Introduction to Predictive Analytics

Correlation and Linear Regression

How will you answer such questions?

- If I study 8 hours a day and my parent's highest education is bachelor's degree in physics; how much would I score in Maths test in 10th board exams?

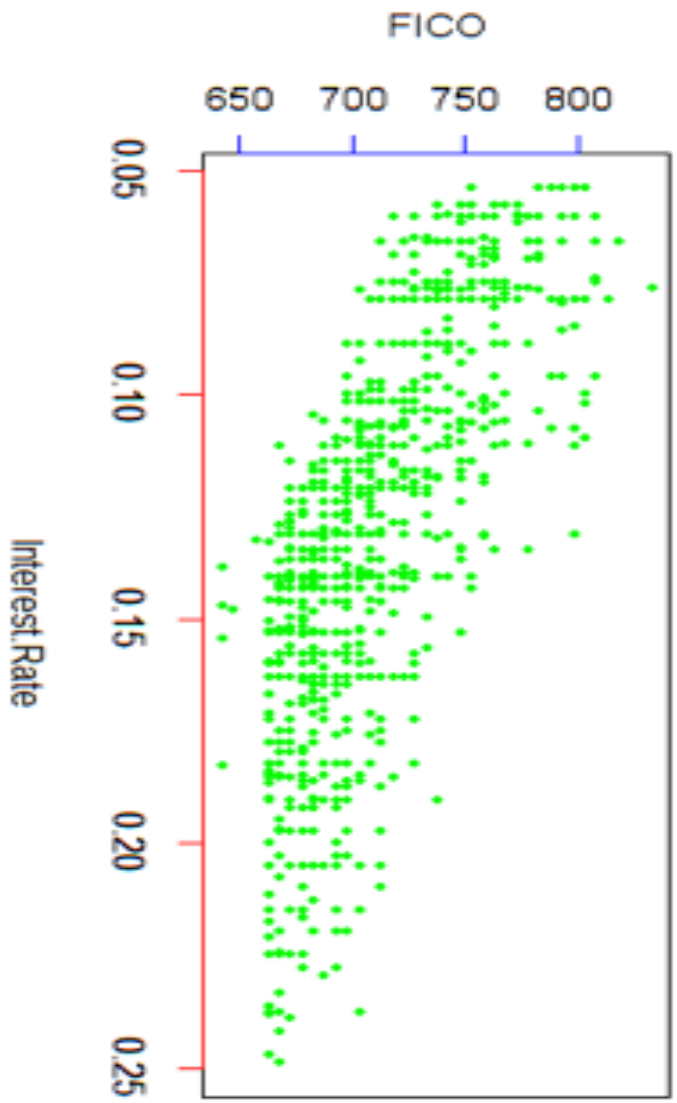
- If temperature today is 32^0C, humidity is 60%, there are 3 consecutive holidays coming up next week then which destinations would customers on my travelling advisory portal search for today?

Right, you will look at the correlations

- linear regression is about extracting a mathematical equation from the data, which tells us how our variable of interest is affected by other variables present in the data.

- Statistically speaking, how are variables in the data are correlated to each other. Lets learn!

Scatter Plot



A scatter plot (or scatter diagram) is a two dimensional graph obtained by plotting one variable against another variable

Useful for

- Depicting the relationship between two variables
- Identify outliers or unusual values
- Identify possible trends
- Identify a basic range of Y and X values

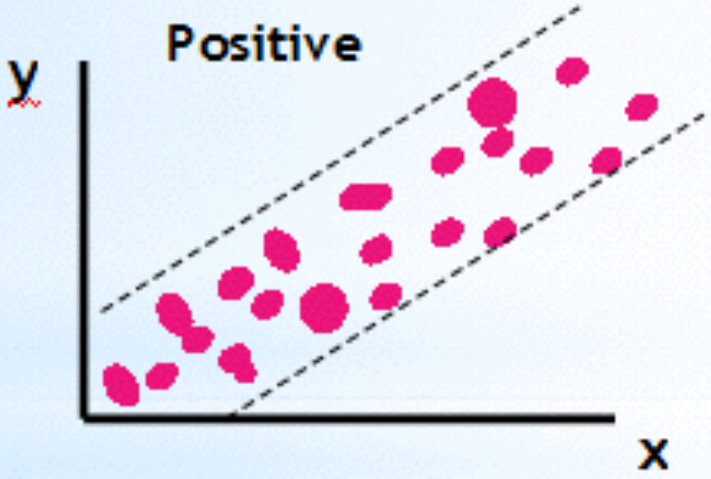
Scatter Plot Examples

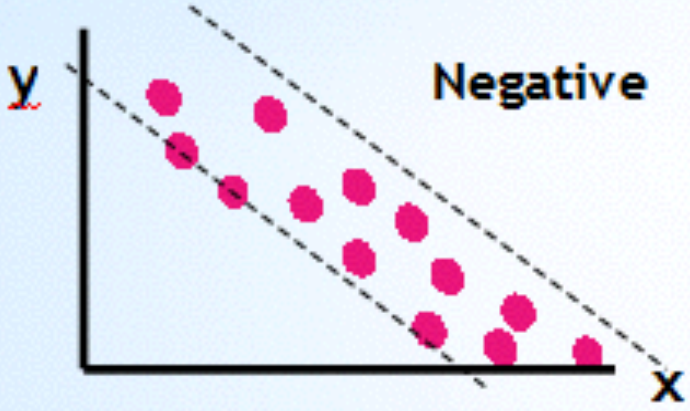
Linear - A straight line describes the relationship

