown above

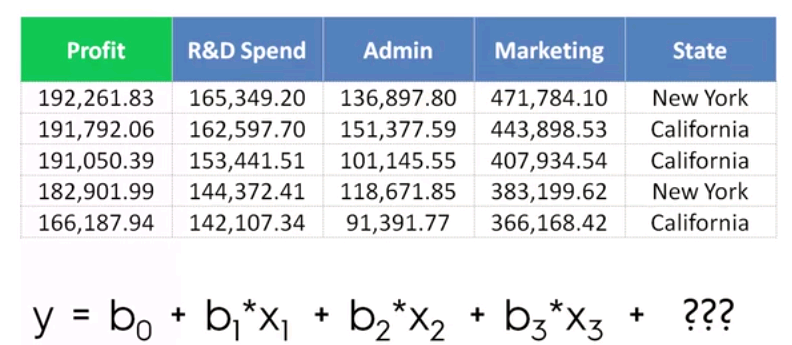
Here the green lines represent the forecast of salary over experience. The green lines are the intervals within which the salaries are being forecasted for the number of years of experience with 95% accuracy.



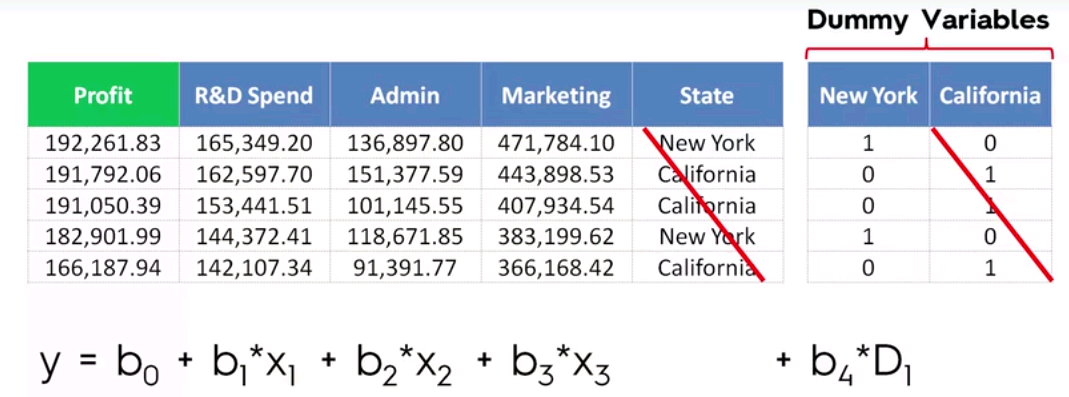
Let us look at Multiple Linear Regression example

1. Open the data file 50Startups.csv in excel and analyze
2. The challenge here is to analyze and create a model to explain, of the 50 companies, **investing in which types of companies would yield a better return on investment.**
3. **That will require us to analyze each of the Independent variables of R&D, Admin, Marketing and State and study their impact on Profit**

By analyzing the data in the excel as shown below, we are faced with the challenge of calculating based on a Qualitative variables

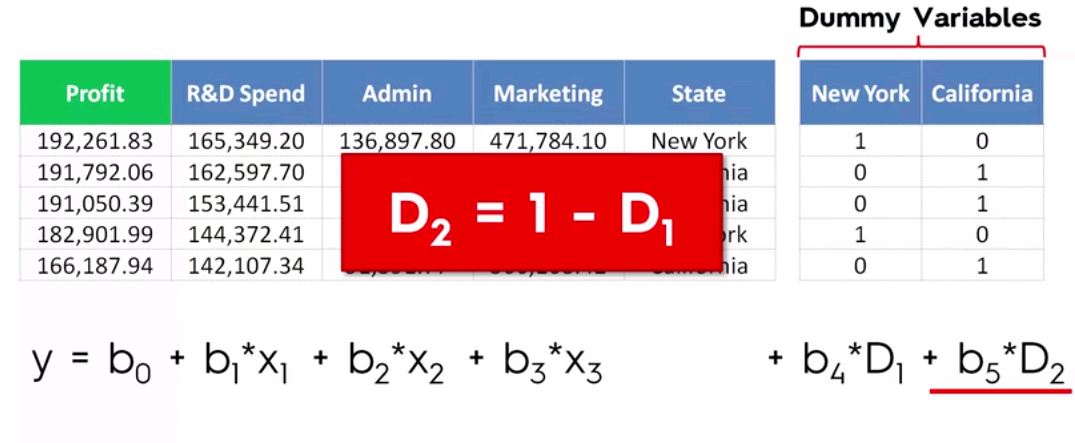


To counter this challenge, we will need to create a Dummy variable that would be Quantitative in nature so it can be used in the calculation and also represent the variable.



As you can see here, of the 2 dummy variables, only 1 has been used. **This is due to a phenomenon called multi-collinearity where inter dependent variables used in the equation can cause the model to behave incorrectly.**

While testing individual IV’s for their significance, it is important that the other IV’s in the equation are kept constant, but with collinearity this cannot be achieved, hence the below formula for dummy variables.



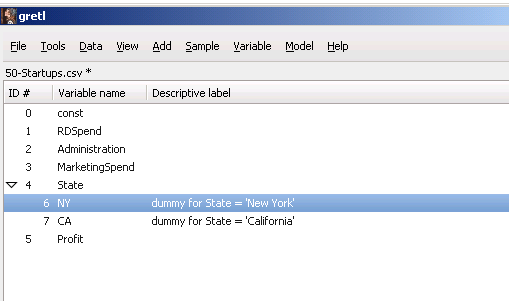
Building a Regression Model

There are 5 ways to build a model

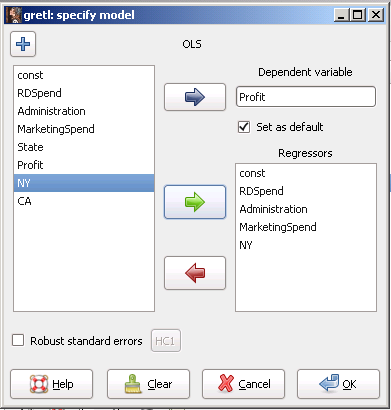
1. All –in
2. Backward Elimination
3. Forward Selection
4. Bi-directional Elimination
5. Score comparison
6. All –in (throw in all your variables)
   1. Use only if u have prior knowledge based on domain knowledge
   2. Some framework in your company or based on suggestion to use all variables
   3. Preparing for Backward elimination type of regression
7. Backward Elimination (Stepwise)
   1. Select a significance level to stay in the model (eg: SL: 0.05) – 5.0%
   2. Fit the full model with all possible predictors
   3. Consider the predictor with the highest P-Value. If P> SL, go to step 4, otherwise go to step 6.
   4. Remove the predictor that has the highest P-Value
   5. Fit model without the variable, refit the model. Repeat steps 3 through 4 until you find a P value that < SL.
   6. MODEL READY (Once all IV’s have P-values less than the significance level)
8. Forward Selection (Stepwise)
   1. Select a significance level to enter the model (eg: SL: 0.05) – 5.0%
   2. Fit all simple regression models y against **x**n. Select the one with the lowest P-value (take all dependent variable and fit it against all independent variables)
   3. Keep the variables and fit all possible models with one extra predictor added to the one you already have. (construct all possible 2 variable linear regression)
   4. Consider the predictor with the lowest P-value. If P< SL, go to Step 3 else step 5.
   5. MODEL READY
9. Bi-directional Elimination (Stepwise)
   1. Select a SL to stay SLSTAY(0.05) and a SL to enter SLENTER(0.05) the model.
   2. Perform the next step of forward selection (new var must have P < SLENTER to enter)
   3. Perform the next step of backward elimination (old var must have P<SLSTAY to stay)
   4. No new var enter and no old var can exit
   5. MODEL READY
10. Score Comparison - All possible Models
    1. Select a criterion of goodness of fit (eg: Akaike criterion)
    2. Construct all possible regression models 2 power n – 1 total combinations (E.g.: 10 columns with have 1023 models)
    3. Select the one with the best fit
    4. MODEL READY

Multiple linear regression using Backward Elimination

* With GRETL open the 50Startups.csv file.
* As discussed above, for State we need to create Dummy Variables
* As per Backward Elimination Step 1, let us set the Significance Level (**SL**) as 0.05.



* Fit the model with all possible predictors



Model 1: OLS, using observations 1-50

Dependent variable: Profit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *t-ratio* | *p-value* |  |
| const | 50416.5 | 6653.54 | 7.5774 | <0.0001 | \*\*\* |
| RDSpend | 0.807956 | 0.0457466 | 17.6615 | <0.0001 | \*\*\* |
| Administration | −0.02362 | 0.0518559 | −0.4555 | 0.6509 |  |
| MarketingSpend | 0.0263692 | 0.0166783 | 1.5810 | 0.1209 |  |
| NewYork | −1332.09 | 2690.18 | −0.4952 | 0.6229 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 112012.6 |  | S.D. dependent var | 40306.18 |
| Sum squared resid | 3.90e+09 |  | S.E. of regression | 9309.026 |
| R-squared | 0.951013 |  | Adjusted R-squared | 0.946659 |
| F(4, 45) | 218.4023 |  | P-value(F) | 7.53e-29 |
| Log-likelihood | −525.2499 |  | Akaike criterion | 1060.500 |
| Schwarz criterion | 1070.060 |  | Hannan-Quinn | 1064.140 |

Above you can see that with Administration, the coefficient is negative. For every dollar spent on administration, company loses 2.3 cents so need to look at reducing administrative spending.

Observe the graph, fitted vs Actual against Administration.

As it can be observed below, the Profit does not have a linear distribution based on the Administrative expenses.



* Repeat Step 3. Remove the predictor with the highest P-Value that is greater than SL. So we will remove the Administration IV.

