Statistical Techniques for Data Science

Regression Analysis

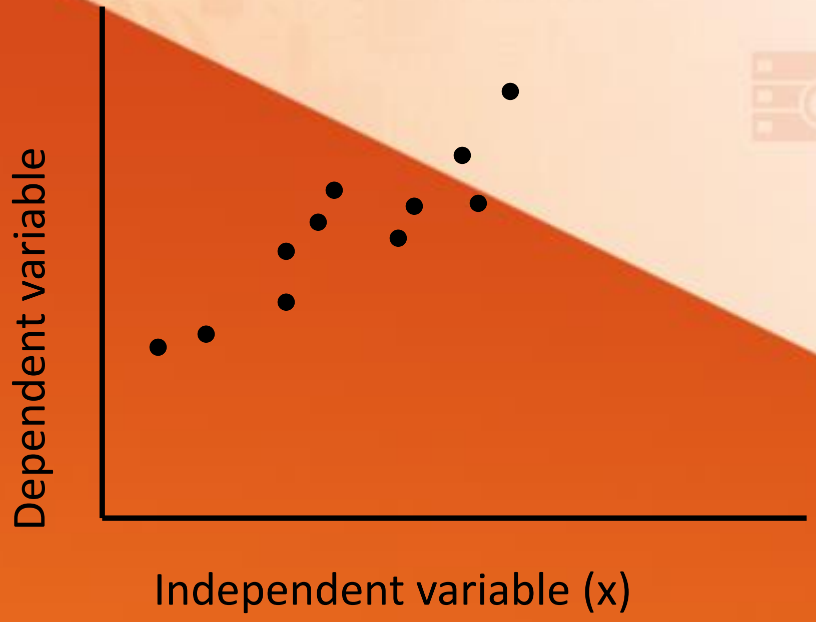
Dr. Subhababa Pal
Manipal Global Academy of Data Science

Objective

manipalglobal
Academy of Data Science

After attending this session, you will be able to –

- **Explain what is Regression**
- **Explain what is Simple Linear Regression**
- **Explain what is Multiple Linear Regression**
- **Describe Multi-collinearity**
- **Explain Logistic Regression**

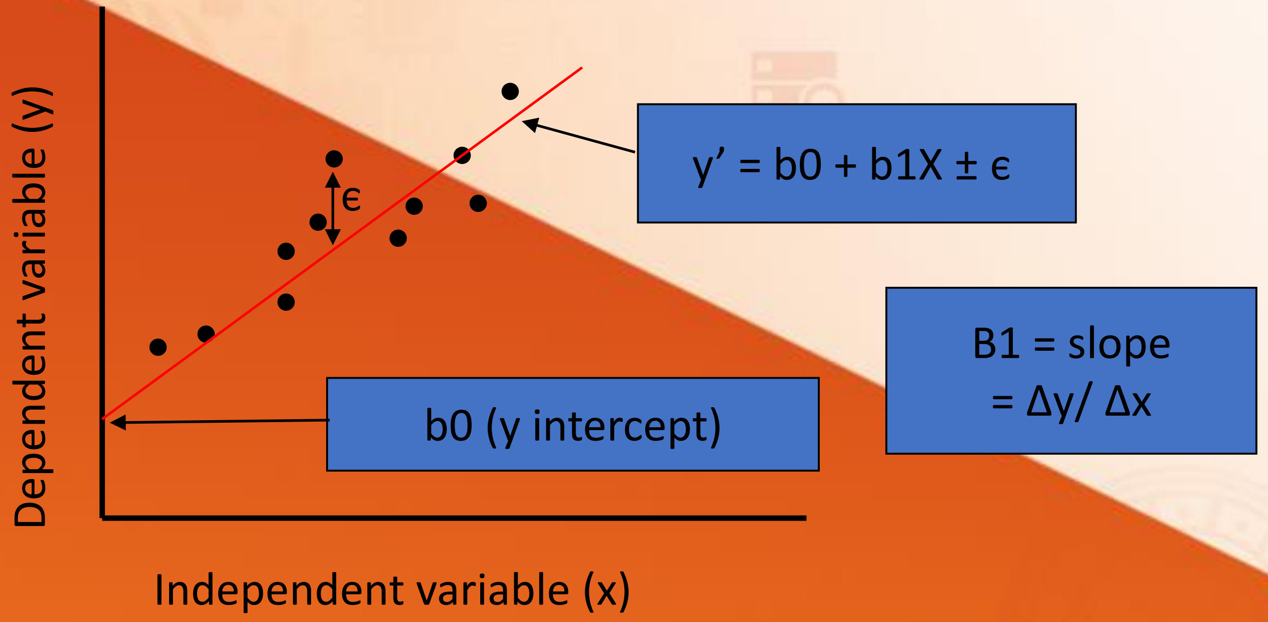


Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables.

Regression is thus an explanation of causation.

If the independent variable(s) sufficiently explain the variation in the dependent variable, the model can be used for prediction.