Curvilinear - Curvature is present in the relationship

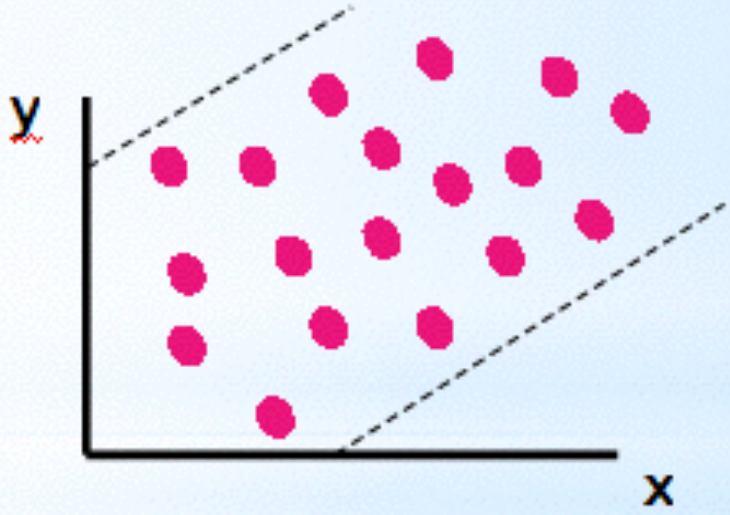
Scatter Plot Examples contd

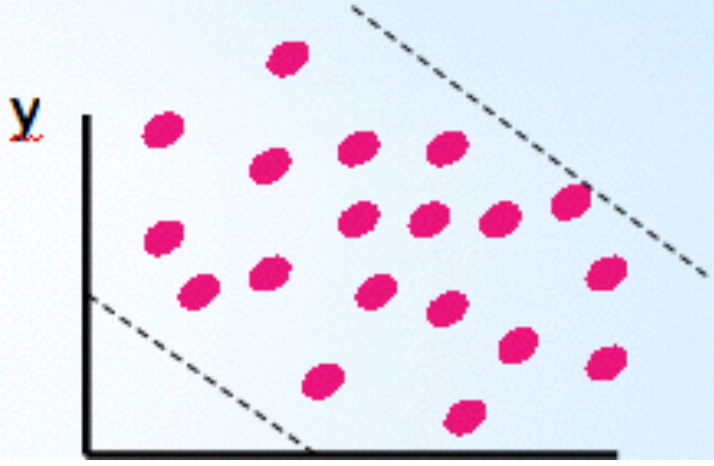
Strong linear relationships



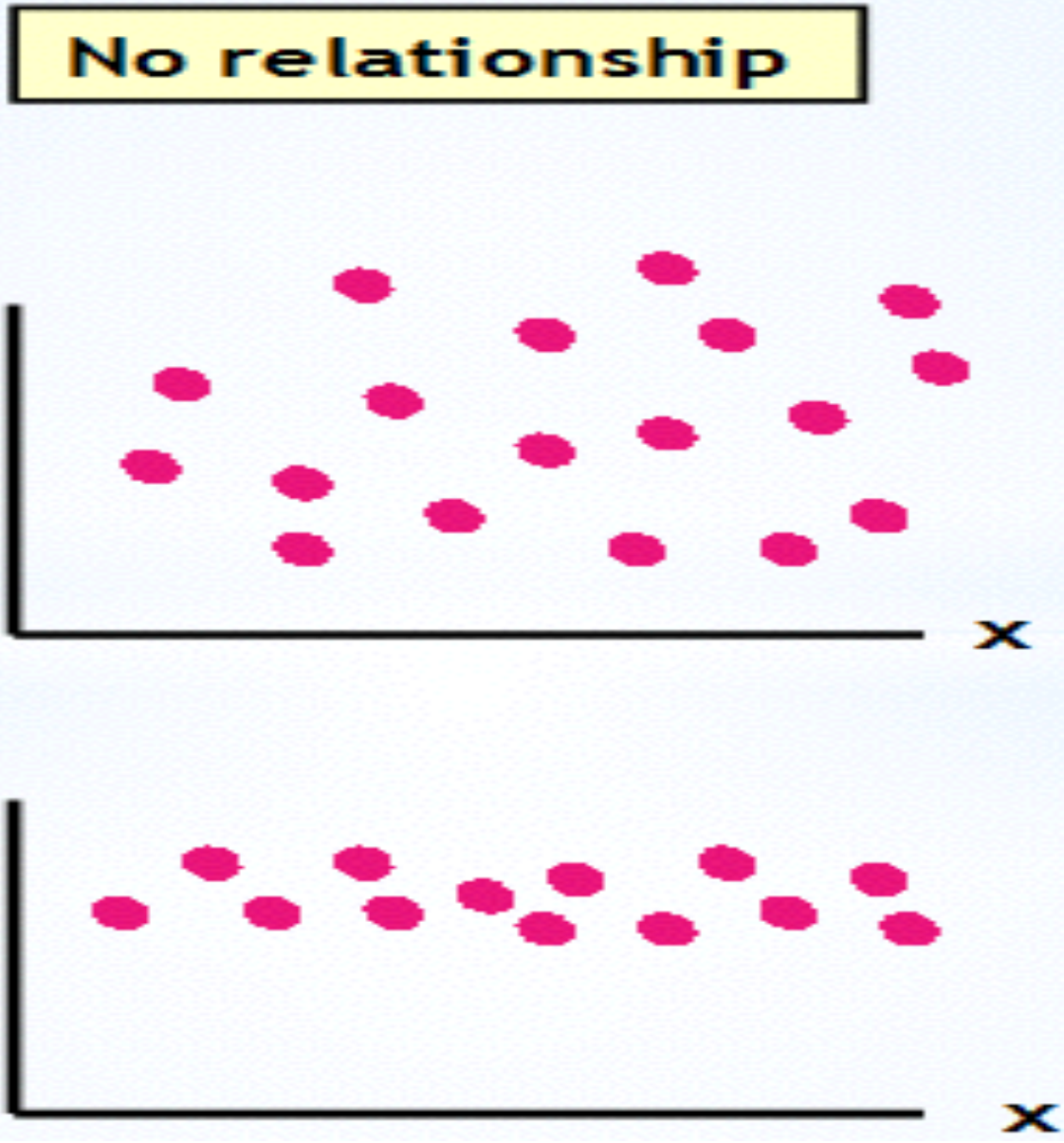


Weak relationships





And some more...



Quantifying Correlation

Linear Correlation Coefficient

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

- Measures linear correlations only
- Can we convert curvilinear relations to linear and use the same measure?
- Can we always do that?
- What can we use this correlation for?

Correlation and Causation

- During summers, in Mumbai, sales of sunglasses and ice creams increase significantly.

- Would you conclude that buyers of sunglasses also love ice-cream?

To find causation..

- Assigning causes comes from business process understanding.

- Data can help but it doesn't determine, you do.

Case Study : The Business Problem

- LoanSmart is a debt advisory firm which advises its clients on which banks to apply to for loans based on their need and personal details.

- LoanSmart goes through all the client characteristics like requested loan amount, duration, detailed credit history, personal information etc. Once done, it reaches a conclusion on which banks should its client apply to for loans.

Contd..

- However now it wants to add a value added service by helping its new clients understand how the banks stack up with respect to each other in terms of their interest rates for the specific client

- Precise criteria used by the banks to determine the rates remains unknown to LoanSmart

Contd..