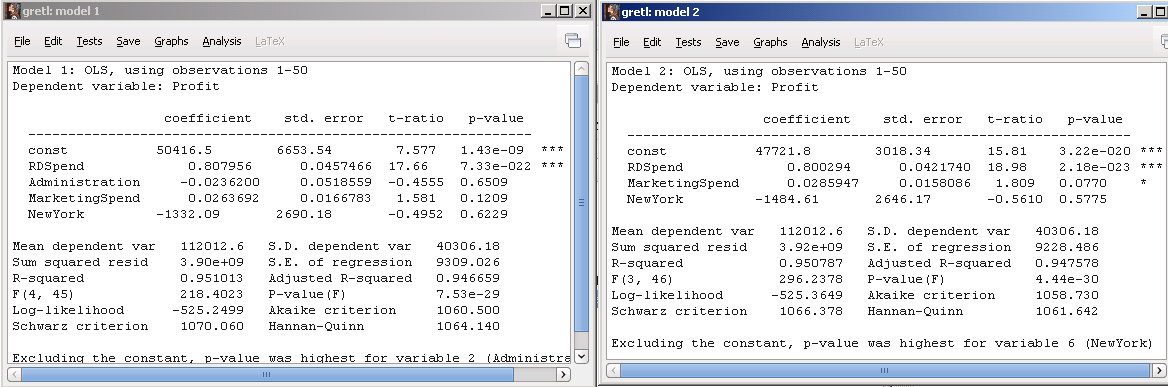
Model 2: OLS, using observations 1-50

Dependent variable: Profit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *t-ratio* | *p-value* |  |
| Const | 47721.8 | 3018.34 | 15.8106 | <0.0001 | \*\*\* |
| RDSpend | 0.800294 | 0.042174 | 18.9760 | <0.0001 | \*\*\* |
| MarketingSpend | 0.0285947 | 0.0158086 | 1.8088 | 0.0770 | \* |
| NewYork | −1484.61 | 2646.17 | −0.5610 | 0.5775 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 112012.6 |  | S.D. dependent var | 40306.18 |
| Sum squared resid | 3.92e+09 |  | S.E. of regression | 9228.486 |
| R-squared | 0.950787 |  | Adjusted R-squared | 0.947578 |
| F(3, 46) | 296.2378 |  | P-value(F) | 4.44e-30 |
| Log-likelihood | −525.3649 |  | Akaike criterion | 1058.730 |
| Schwarz criterion | 1066.378 |  | Hannan-Quinn | 1061.642 |

Analysis of both models side by side.

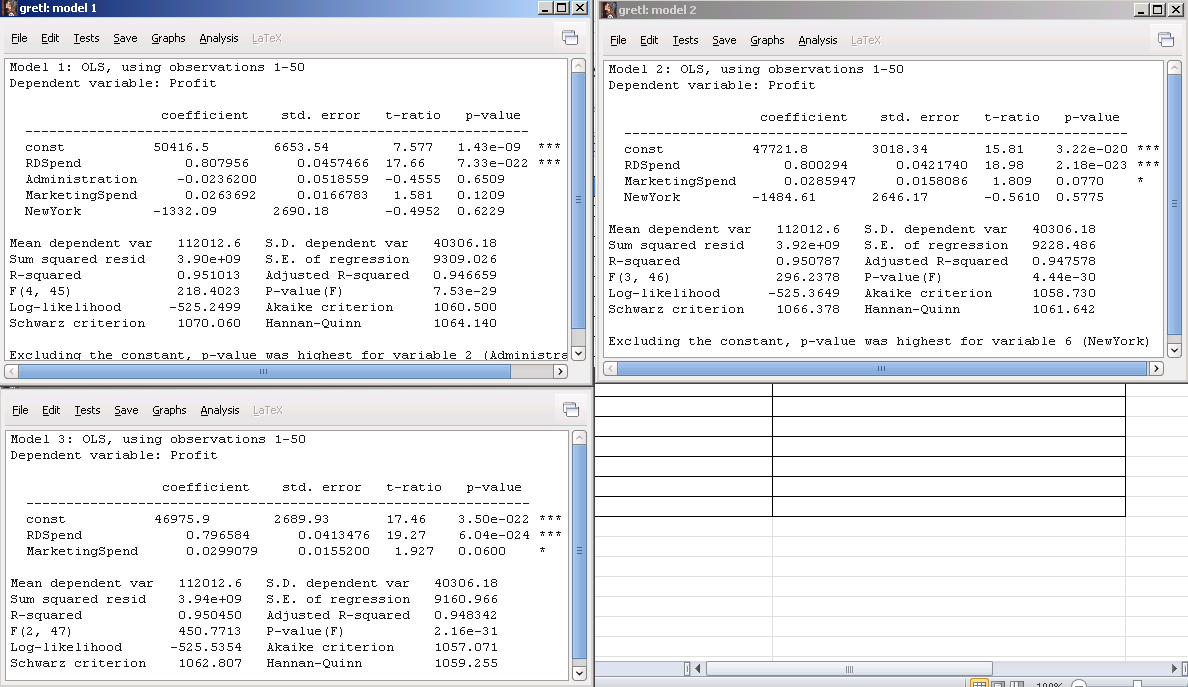


Observations

* In the right model, a star has appeared for Marketing Spend and the P-value is now closer to the SL of 0.05.
* Now the State – New York (dummy value) has the highest P-value.
* For New York, the profit is lower than 1484 $ compared to the other region, that is California.
* As shown in graph below, the plot does not show any difference to Profit being in New York or California.
* If there was a substantial difference then we would observe more red crosses in any one of the plots but here we can see the red crosses are almost evenly distributed between both States.



* Repeat step 3 by removing the IV New York.



Model 3: OLS, using observations 1-50

Dependent variable: Profit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *t-ratio* | *p-value* |  |
| Const | 46975.9 | 2689.93 | 17.4636 | <0.0001 | \*\*\* |
| RDSpend | 0.796584 | 0.0413476 | 19.2656 | <0.0001 | \*\*\* |
| MarketingSpend | 0.0299079 | 0.01552 | 1.9271 | 0.0600 | \* |

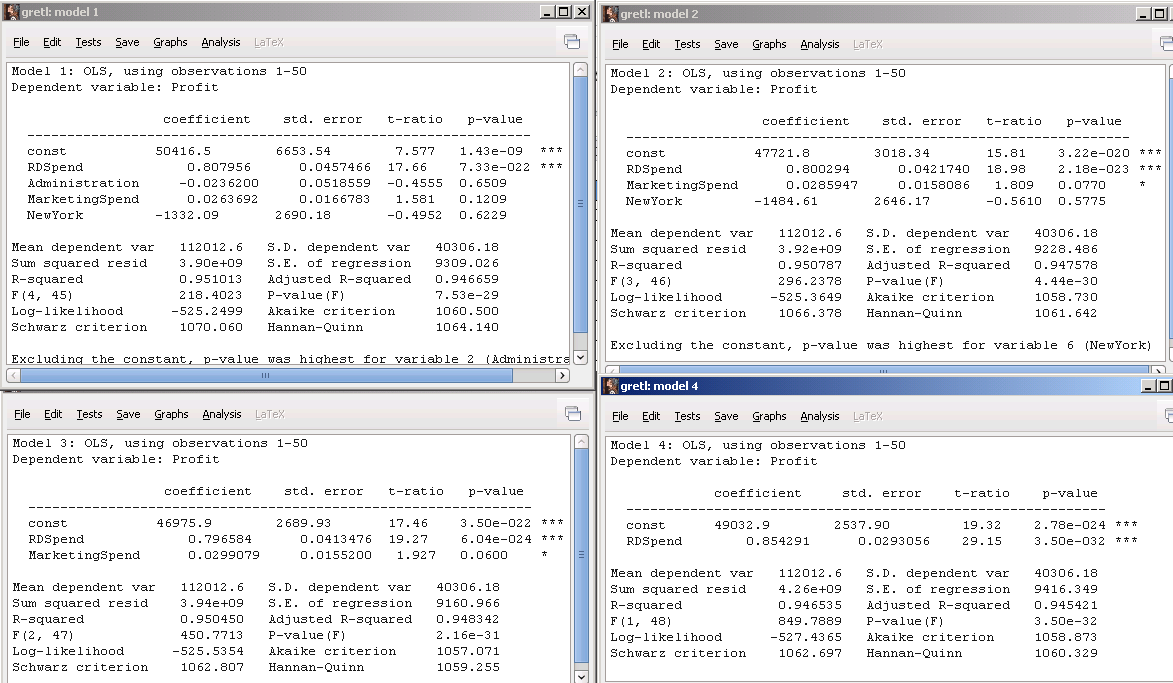
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 112012.6 |  | S.D. dependent var | 40306.18 |
| Sum squared resid | 3.94e+09 |  | S.E. of regression | 9160.966 |
| R-squared | 0.950450 |  | Adjusted R-squared | 0.948342 |
| F(2, 47) | 450.7713 |  | P-value(F) | 2.16e-31 |
| Log-likelihood | −525.5354 |  | Akaike criterion | 1057.071 |
| Schwarz criterion | 1062.807 |  | Hannan-Quinn | 1059.255 |

Observations

* P-Value for Marketing spending is now closer to the SL value of 0.05.
* Though below we can observe the linear distribution is broken at places while comparing profit vs Market spending.



* Repeat step 3 by removing the IV Marketing Spending as it is still not below the SL.