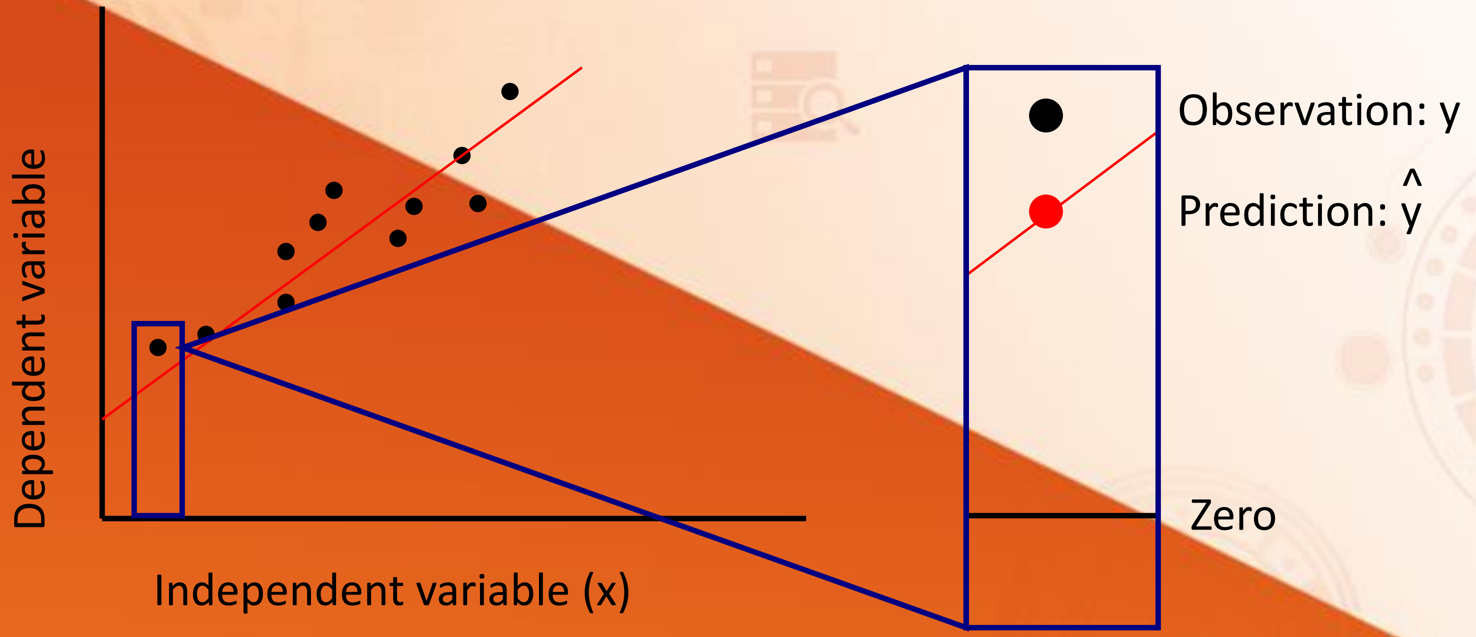
Simple Linear Regression



The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.

Simple regression fits a straight line to the data.

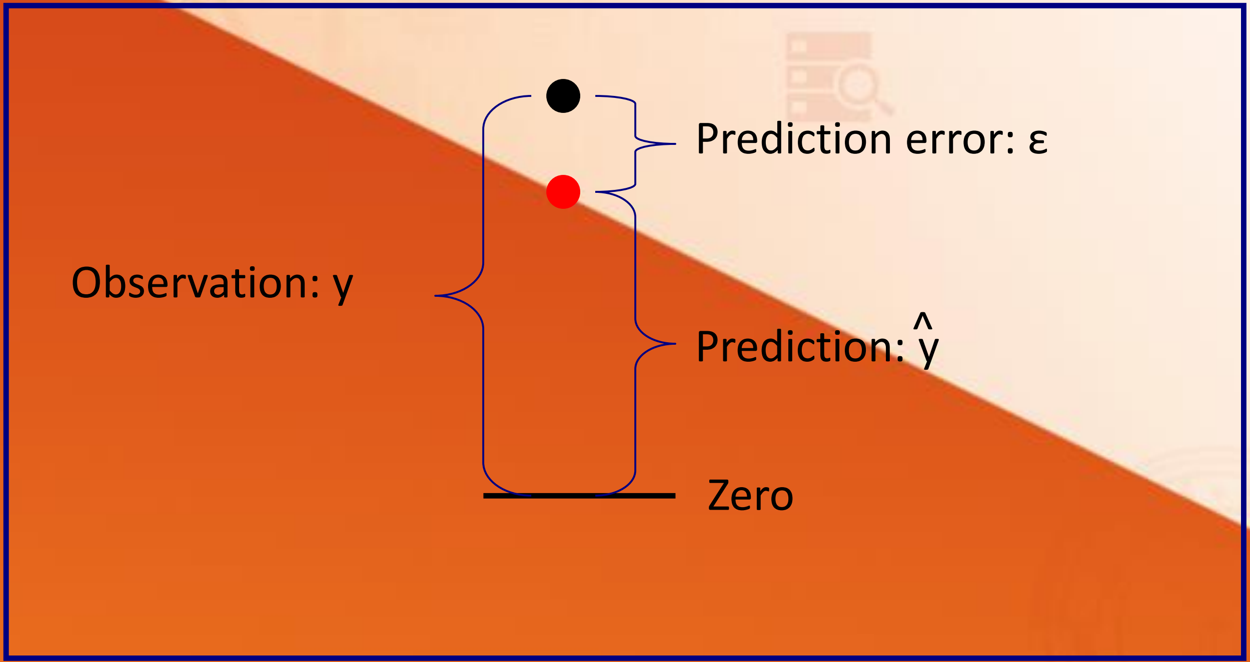
Simple Linear Regression



The function will make a prediction for each observed data point.