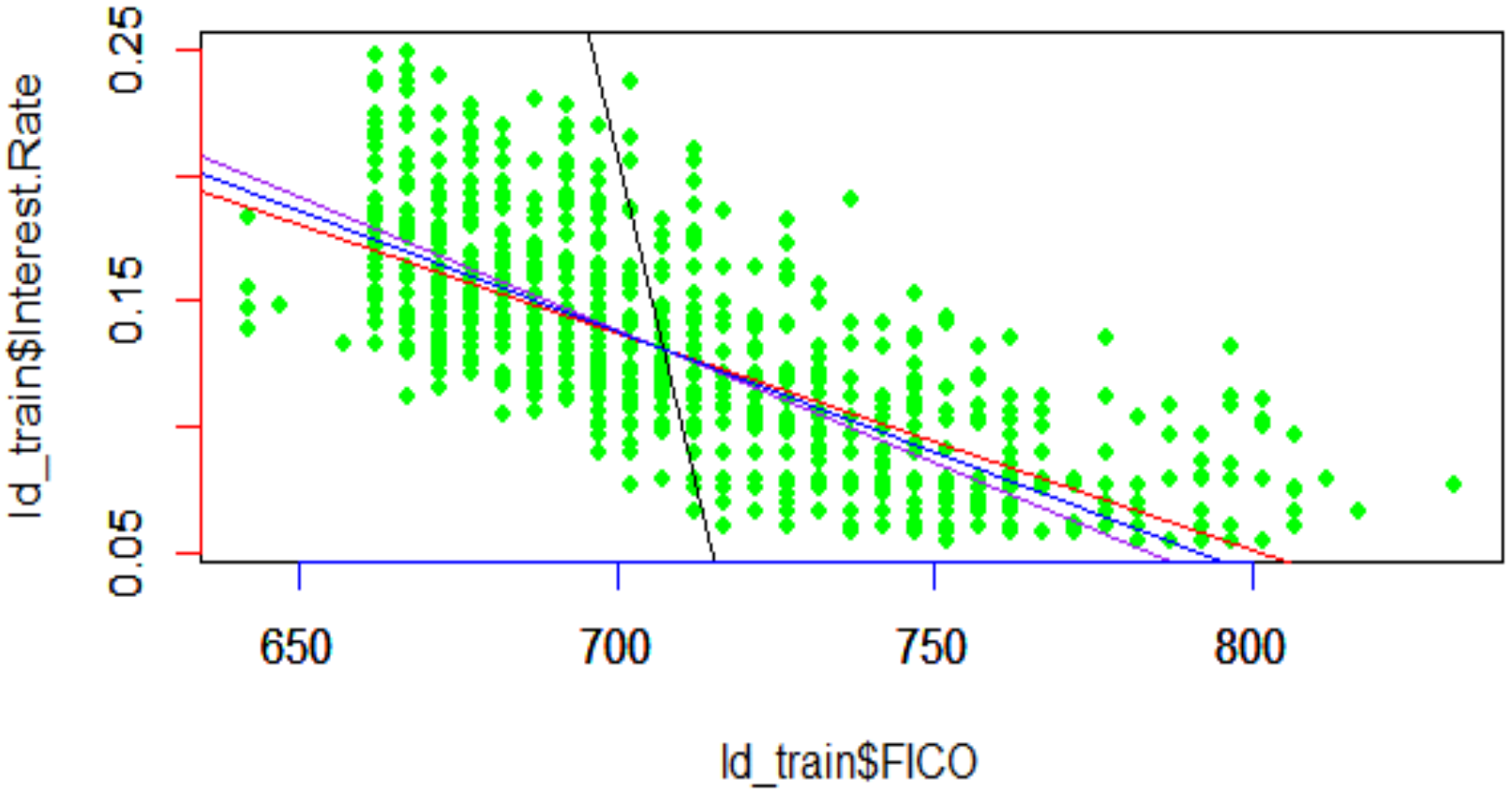
- However, LoanSmart has all the data of its previous clients and the banks they got the loan from and the interest rates offered.

- They want you to help them out on this problem by creating a proof of concept using the data of clients who took loans from one particular bank ABC Capital Ltd. before proceeding to work on data from other banks.

Case Study : the Analytics Problem

- "loandata.csv" contains data for 2,393 customers of a lending company ABC Capital Ltd.
- Take a look at the data and discuss:
  - What is the exact business problem to solve?
  - What is the approach we shall use to solve the problem?
- Derive the analytics solution and bring out the business solution

Data scatter Plot



Idea of a relationship to a mathematical equation!

Since we have assumed that there exist a linear relationship, we can write it this way

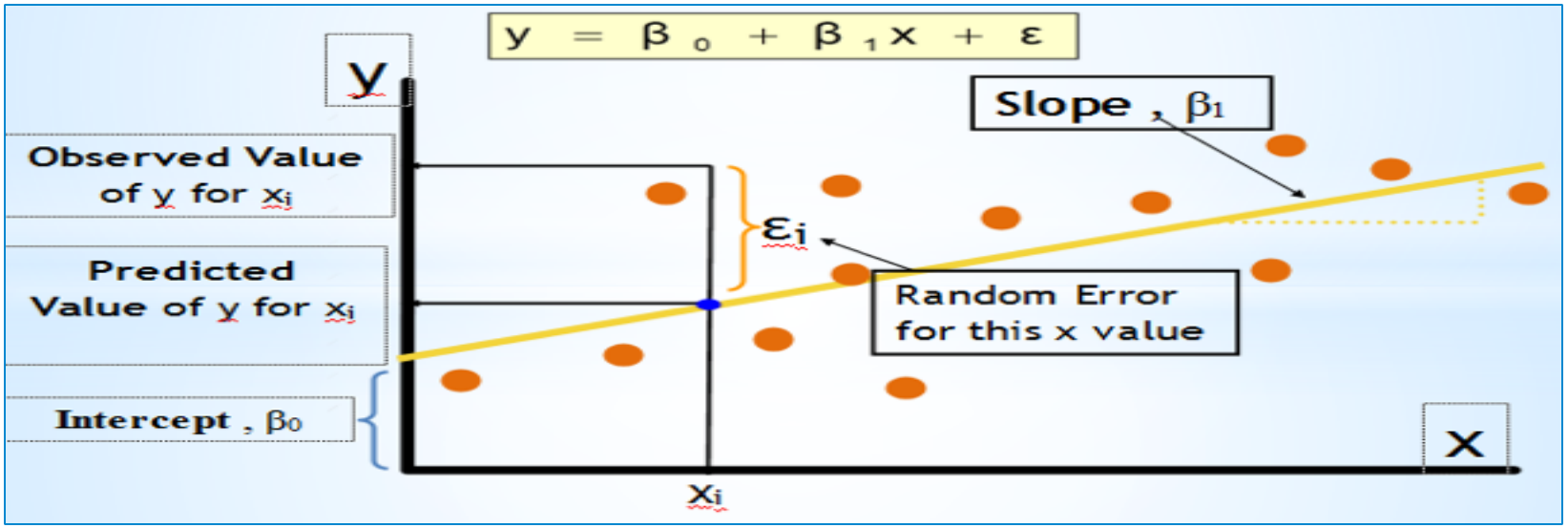
- $Y = a + bX$, but this is still incomplete. Why?