Model 4: OLS, using observations 1-50

Dependent variable: Profit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *t-ratio* | *p-value* |  |
| Const | 49032.9 | 2537.9 | 19.3203 | <0.0001 | \*\*\* |
| RDSpend | 0.854291 | 0.0293056 | 29.1511 | <0.0001 | \*\*\* |

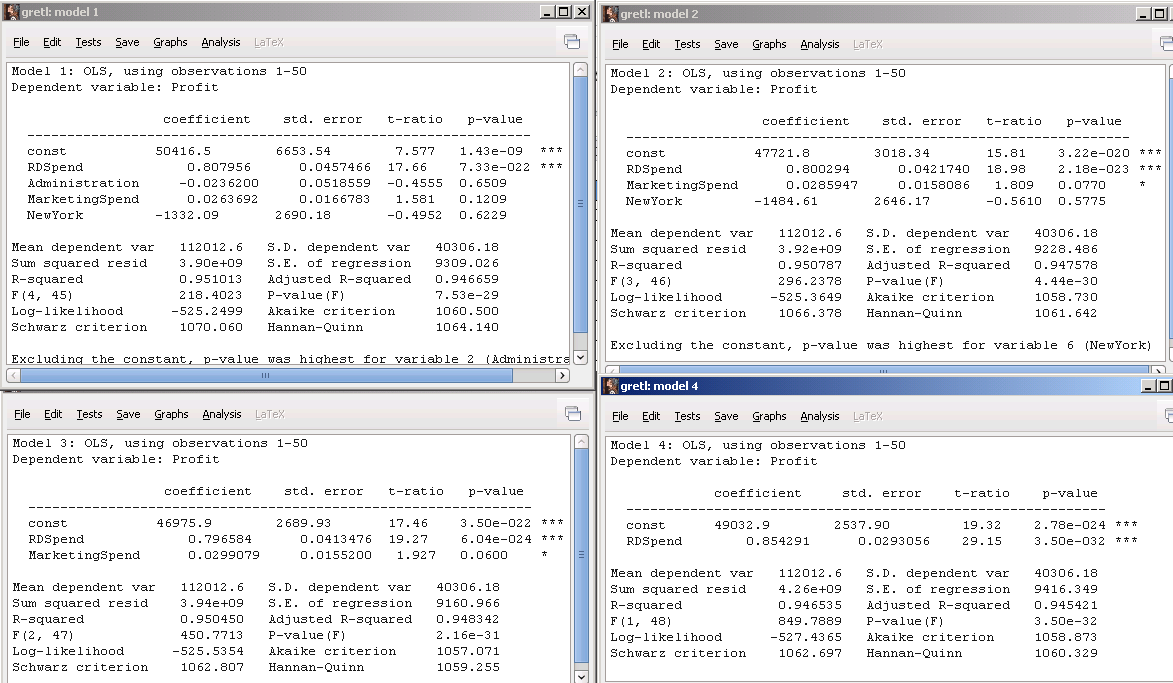
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 112012.6 |  | S.D. dependent var | 40306.18 |
| Sum squared resid | 4.26e+09 |  | S.E. of regression | 9416.349 |
| R-squared | 0.946535 |  | Adjusted R-squared | 0.945421 |
| F(1, 48) | 849.7889 |  | P-value(F) | 3.50e-32 |
| Log-likelihood | −527.4365 |  | Akaike criterion | 1058.873 |
| Schwarz criterion | 1062.697 |  | Hannan-Quinn | 1060.329 |



* As per the backward elimination steps, we have completed the model and we have been able to identify the best regressor that fits the model and is the best reflector of Profit.
* If we can take a step back though and think from the perspective of a business stakeholder, you would think, “Yes we are fine with reducing Administrative costs but what about all those marketing campaigns that generated so much profit”.
* This is where we need to think beyond the arbitrary approach of eliminating IV’s and see if there is another approach.

Adjusted R square Analysis

* R – Square helps determine the goodness of fit of the model. The closer it tends to 1, the better it is considered.
* As you see in above examples the more variables that get added, more the R-Square grows. So R Square is considered biased.



This is where **Adjusted R square** comes into play.

As you can see in the above scenarios, adjusted R square is the best in Model 3 where it has the marketing spending regression included compared to in Model 4, where the adjusted r square value reduces substantially by the removing of Marketing Spending.

This shows that Marketing spending plays a vital role in the goodness of fit of determination of profit along with RnD Spending.

So it is important to observe the adjusted R square along with P-Value. If adjusted R square is reducing it is necessary to stop and reflect on the backward elimination and adjust the variables accordingly.

Akaike criterion is another criterion to observe. This criteria shows a better fitting model when the value is lower.

Interpreting Coefficients

Coefficients determine the impact on the dependent variable by the per unit increase in value of the independent variable.

The positive coefficient value impacts the dependent variable in a positive way so is the vice versa.

But we also need to consider the **magnitude of the coefficients** before coming to a conclusion on the values.

Let us say in the below scenario, the RnDSpend and Marketing spend where not similar unit of measure, that is rndspend was determined in the unit of 1000 $ and Marketing in actual $ so in that case it would reflect totally different impact.

Always say magnitude in terms in terms of Unit of measure.

So the right way to represent is “RnD spend has a greater impact based on per unit spent compared to per unit of Marketing spend.”

So in the below scenario, a unit spend of 1$ in RnD spending, there will be an increase in profit of 79 cents.

A unit spending of 1$ in Marketing spending, there will be an increase in profit of 2.9 cents.

In model 1, For New York, the negative coefficient indicates that the profit is lower than 1332 $ compared to the other region, that is California but since the associated p-value does not indicate the IV to be statistically significant, we would need to ignore the coefficient.

Note: Coefficients always add an additional effect to the value of previously added IV’s.

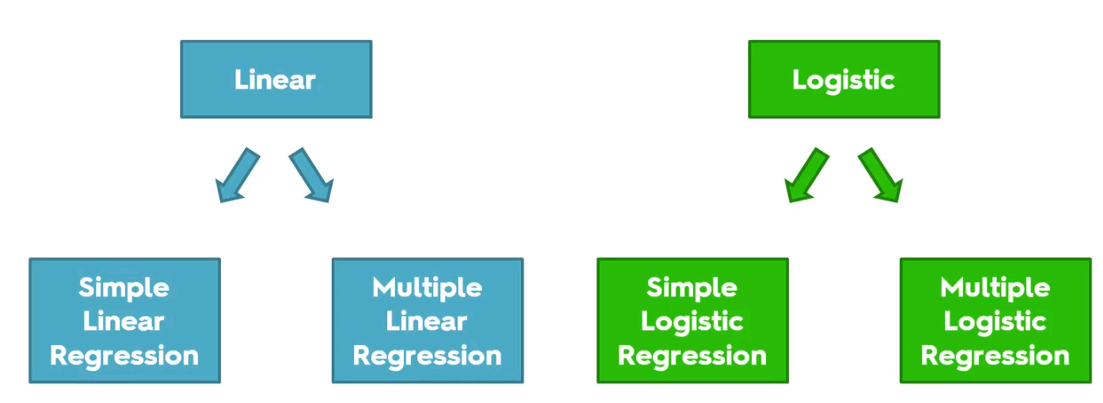
Business Conclusion

***If you are a venture capitalist, the analysis is that you should invest on companies that spend more on RnD and marketing to yield better profits or invest in companies that have a strong focus on RnD.***

***Here you can also interpret that marketing spending impact might be lower due to increased cost with respect to the marketing activity itself which impacts the profit margin compared to RnD.***

***This is also evident in the plot where the profit is not consistent or the linear distribution is unevenly disturbed at some sections of marketing spending.***

Logistic Regression



Logistic Regression is used when the Dependent Variable is Qualitative in nature, which means that expected result from the Regression process is a Straight Yes or No, so this is a process of modeling a Binary outcome

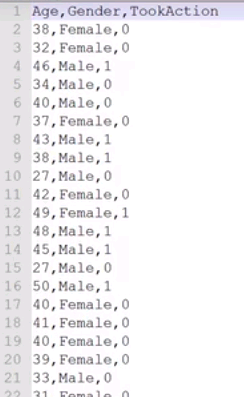
Logistic Regression tries to

* ***Model the probability of an event*** occurring depending on the values of the independent variables, which can be qualitative or quantitative in nature.
* ***Estimate the probability of an event*** occurring for a randomly selected observation versus the probability that the event does not occur.
* ***Predict the effect of a series of variables*** on a binary response variable
* ***Classify observations by estimating the probability*** that an observation is in a particular category.

Let us analyze the EmailOffer.csv data file.

As we can see this data provides some information about email offers sent to Men and Women of various ages and their response to the offer by taking action. This action can mean different things. E.g.: Whether they clicked on a link in the email or they purchased the product offered.

So TookAction is a categorical variable (that is a binary 0/1).



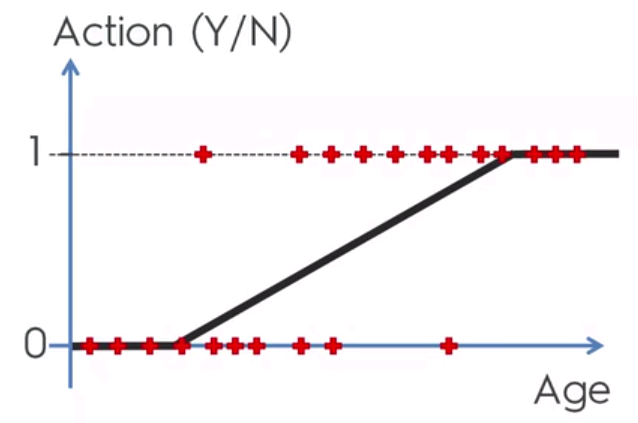


Observations:

* As shown below while plotting age against tookaction, it shows that the linear regression line cannot accurately estimate the model.
* On the y axis, the regression line is plotting values starting below 0 as well as above 1
* This model is not completely illogical. Here we can see that people in the age group below 40 are more skewed towards left and above age group of 40 are skewed to right.



If we normalize the extreme values plotted on the positive and negative scale, the regression line should look like below.



* We will come across situations where we need to predict if something is going to happen or not.

E.g.: If a drug is helping a patient or not.

A simple A-B test is not sufficient here as there are 100’s of independent variables like Age, geography, medical history and genetic makeup/ history that needs to be analyzed.

* Another example is in customer centric organizations where we need to categorize/ segment the customers based on how they respond to offers, based on their behavioral patterns like they use mobile or visit stores to purchase / browse products.