The observation is denoted by y and the prediction is denoted by $\hat{y}$.

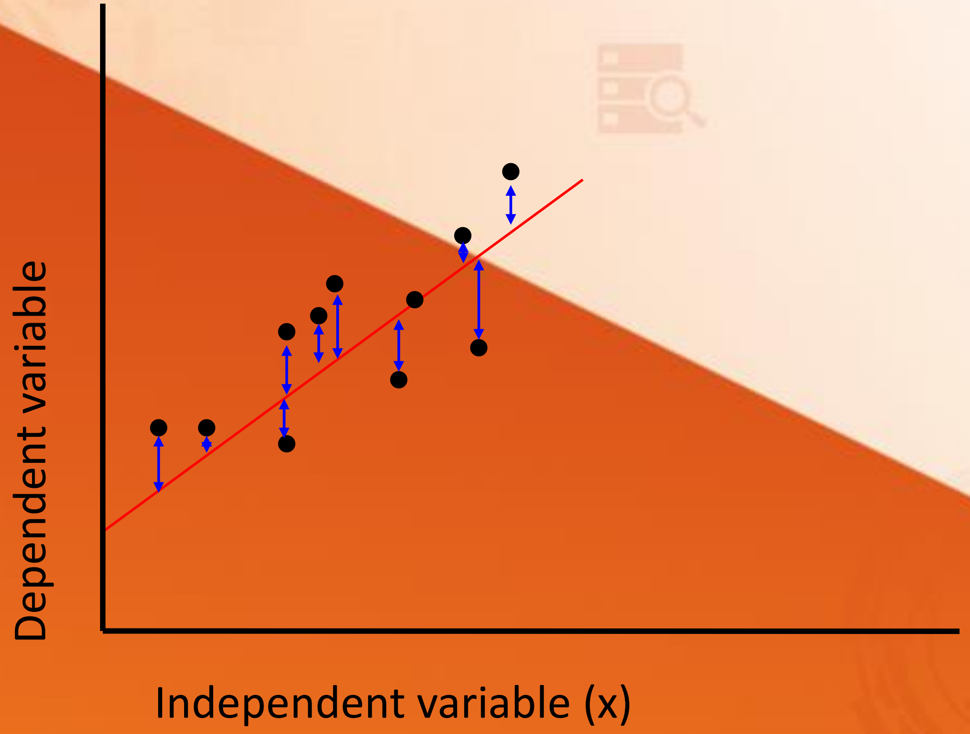
Simple Linear Regression



For each observation, the variation can be described as:

$$y = \hat{y} + \varepsilon$$

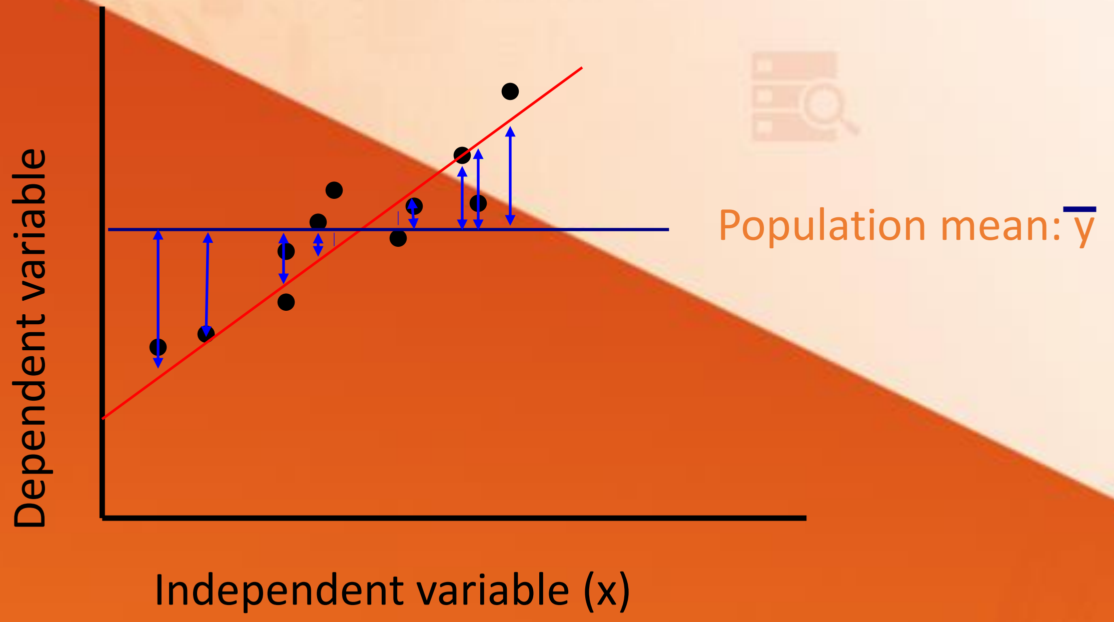
Actual = Explained + Error



A least squares regression selects the line with the lowest total sum of squared prediction errors.