Adding Uncertainity

- We know that this is not a perfect relationship, each observation has some error associated with it.

- Y=a + bX + e , a and b are the parameters of this equation which we can estimate from the data that we have. How? Which of those lines is best?

Minimizing Sum of Square of errors



$$\Sigma e^2 = \Sigma(y-\beta_0-\beta_1 X)^2$$

We minimize this error sum w.r.t. To parameters β_0 and β_1.

Estimates of Parameters

The Least Squares Equation

- The formulas for b_1 and b_0 are:

$$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

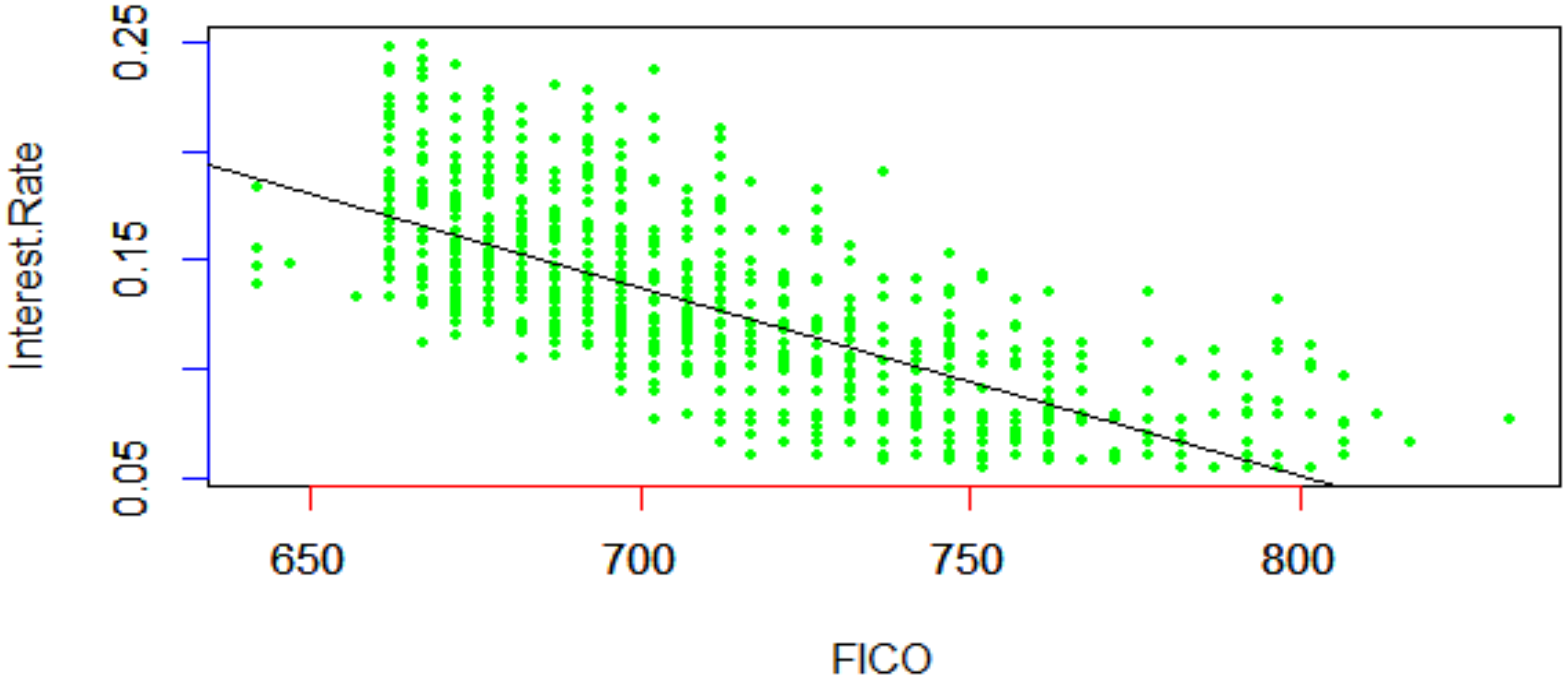
$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

No! you are not supposed to mug up this formulae. This is just to give you hint about what goes on at the back end inside the software that you are using.

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

Data Scatter plot with fitted Regression Line



Is This Enough?

- What did we just predict?
- What about errors associated with these prediction?
  - Is there a way to quantify them?