Probability

***Describes the possibility of an outcome of interest***

P = outcome of interest

all possible outcomes

***Fair coin flip***

P (***heads***) = 1/2 = 0.5

***Fair die flip***

P (***1 or 2***) = 2/6 = 1/3 = 0.333

***Deck of playing cards***

P (***diamond card***) = 13/52 = 1/4 = 0.25

Odds

***Describes the probability of an event occurring against the probability of an event not occurring.***

Odds = p (occurring)

p (not occurring)

Odds = p / 1-p

Going back to the previous example, apply the probabilities derived to this odds function

Fair coin flip

Odds (***heads***) = 0.5/ (1-0.5) = 1 or 1:1 (ratio)

Fair die flip

Odds (***1 or 2***) = 0.333/ 0.666 = 1/2 = 0.5 or 1:2 (ratio)

Deck of playing cards

Odds (***diamond card***) = 0.25/ 0.75 = 1/3 = 0.333 or 1:3 (ratio)

Odds Ratio

***Describes the ratio between two odds***

***Fair coin flip***

P (***heads***) = 1/2 = 0.5

Odds (***heads***) = 0.5/ (1 - 0.5) = 1 or 1:1 (ratio)

***Loaded coin flip***

P (***heads***) = 7/10 = 0.7

Odds (***heads***) = 0.7/ 0.3 = 2.333

***Odds Ratio = odds1 / odds2 = (p1/1-p1)/ (p0/1-p0)***

***Odds Ratio =*** (0.7/ 0.3) **/** (0.5/ 0.5) = ***2.333***

**Conclusion**

***The odds of getting a Heads on the loaded coin is 2.333 times higher than the fair coin***

Odds Ratio in Logistic Regression

***Odds ratio for a variable in Logistic regression represents how the odds change with a 1 unit increase in that variable holding all other variables constant.***

Note: The odds can have a larger magnitude even if the underlying probabilities are low.

Logistic Regression explained Step by Step

* ***The dependent variable in logistic regression follows the Bernoulli distribution having an unknown probability, p.***
* ***Bernoulli distribution is just a special case of Binomial Distribution where n = 1 (just one trial). Success = “1” and Failure =” 0”.***
* ***So if the probability of success is p, then Failure is q= 1 – p.***
* ***In Logistic regression we are estimating an unknown p for any given linear combination of the independent variable.***
* ***We are trying to LINK the independent variables that can be quantitative or qualitative in nature to the outcome which is an unknown Probability p.***
* ***The independent variables need to be linked together to the Bernoulli distribution and that Link is called Logit.***
* **LOGIT**

Unlike in Binomial distribution where we know the probability p, the goal of Logistic regression is to estimate p for a linear combination of independent variables.

Estimate of **p** is **p- hat** represented by the symbol 

As discussed in section above, we need a Link function to tie together the linear combination of variables to the binomial (Bernoulli) distribution that would result in a domain from 0 to 1.

The natural log of the odds ratio, the logit, is that link function.

**Ln (odds)**

This can be rewritten in its natural form as

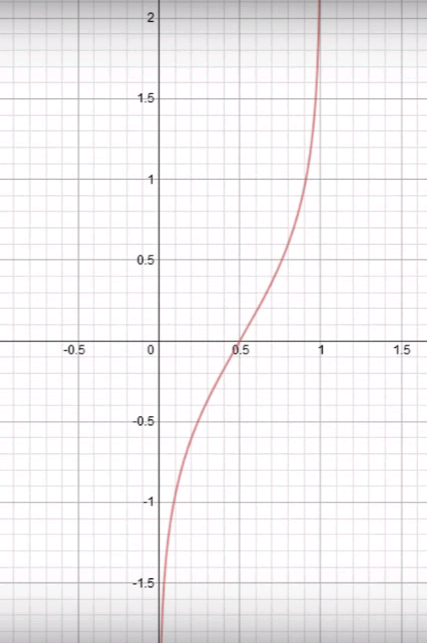
**ln (p / (1-p)) This is called the Logit(p)**

**This can also be written as**

**ln (p) – ln (1 – p) = Logit(p)**

Note: **Log**e x is the same as ln x

By plotting ln (p / (1-p)) on a graph, this is the result.



**Example**:

**At x = 0**

ln (0 / 1 -0) = ln (0 / 1) = ln(0) = Unidentified (We can see for the probability of 0, the graph tends to 0, but does not really reach it).

**At x = 1**

ln (1 / 1 -1) = ln (1 / 0) = ln(UNF) = Unidentified (We can see for the probability of 1, the graph tends to 1, but does not really reach it).

***So these 2 examples above are called the log odds as they define the probability***

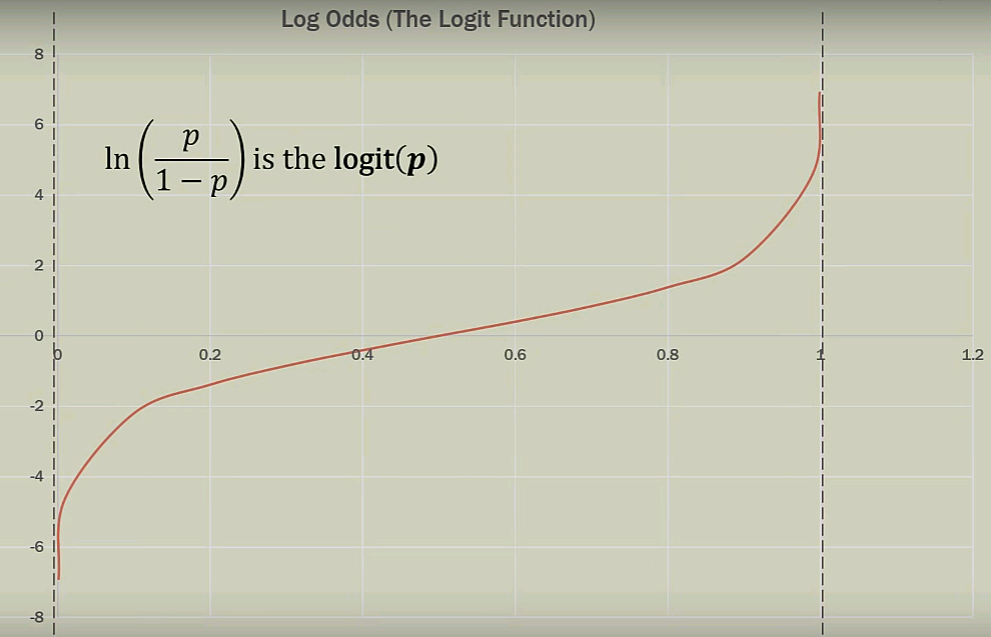
**At x = 0.5**

ln (0.5 / 1 -0.5) = ln (0.5 / 0.5) = ln(1) = 0

***So this explains that when the odds are even, then the Logit is “0”.***

* **Sigmoid Function**

If we spread the frequency distribution of x values a little more on the above graph, then we would see the graph below makes an **S** curve or also known as a **Sigmoid curve**.



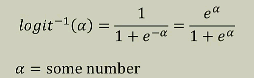
* **The Inverse Logit**

By observing previous plotting of graphs in this material, it can be observed that the probabilities or the Dependent Variables are always plotted against the Y – axis unlike the signmoid curve above.

To do this, we have to take the Logit Link function and run 0 and 1 probabilities against the Y – axis.

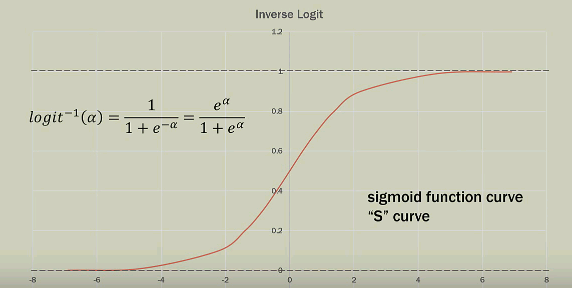
**Logit(p) = ln (p / (1-p))**

**After inverse becomes**



The “some number” will be the linear combination of variables and their coefficients.

Also note that the trend of the inverse logit will return the probability of being a “1” or in the “event occurs” group (that something happened than saying it did not happen)



So by summarizing the above steps, we derive the Logistic equation as shown below.



Logistic Regression Equation!!

Apply the value of y to sigmoid function

**Note**:

Alpha in the inverse logit function has been replaced with y to indicate plotting against Y-axis to get the S curve.

Going back to the definition of Logistic regression

***In logistic regression we are estimating an unknown p for any given linear combination of the independent variable***

So we are trying to estimate a probability

To do this let us solve the logistic regression equation further by taking an antilog of the equation.