This value is called the Sum of Squares of Error, or SSE.

Calculating SSR



The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

The Total Sum of Squares (SST) is equal to SSR + SSE.

Mathematically,

SSR = $\sum(\hat{y}-\bar{y})^2$ **(measure of explained variation)**

SSE = $\sum(y-\hat{y})^2$ **(measure of unexplained variation)**

SST = $\sum(y-\bar{y})^2$ **(measure of total variation in y)**

The Coefficient of Determination

The proportion of total variation (SST) that is explained by the regression (SSR) is known as the Coefficient of Determination, and is often referred to as R .

$$R^2 = \frac{SSR}{SST}$$

The value of R can range between 0 and 1, and the higher its value the more accurate the regression model is. It is often referred to as a percentage.

Standard Error of Regression

The Standard Error of a regression is a measure of its variability. It can be used in a similar manner to standard deviation, allowing for prediction intervals.

y ± 2 standard errors will provide approximately 95% accuracy, and 3 standard errors will provide a 99% confidence interval.

Standard Error is calculated by taking the square root of the average prediction error.

$$\text{Standard Error} = \sqrt{\frac{SSE}{n-k}}$$

Where n is the number of observations in the sample and k is the total number of variables in the model

The output of a simple regression is the coefficient β and the constant A. The equation is then:

$$y = A + \beta * x + \varepsilon$$

where ε is the residual error.

β is the per unit change in the dependent variable for each unit change in the independent variable. Mathematically:

$$\beta = \frac{\Delta y}{\Delta x}$$

More than one independent variable can be used to explain variance in the dependent variable, as long as they are not linearly related.

A multiple regression takes the form:

$$y = A + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

where k is the number of variables, or parameters.

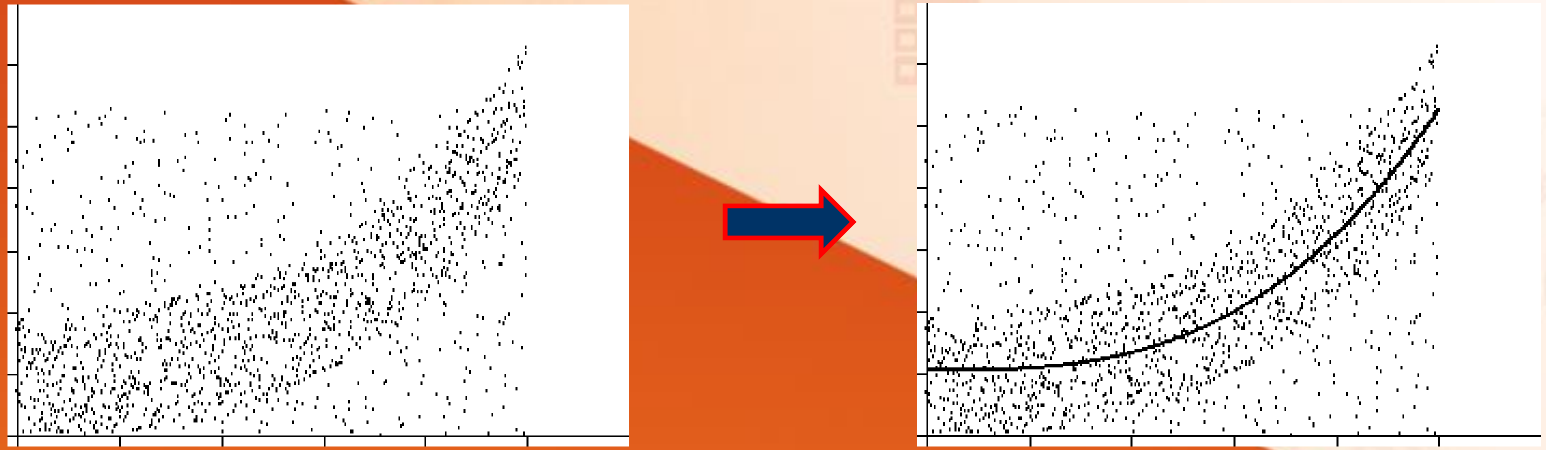
Multicollinearity is a condition in which at least 2 independent variables are highly linearly correlated. It will often crash computers.

Example table of Correlations

	Y	X1	X2
Y	1.000		
X1	0.802	1.000	
X2	0.848	0.578	1.000

A correlations table can suggest which independent variables may be significant. Generally, an ind. variable that has more than a .3 correlation with the dependent variable and less than .7 with any other ind. variable can be included as a possible predictor.

Nonlinear Regression



Nonlinear functions can also be fit as regressions. Common choices include Power, Logarithmic, Exponential, and Logistic, but any continuous function can be used.

Linear Regression

- Simple Linear Regression → The case where only one explanatory Variable is present
- Multiple Linear Regression → The case where multiple explanatory Variables are present.
- Data is modeled using linear predictor functions
- Unknown model parameters are estimated from the data.
- First Type of Regression model studied rigorously
- Goal of any regression model is to fit a predictive model to an observed data.