Linear regression Assumptions

- Linearity [More of a necessity]
- Normality of errors
- Homoscedasticity
- Independence of errors and predictor variables

Normality of Errors

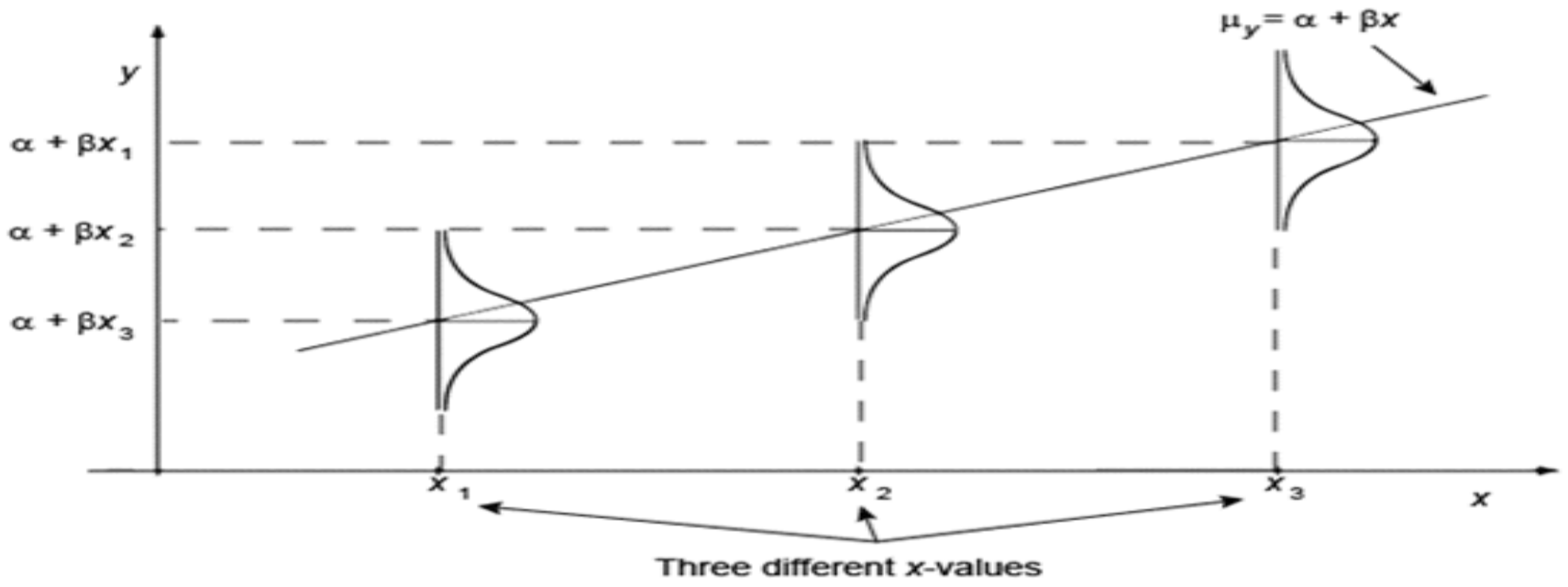
- Errors follow ~ $N(0,\sigma^2)$ for each given X/ predictor value.

- This allows us to build a confidence interval around our predicted value. [remember $Y=Y_{pred}$ +error]

Contd…

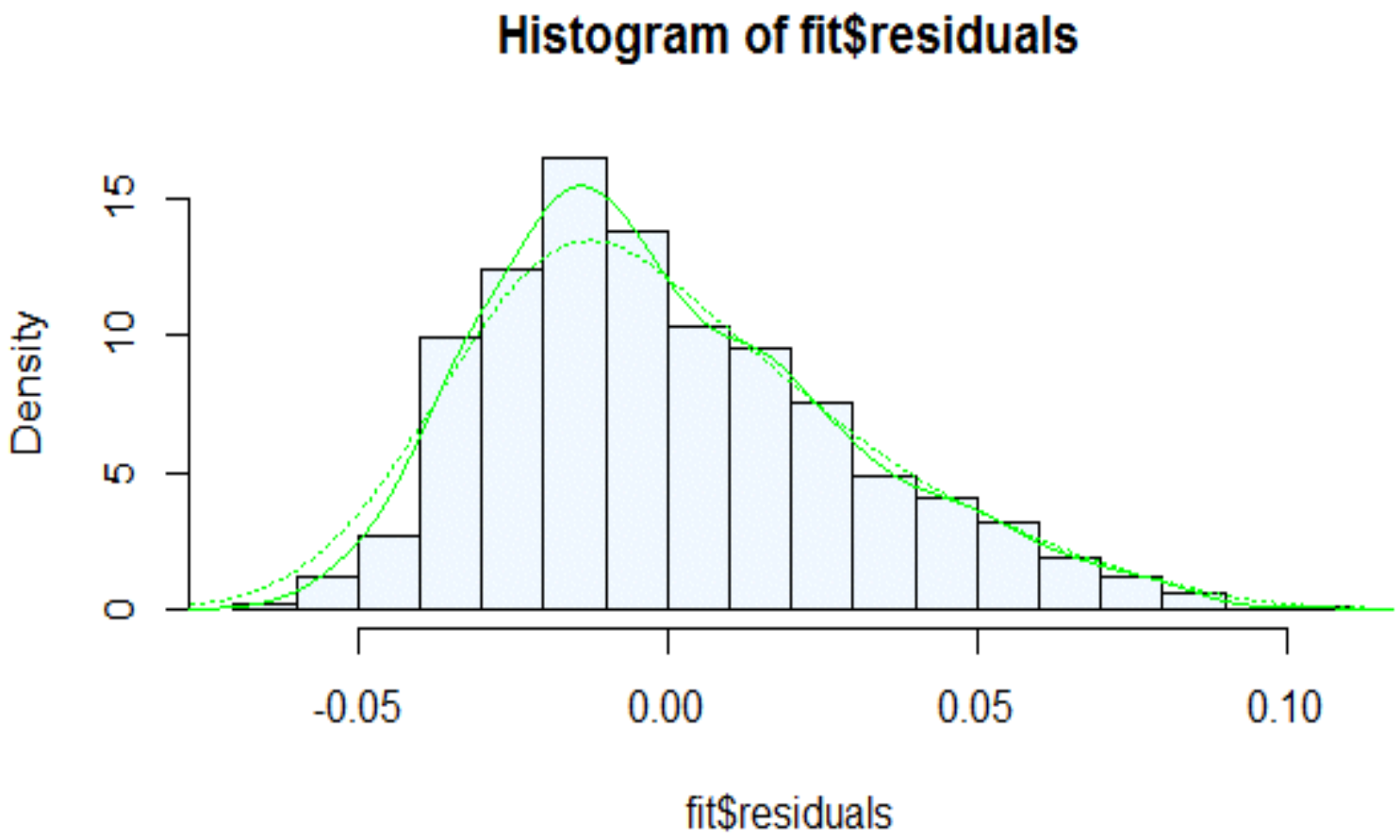
- Variance σ^2 associated with each predictor value can be estimated by looking at errors associated with that particular predictor value.
  - But in real life data, we rarely have multiple occurrences of any predictor value. Way out?