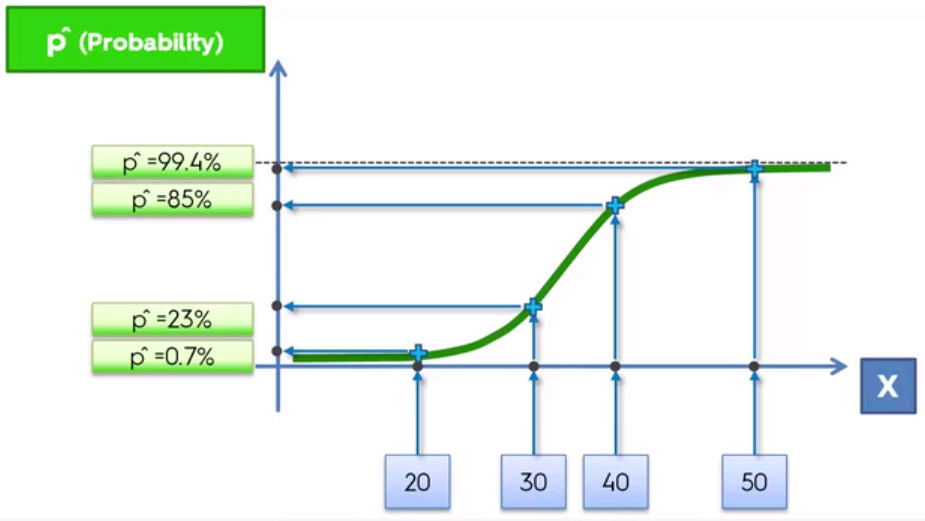
That gives us

**p / (1-p)** = **e b0+ b1x**

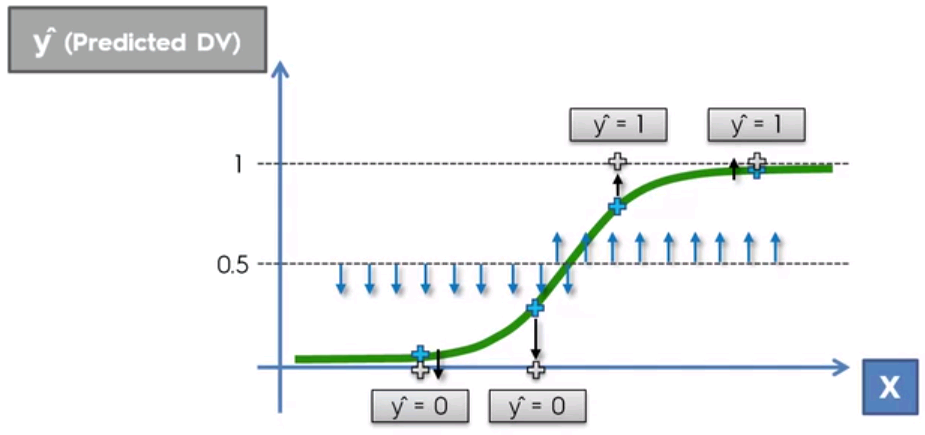
**As we continue to solve for p, we arrive at the p^ (estimation of probability) also known as the Estimated Regression Equation**

**p^ = e b0+ b1x / (1 + e b0+ b1x )**

Graphical representation of the above formula for p^can be interpreted as below.



If we want to just predict the y (y ^) values instead of the percentage of the probabilities then as shown below for any value below the median (0.5) on the curve, y ^= 0 and all values above curve, then y ^= 1.



False Positive and Negatives



**Y (actual DV)**

**p** (predicted values)

0.5

Observations:

* The blue crosses form the logistic regression “S” curve predicting the probability of people taking action for every unit change in age.
* As discussed previously, for y^ any predictions (blue cross) below the 50% line would mean that email offer will not be taken and anything above the 50% line, offer will be taken.
* The blue arrow pointing up indicate the prediction that the particular person will not take the offer and the blue arrow pointing down predicts that the person will take the offer and both predictions are correct (as the predictions are within the 50% of either side).
* The red arrow on the right where the prediction points to the offer being taken is refferred to as ***False Positive.* (Type I error)**
* The red arrow on the left where the prediction points to the offer not being taken is referred to as **False Negative. (Type II error)**

**Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Y^** | |
|  |  | **0** | **1** |
| **Y** | **0** | **35** | **5** |
| **1** | **10** | **50** |

**Observation:**

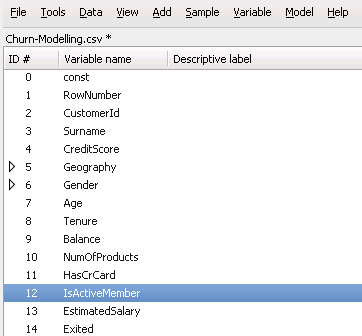
* For actual **Y** value of 0 (offer not taken) , predicted **Y^ shows** correct results for 35 people .
* Simiarly for actual **Y** value of 1 (offer taken) , predicted **Y^ shows** correct results for 50 people.
* For actual **Y** value of 0 (offer not taken) , predicted **Y^ shows** incorrect results for 5 people. These are False Positive or Type I errors.
* For actual **Y** value of 1 (offer taken) , predicted **Y^ shows** incorrect results for 10 people. These are False Negative or Type II errors.
* Accuracy rate = total correct/ total values = 85/100 = 0.85. Model is 85% accurate.

**Geo-demographic Segmentation**

**Lets analyze the Churn-Modeling.csv using excel**

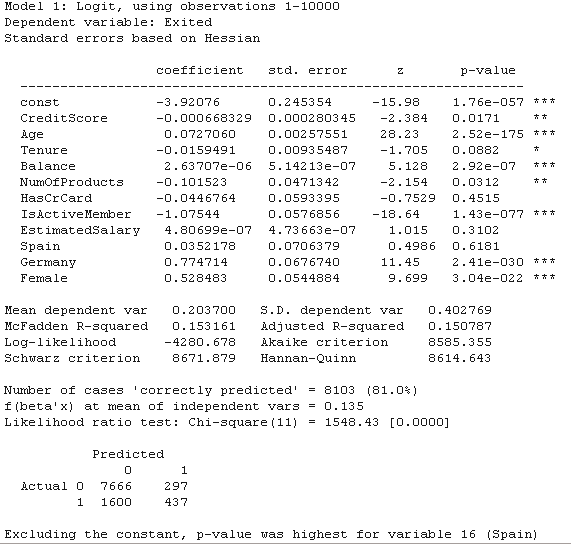
**Open the file with GRETL**

**Create the necessary dummy variables**



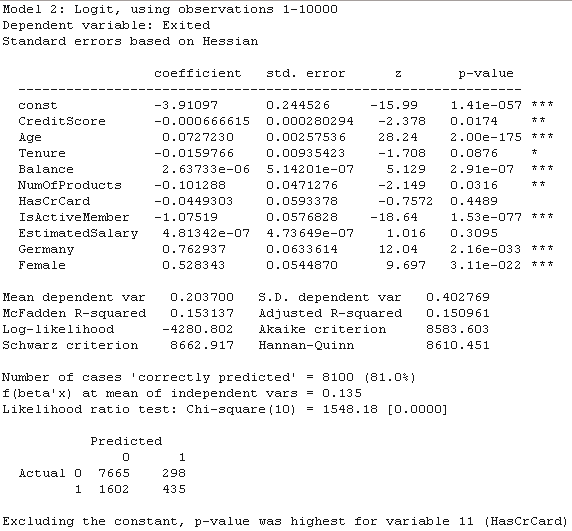
**As in previous exercise, use backward elimination to derive the best possible Logistic Regression model.**

Analyze the variables and select a set of IV’s that can best determine the dependent variable.

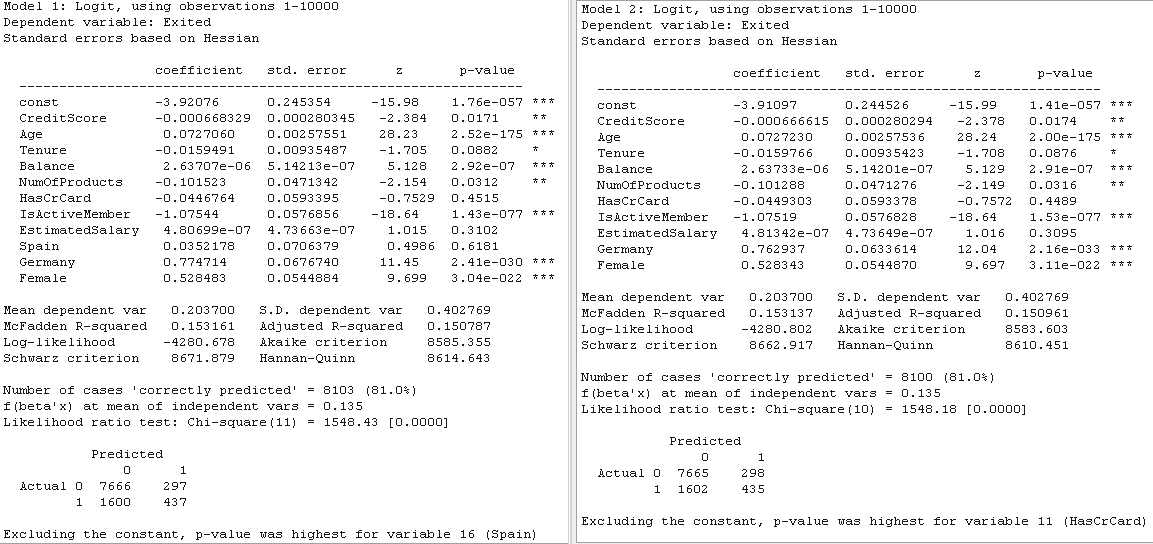


Observe the p-values of all IV’s, the r Squared and Adjusted R – Square

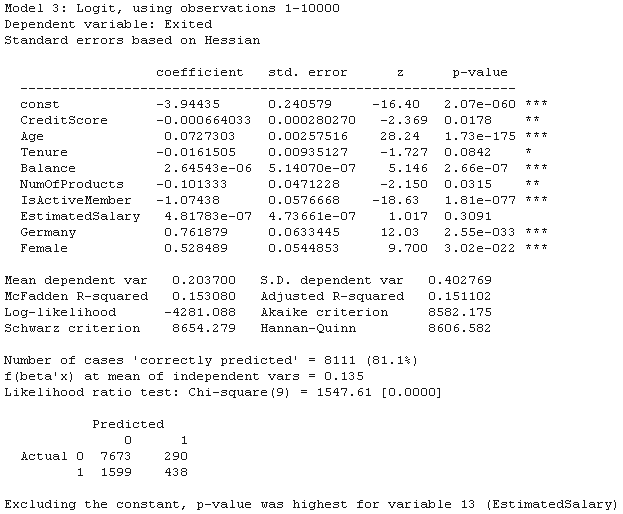
As per Model suggestion, eliminate the IV – Spain and rerun Backward elimination



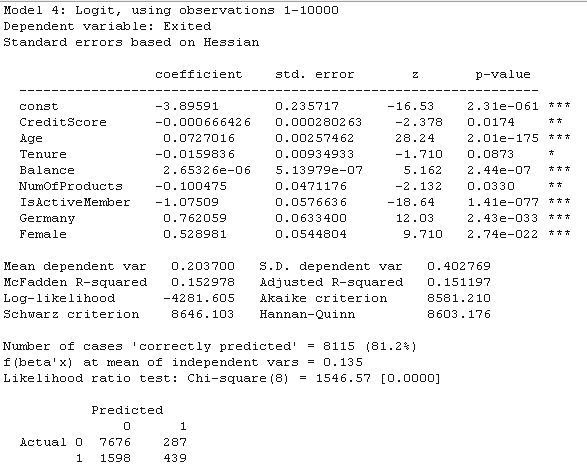
Compare the p values, coefficients and R squared against the previous model