Different Steps in Regression

manipalglobal
Academy of Data Science

- Step 1: Create the training (development) and test (validation) data samples from original data.
- Step 2: Develop the model on the training data and use it to predict the distance on test data
- Step 3: Review diagnostic measures.
- Step 4: Calculate prediction accuracy and error rates

Linear Regression – R Code

```
# Create Training and Test data -
  set.seed(100)  # setting seed to reproduce results of random
  sampling
  trainingRowIndex <-sample(1:nrow(cars), 0.8*nrow(cars))  # row
  indices for training data
  trainingData <-cars[trainingRowIndex, ]  # model training data
  testData  <-cars[-trainingRowIndex, ]   # test data
```

Linear Regression – R Code

```
# Build the model on training data -
  lmMod <-lm(dist ~speed, data=trainingData)  #
  build the model
  distPred <-predict(lmMod, testData)  # predict
  distance


summary (lmMod)  # model summary

#>
  #> Call:
  #> lm(formula = dist ~ speed, data =
  trainingData)
```

Linear Regression – R Code

```
#> Residuals:
  #>     Min      1Q  Median      3Q     Max
  #> -23.350 -10.771  -2.137   9.255  42.231
  #>
  #> Coefficients:
  #>             Estimate Std. Error t value Pr(>|t|)
  #> (Intercept)  -22.657      7.999  -2.833  0.00735 **
  #> speed          4.316      0.487   8.863 8.73e-11 ***
```

Linear Regression – R Code

- #> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  #> Residual standard error: 15.84 on 38 degrees of freedom
  #> Multiple R-squared: 0.674, Adjusted R-squared: 0.6654
  #> F-statistic: 78.56 on 1 and 38 DF, p-value: 8.734e-11
  AIC (lmMod) # Calculate akaike information criterion
  #> [1] 338.4489

Linear Regression – R Code

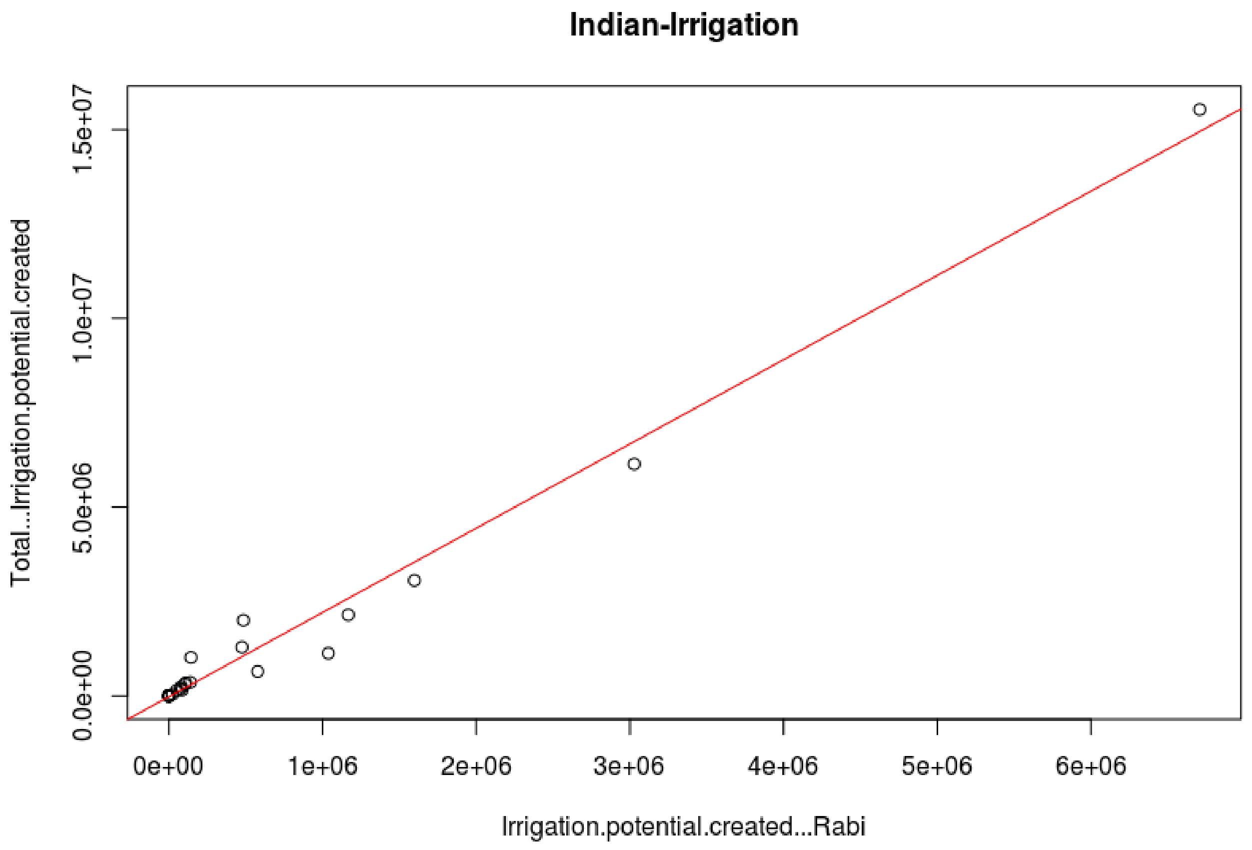
From the model summary, the model p value and predictor's p value are less than the significance level, so we know we have a statistically significant model. Also, the R-Sq and Adj R-Sq are comparative to the original model built on full data.