Visually understanding Normality



- Response need not follow normal distribution as a whole. Following from assumption of normality of errors, Response follows normal distribution for each predictor value.
- $Y_i \sim N(\mu_{yi}, \sigma^2)$

How do our Errors look?



Homoscedasticity

- Assumption is that variance of errors remains constant across predictor values.

- This enables us to estimate error variance, by pooling errors associated with all the observations.

Independence of Errors

- If you see a pattern between predictor variable and errors, what does this imply?

- There might exist a relationship between errors and predictor variable, lets say error=f(x) + e’

Contd..

- Putting this back to original equation which we assumed, $Y=\alpha + \beta X + f(X)+e'$

- This implies that relationship between response and predictor is “non-linear” as opposed to our assumption of linearity

Consequence and usage of normality assumption

- One direct usage is to get confidence intervals for prediction.

- Another consequence is that estimate ofβfollows normal distribution with variance s_βAnd mean β[the true value of parameter]

Contd..

- Null hypothesis for linear regression is that there is no relationship between predictor and response, or

- $\beta=0$

Contd..

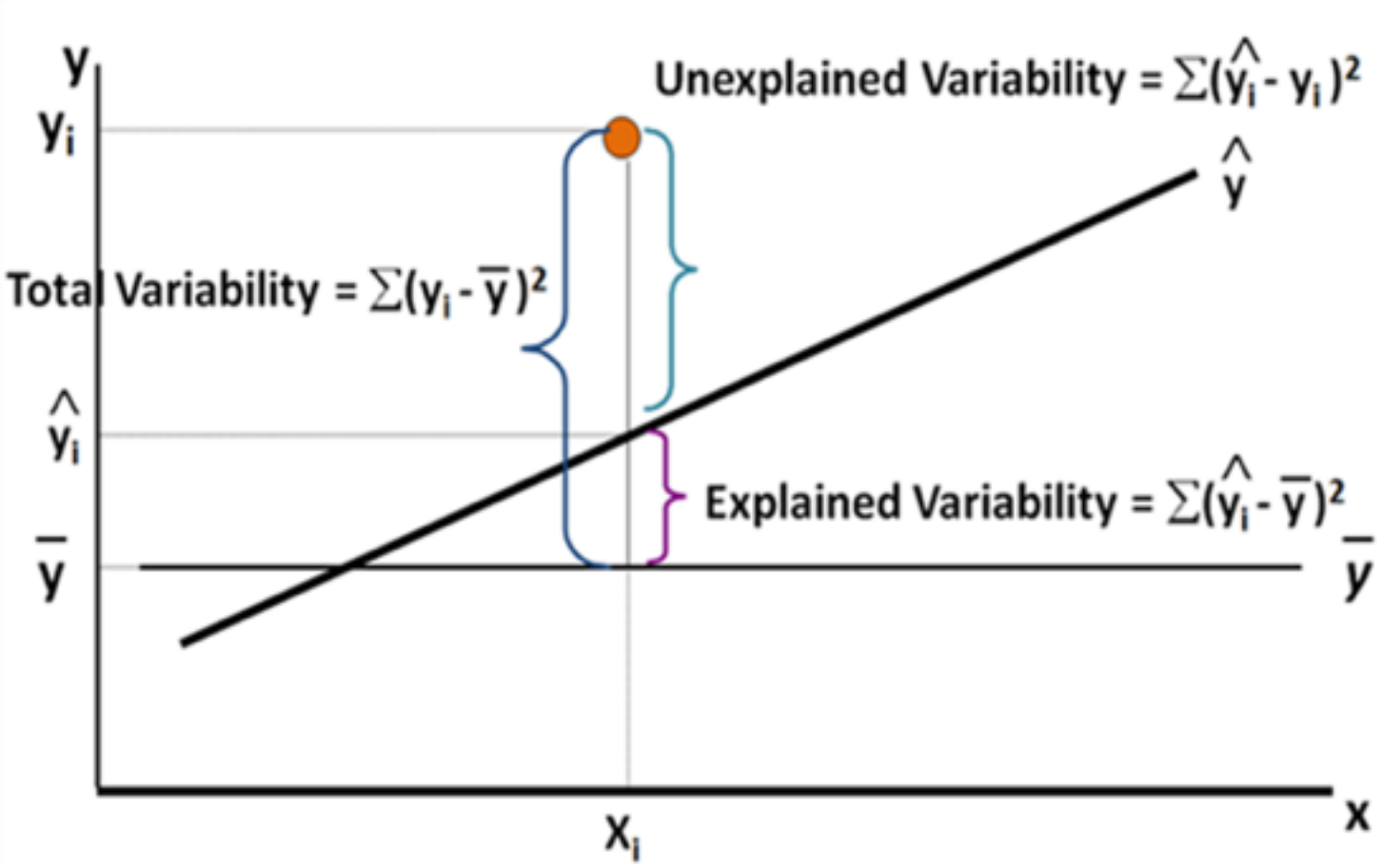
- From here we can construct a test statistic ($\beta_{Estimate}/s_{\beta}$)Which follows t-distribution with degrees of freedom n-2

- P-values for this hypothesis test tell us whether our parameter estimate is non-zero by chance or not.

How good is your model? Explained Variability

- Linear equation that we obtain doesn't ever perfectly fit the data at hand
- Framework of least square estimation focuses on minimizing the total error
- If average of target variable is A, all data points will have values around A with some variation. Let's say they take a value y. Variation would be = $|y-\mathcal{y}|$
- Let's say predicted value by the regression line is P. This will also have variation from the average = $|\mathcal{y} - \mathcal{y}|$
- This implies that although regression did account for some variation from the average for this particular value, there was some part of the variation which it could not account for which is = $|y- \mathcal{y}|$
- We shall see soon how these variations are written as sum of squares

Explained and Unexplained Variation, explained!



Coefficient of determination