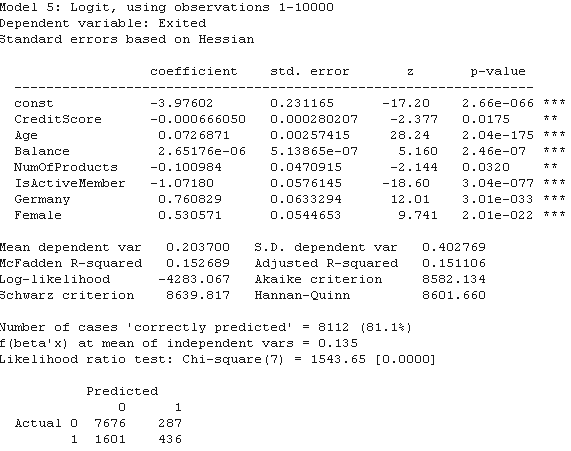
As suggested proceed with backward elimination



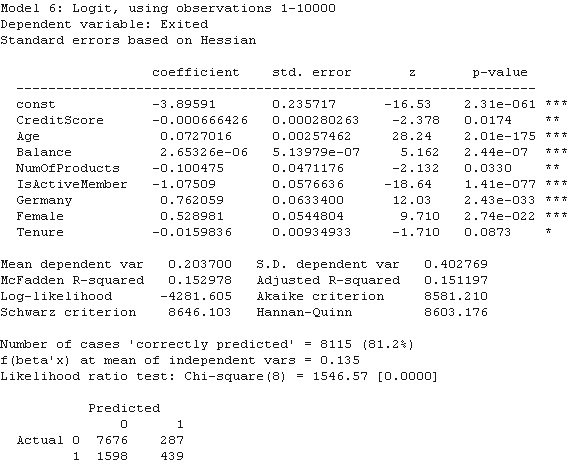
Compare with previous models and run the backward elimination by removing the IV with highest p-value



Since the IV Tenure has highest p-value and still below the 5% Significance Level, lets eliminate this.



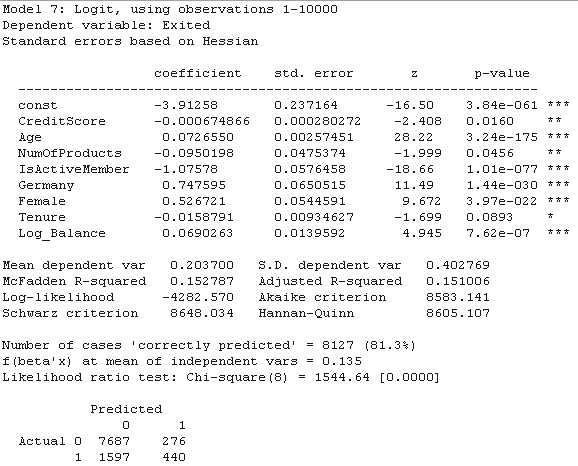
Make the observations.



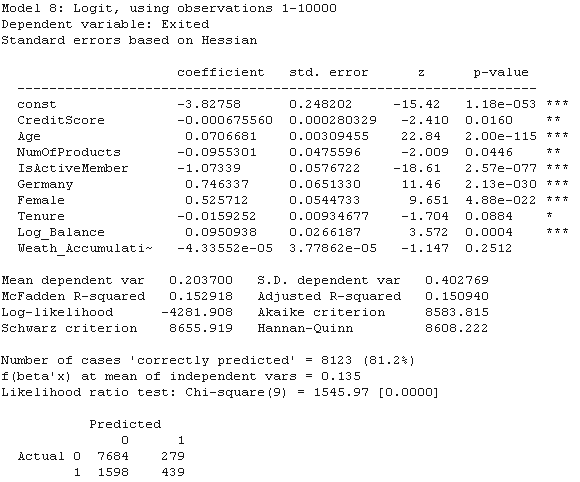
Create new variable Log Balance.

Explain Log concepts

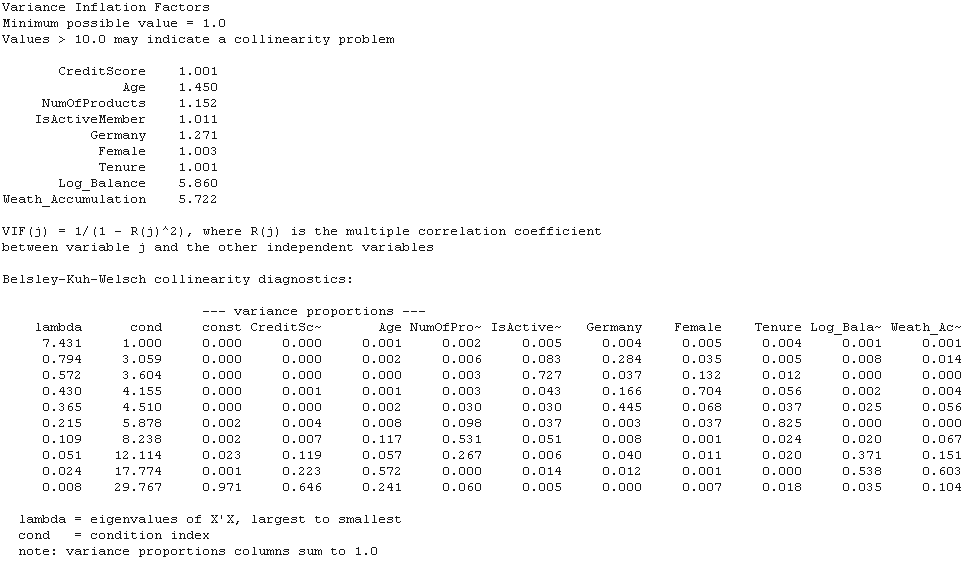
Add the Log Balance variable in Model



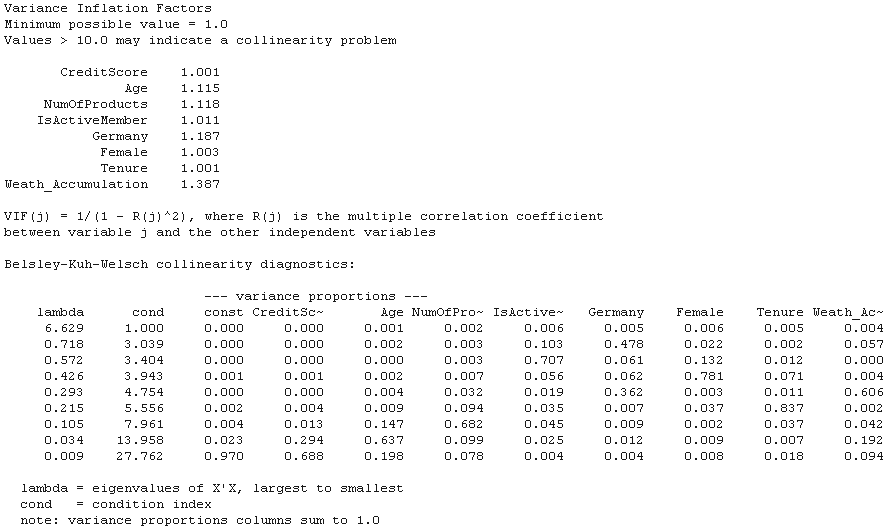
Create Wealth Accumulation calculation and add to model



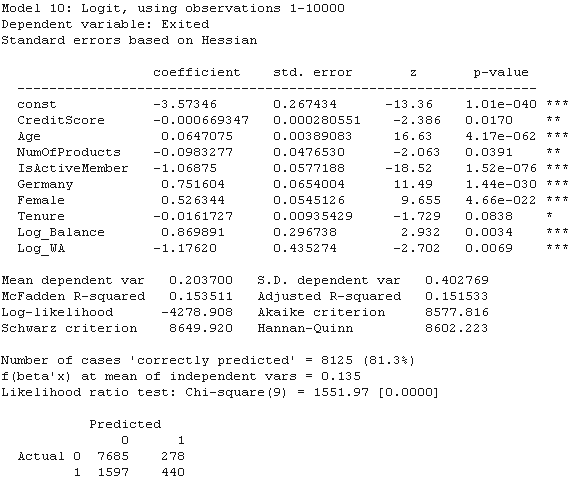
Check colinearity between Log Balance and Wealth Accumulation