A simple correlation between the actuals and predicted values can be used as a form of accuracy measure. A higher correlation accuracy implies that the actuals and predicted values have similar directional movement,

Simple Linear Regression R code

```
x<-read.csv("datafile.csv", header = TRUE, sep = ",")
head(x)
attach(x)
fit <-lm(Total...Irrigation.potential.created~Irrigation.potential.created...Rabi, data=x)
summary(fit)
fitted(fit)
plot(Total...Irrigation.potential.created~Irrigation.potential.created...Rabi, data=x, main="Indian-Irrigation")
abline(fit, col="red")

# diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit)
detach(x)
```

Simple Linear Regression R Plot



http://tarah.ai

Multiple Linear Regression R code

```
x<-read.csv("datafile.csv", header = TRUE, sep = ",")
head(x)
attach(x)

#two predictor model
two_pred_mod <- lm(Total...Irrigation.potential.created~Irrigation.potential.created...Rabi+Irrigation.potential.created...Perennial, data=x)
two_pred_mod

#Three predictor model
three_pred_mod <- lm(Total...Irrigation.potential.created~Irrigation.potential.created...Rabi+Irrigation.potential.created...Perennial+Irrigation.potential.created...Kharif, data=x)
three_pred_mod
```

Logistic Regression

- A regression model where the dependent variable is categorical.
- Here let us consider a dependent variable which is binary
- The binary logistic model is used to predict a binary response based on one or more predictor (or independent) variables (features), making it a probabilistic classification model in the parlance of machine learning, or a qualitative response/discrete choice model in the terminology of economics.
- Logistic Regression is fairly mathematical, and we will focus on the basic principle that underline Logistic Regression

Logistic Regression

manipalglobal
Academy of Data Science

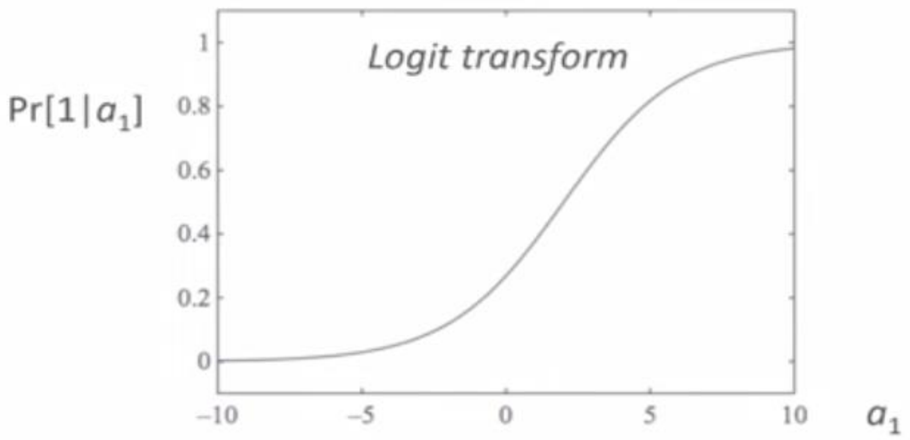
- Uses probabilities rather than actual values
- Instead of predicting whether its going to be a “yes” or a “no”, it is better to predict the probability with which you think it’s going to be ‘yes’ or a ‘no’.
- Ex: If an student is 95% likely to pass rather than he/she is definitely going to pass the exam.
- There are some other algorithms like NaiveBayes, J48 which uses probabilities

manipalglobal Academy of Data Science

Logistic Regression

- In Linear Regression, we calculate a linear function and then a threshold
- In a Logistic Regression, we estimate the probabilities of the dependent variable directly.

$$\Pr[1 \mid a_1, a_2, \ldots, a_k] = 1/(1 + \exp(-w_0 - w_1 a_1 - \ldots - w_k a_k))$$



Logistic Regression

- In Logistic Regression we have to choose the weights to maximize the log-likelyhood
- Sometimes the numbers that come on the regression lines are negative, it is helpful to use logistic regression in such scenarios.
- In Linear Regression we have a linear sum, but in a logistic regression, we embed the sum in formula as given below

$$\Pr[1 \mid a_1, a_2, \ldots, a_k] = 1/(1 + \exp(-w_0 - w_1 a_1 - \ldots - w_k a_k))$$

- The output lo

Logistic Regression

Confusion Matrix

A **confusion matrix**, also known as a contingency table or an error **matrix** , is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching **matrix**).

```
=== Confusion Matrix ===

   a   b   <-- classified as
 180  22 |   a = tested_negative
  46  59 |   b = tested_positive
```