- The portion of the total variation in the dependent variable that is explained by variation in the independent variable

- The coefficient of determination is also called R-squared and is denoted as R^2

- R^2=SSR/SST =Explained Variation/Total Variation

Adding more predictors : Multiple Linear Regression

- This is a simple extension of the simple linear regression .
- Equation becomes:
  - $Y=b_0+b_1X_1+b_2X_2.....+error$
- We estimate values of parameters by minimizing sum of square of errors [same as in simple linear regression]

Multi-co-linearity

- Remember those linear equations which we solved to get values for parameter estimate?
  - They don't have a unique solution if any of the predictor variables are perfectly correlated

- If there exist a partial correlation, the above problem doesn't exist, but it has some other serious repercussions.

Contd..

$S_\beta = S_\beta' * (1/(1-R_j^2))$ for MLR, where R_j^2 is R^2 for the linear model when j^{th} predictor is taken as a response and rest of the predictors as predictors. Here S_β' is the variance if it was SLR.

Effects of multi-co-linearity

- Because of variance inflation [measured by VIF], t-values are decreased [not because of lowered significance but because of multi-co-linearity!!] , which results in artificially inflated p-values. P-values are not reliable anymore!

- Interpreting the model becomes difficult. How?

Remedies!

- Get rid of the correlated predictors
  - One by one. [why not all of them at once?]

- Instead of those variables, use a combination of the variables. [Principle component analysis, Variable Clusters]. We'll not go into details of PCA or Varclus here.

Train and Test

- Why did we build this model?
  - To use it for prediction of response for unknown/ unseen/future scenarios .
- How do we simulate that with the data at our hand?
  - We split the data into two parts, call them; train and test [model and validation etc]
- Procedure to use: Proc Surveyselect

Validation

- The validation or test data set should be similar to the train dataset. How to verify?
  - Check univariate statistics in both the dataset
  - Do ttest, or chisq tests to see if variable behaviour remains same across train and test
  - Build model and compare coefficients

Validation : Consistency

- Performance of the model built on train should be consistent on test dataset.

- Not only under-performing [on test data] is bad, over-performing is bad as well! ***Consistency*** is the key.

Lets solve the case study