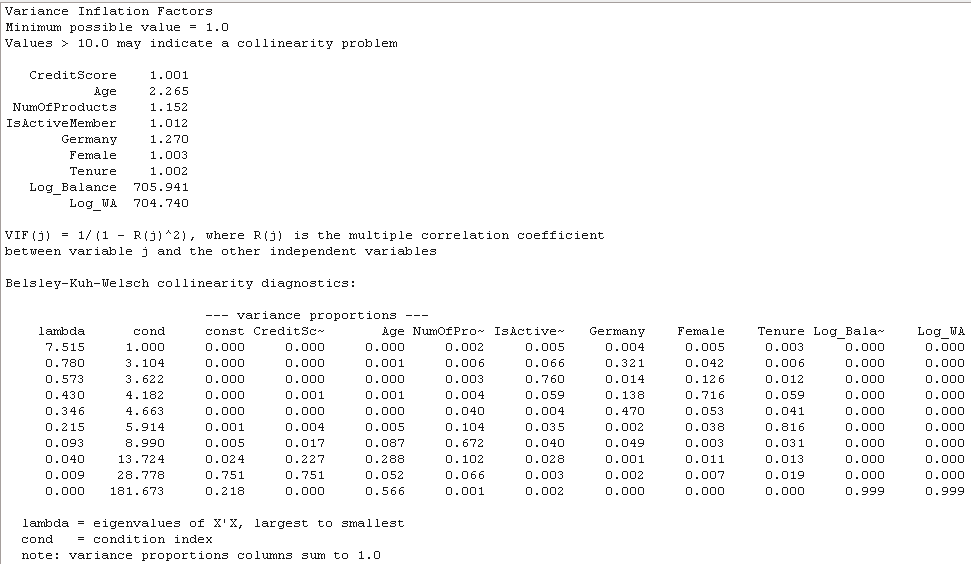






Create the Log of Wealth Accumulation

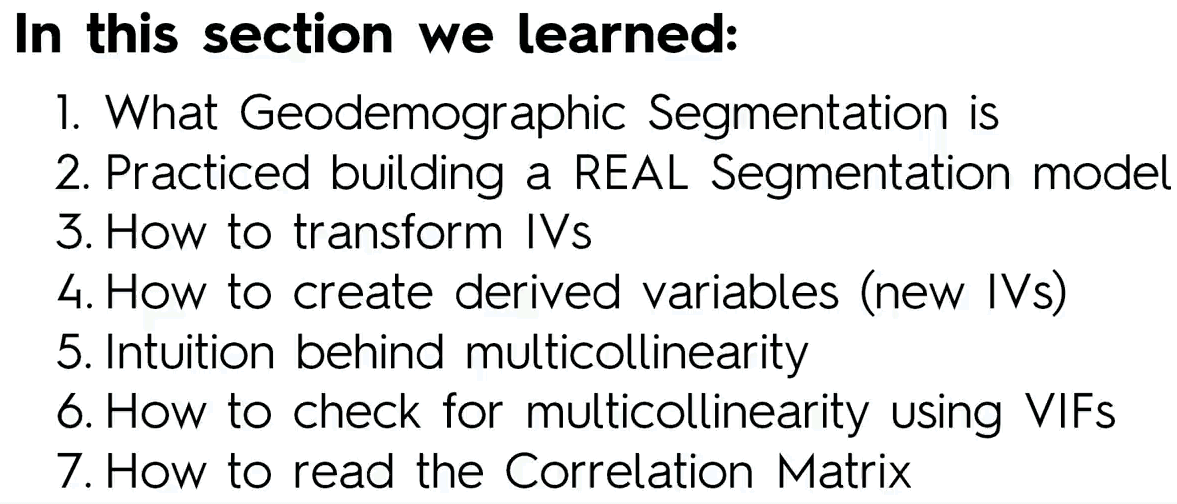




Correlation coefficients, using the observations 1 - 10000

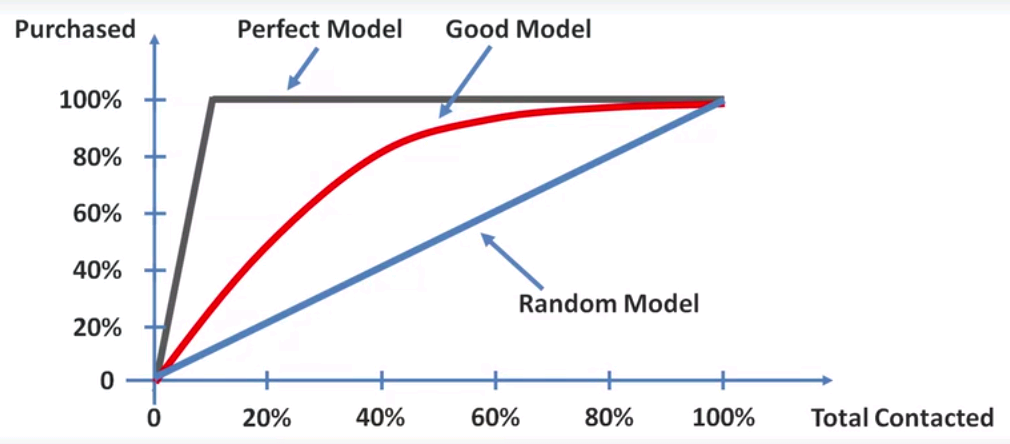
5% critical value (two-tailed) = 0.0196 for n = 10000

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Log\_WA | Weath\_Accumulation | Log\_Balance | Age |  |
| 1.0000 | 0.8889 | 0.9984 | -0.0075 | Log\_WA |
|  | 1.0000 | 0.8651 | -0.2463 | Weath\_Accumulation |
|  |  | 1.0000 | 0.0345 | Log\_Balance |
|  |  |  | 1.0000 | Age |



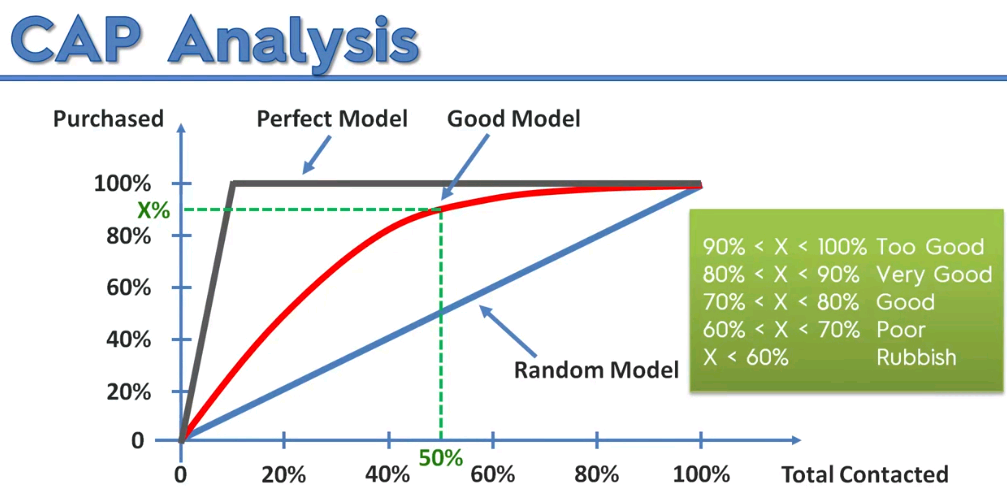
Model Assessment

Cumulatve Accuracy Profile (CAP)



Creating a CAP curve using excel

Use the CAP curve template



Overfitting

If the model is predicting an x% of 90 to 100%, there could be some forward looking variables and the model might not function as it should when you apply a testing data or a new set of data to the model.

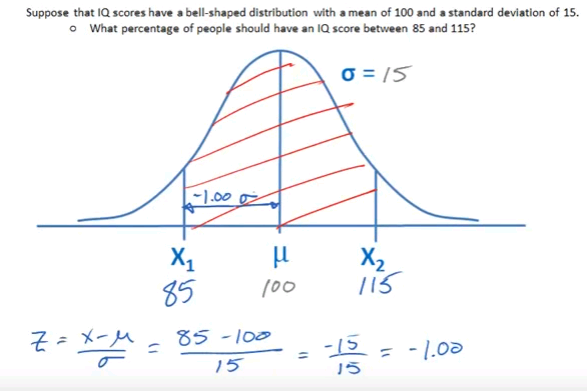
Seperating Training and Test data sets

**Insights from CAP**



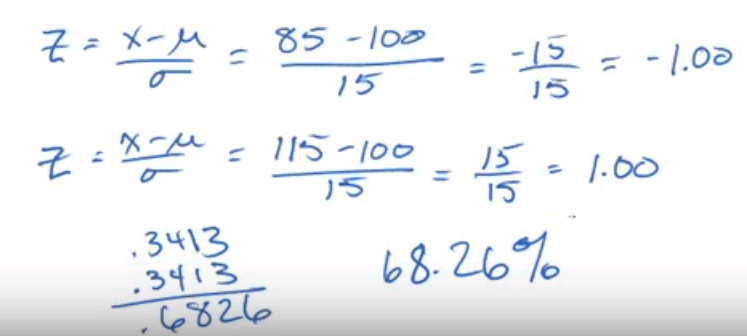
**Addition Reference Material**

**Z Score**



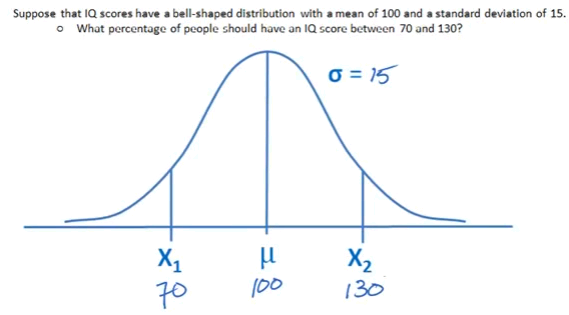
Referring to the z table

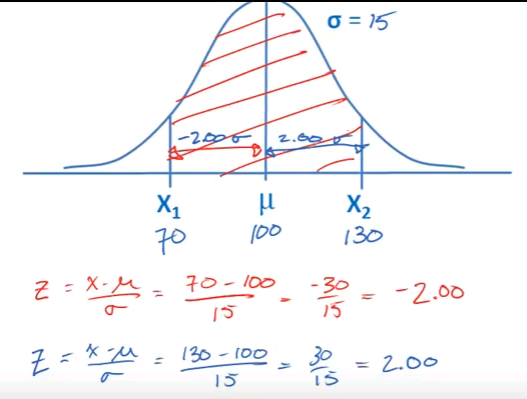
<https://en.wikipedia.org/wiki/Standard_normal_table>