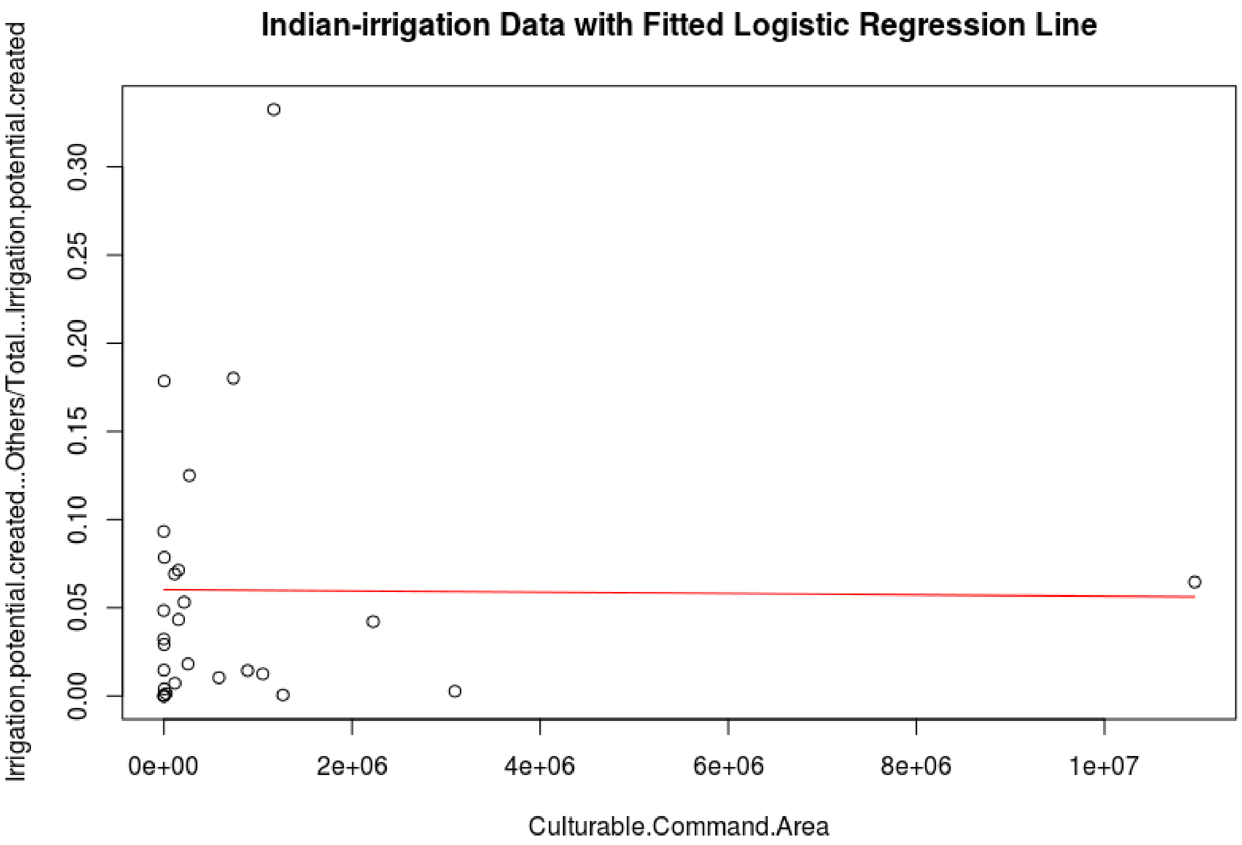
Logistic Regression R code

```
mydata <- read.csv("binary.csv")
## view the first few rows of the data
head(mydata)

#To get the basic derivatives of data
summary(mydata)

#to get the standard deviations of data
sapply(mydata, sd)

## two-way contingency table of categorical outcome and predictors
## we want to make sure there are not 0 cells
xtabs(~ admit + rank, data = mydata)
```

Logistic Regression – R Code

```
#Logistic regression needs a categorical output variable
mydata$rank <- factor(mydata$rank)
mylogit <- glm(admit ~ gre + gpa + rank, data = mydata, family =
"binomial")

#Checking the model created
summary(mylogit)
```

Logistic Regression R Plot



http://tarah.ai

manipalglobal Academy of Data Science

Ordinary Least Regression

- In statistics, ordinary least squares (OLS) or linear least squares is a strategy for assessing the obscure parameters in a linear regression model
- The objective of minimizing the entirety of the squares of the contrasts between the watched reactions in the given dataset
- Those anticipated by a straight capacity of an arrangement of illustrative factors

Ordinary Least Regression R code

```
# The X matrix for this problem:
X.matrix <- cbind(rep(1,length=length(Culturable.Command.Area )),Irrigation.potential.created...Kharif , Irrigation.potential.created...Rabi , Irrigation.potential.created...Perennial)

# Getting the fitted values for the ridge-regression fit:
fitted.vals <- X.matrix %*% c(43.840113, 2.117493, -0.959731, -1.018061)

# Getting the SSE for the ridge-regression fit:
sse.ridge <- sum((Culturable.Command.Area-fitted.vals )^2); sse.ridge

# The original least-squares fit:
bodyfat.reg <- lm(Culturable.Command.Area ~ Irrigation.potential.created...Kharif + Irrigation.potential.created...Rabi + Irrigation.potential.created...Perennial)
```

Model Validation

- Process of deciding if the results obtained from quantifying hypothesized relationships between variables are Acceptable or not
- Multiple Ways of Model Validation
  - Using R^2
  - Analysis of Residuals
  - Graphical Analysis of Residuals
  - Quantitative Analysis of Residuals
  - Data Splitting and Testing
  - Out of Sample Evaluation a.k.a Cross-Validation