68.26% of the population has a probability of having a IQ between 85 and 115

**Second example**





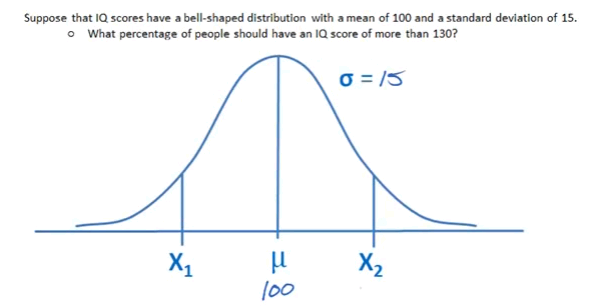
Referring to the z table

<https://en.wikipedia.org/wiki/Standard_normal_table>

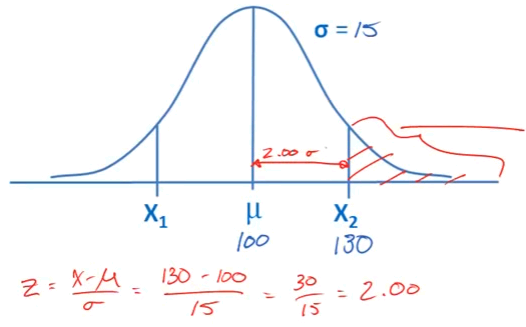


**95% of the population has a probability of having a IQ between 70 and 130**

**Third example**



**What percentage of the population would fall in the shaded area shown below**

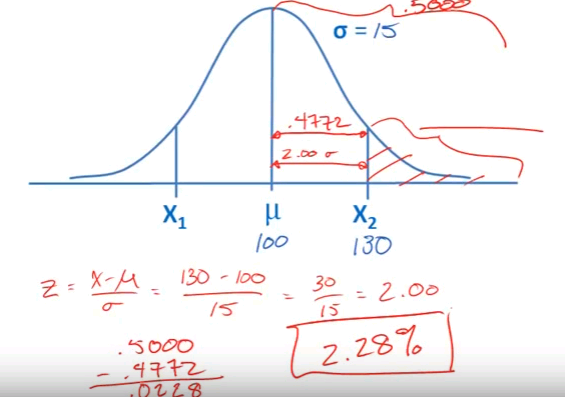


Referring to the z table, the distance between x value and mean is .4772

<https://en.wikipedia.org/wiki/Standard_normal_table>

**50% of the area is in the right of the mean**

**Take the 50% and substract the 0.4772. that is 0.5 – 0.4772 = 0.0228**



**This means 2.28% of the people have an IQ that is greater than 130.**

**P Value - Theory**

Taking the example of probablitily of a **coin toss**

Lets take the Hypothesis testing

So the **Null hypothesis** in this case would be H0

**H0** = It is a **Fair coin.** There is nothing wrong with the coin

And the Alternate Hypothesis would be

**HA =** There is something wrong with the coin. It is not a fair coin or the coin is a **loaded coin**

So to test if the coin follows the H0 or HA, we have to test the coin by making observations of the coin toss

|  |  |
| --- | --- |
| **Observation** | **P- Value** |
| 1 Tail – 1st Trial | 1st bservation is Tail, so it is 50% likely. **So let us set P value to** **0.5.**  We are not convinced if this is a trick coin with one flip. |
| 2 **Tail** | So still we are not convinced that it is a tricky coin. It is a high probability that the coin would flip on the same side twice.  But the likelihood of the Tail flip on 2nd toss is lesser than 1st, **so lets say**  **the P value to 0.25** |
| 3 Tail | 3rd time Tail is a bit unbelievable. So is this a tricky coin? There is a doubt.  But still it is possible in real life scenario for 3 continous Tail flips. So let us set **the P Value to 0.12** |
| 4 Tail | Very unlikely but still can happen.  **The P Value to 0.06** |
| 5 Tail | **P Value 0.03** |
| 6 Tail | **P value 0.01** |

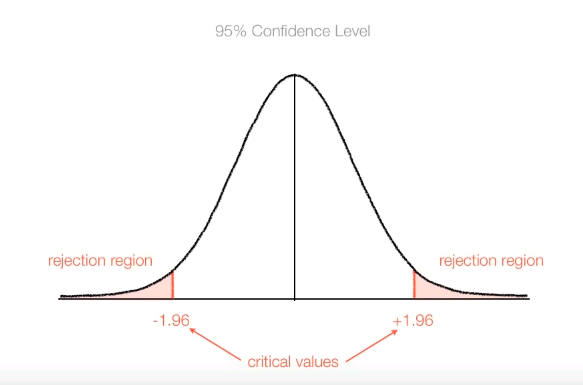
When P Value falls below 0.05, it is defeating the null hypothesis that the coin is fine and there is nothing wrong with the result.

So I think the coin is a Loaded coin – **A Trick coin – Most likely with Tails on both ends.**

**Data for Simple Linear regression**

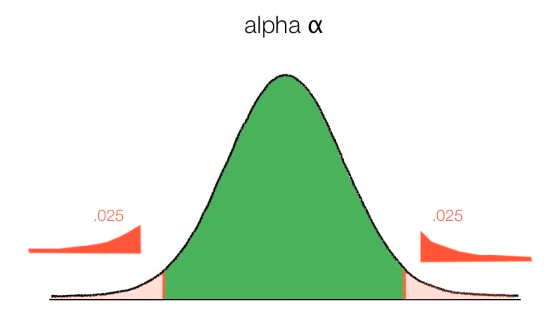
This startup company has been randomly defining salaries for employee. They might taking expereince and other market rates into consideration, but they are not following a model by which to define these salaries and increments

So they need our help to define a model using which they can define salaries for laterla joinees or use the model as a benchmark for increments and annual appraisals.



Here if we take some test of an experiment to calculate the z score and we get a Z score of 2.6.

**Then 2.6 is in the rejection region. That means we reject the NULL hypothesis.**

**As shown below if I get values in the rejection region in red, that means I reject the null hypothesis.-**

**Probability**

***Describes the possibility of an outcome of interest***

P = outcome of interest

all possible outcomes

***Fair coin flip***

P (***heads***) = 1/2 = 0.5

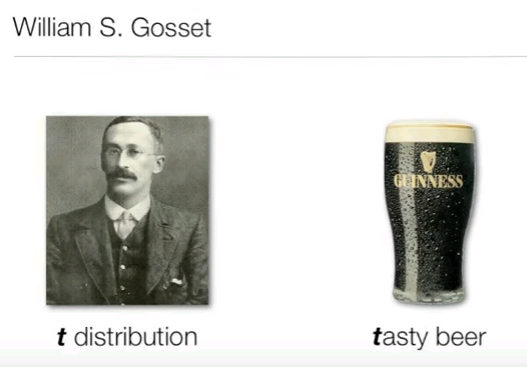
***Fair die flip***

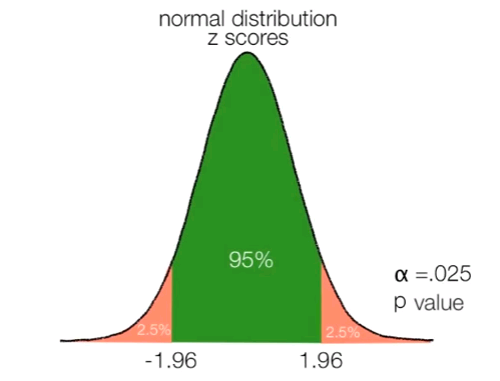
P (***1 or 2***) = 2/6 = 1/3 = 0.333

***Deck of playing cards***

P (***diamond card***) = 13/52 = 1/4 = 0.25

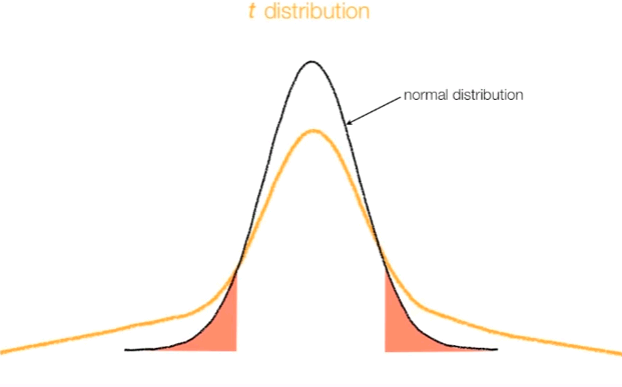
**T distribution**





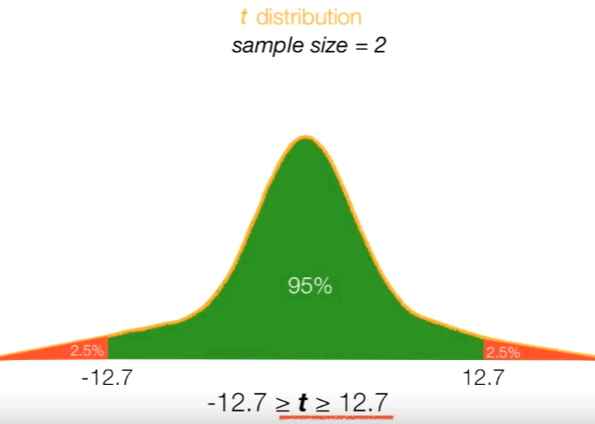
**1.96 are the z scores.**

**If -1.96 >= Z >= 1.96 (that is Z is greater than 1.96 and less than -1.96) then we reject the NULL hypothesis.**

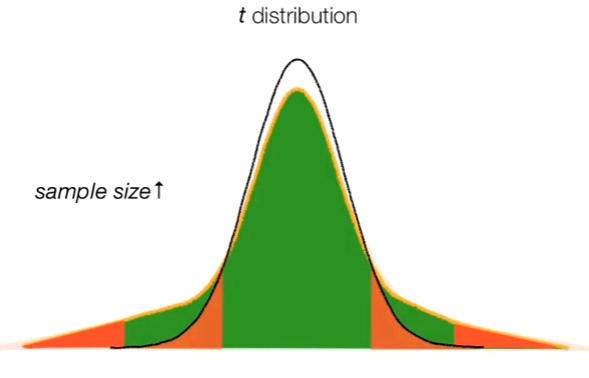


**T distrunition is used when sample size are very small as shown below.**

**And the null hypothesis is rejected if t is less than -12.7 and greater than 12.7**



**And as the sample size goes up, the t distribution tends to become a normal distribution.**



**Degrees of freedom = sample size – 1**