Model Validation

- R^2 → Also known as Coefficient of Determination is a number that indicates the proportion of variance in the dependent variable which is predictable from the dependent variable
- Analysis of Residuals → Residuals are the differences between the predicted values and the actual.
- Graphical analysis of residuals: A scatter plot between the actuals and predicted values is a solid example

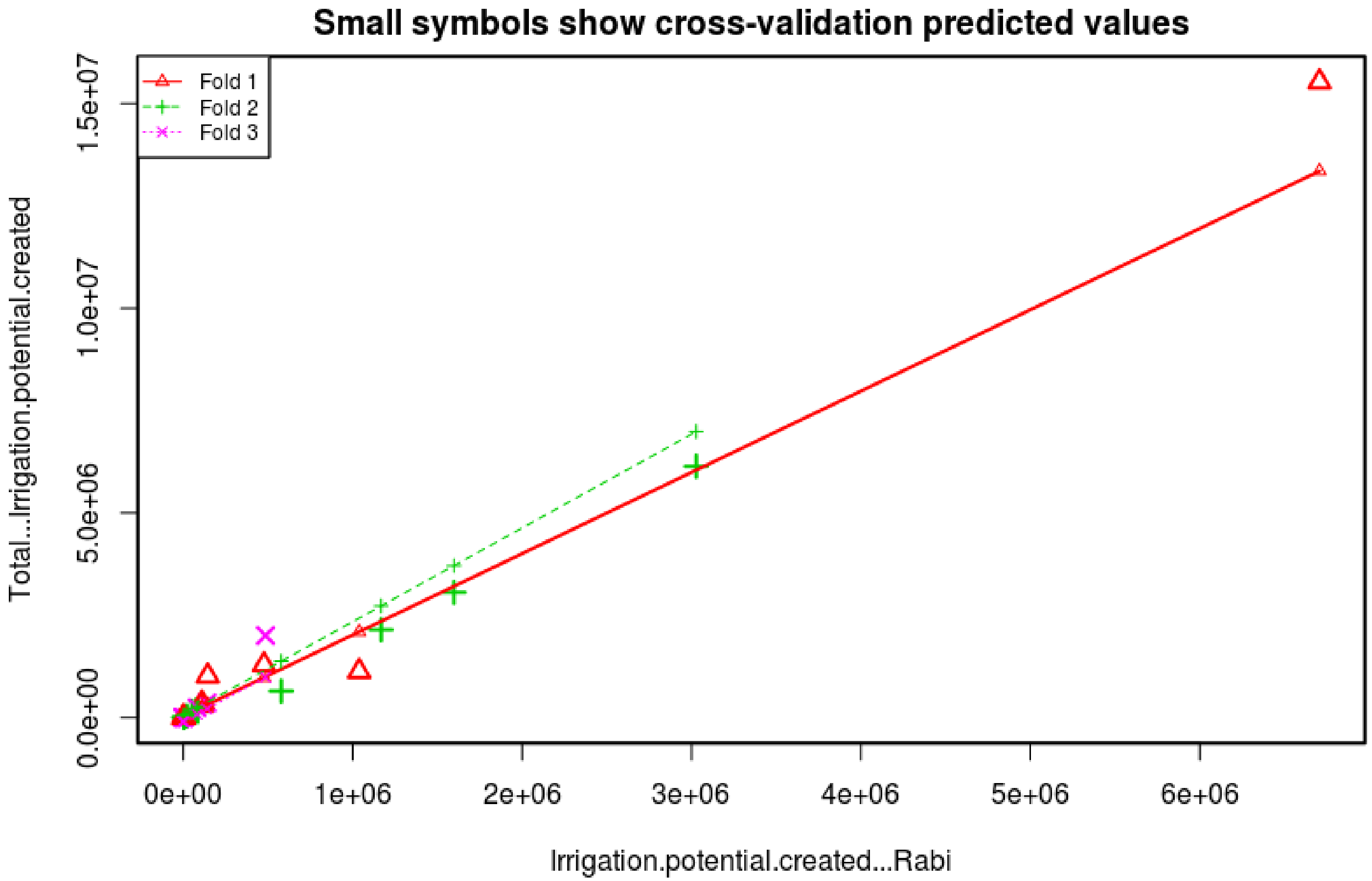
Model Validation

- Quantitative Validation
- Out-of-sample Evaluation
- Data Splitting

Model validation

```
x<-read.csv("datafile.csv", header = TRUE, sep = ",")
head(x)
attach(x)
fit <-lm(Total...Irrigation.potential.created~Irrigation.potential.created...Rabi, data=x)
predict(fit, x[1, ])

library(DAAG)
cv.lm(x, form.lm = formula(Total...Irrigation.potential.created~Irrigation.potential.created...Rabi))
```

Model validation



http://tarah.ai

Applications of Regression

- Application of Regression Analysis in Business
- A pharmaceutical organization utilized regression to evaluate the solidness of the dynamic fixing in a medication to anticipate its time frame of realistic usability
- A charge card organization connected relapse examination to anticipate month to month blessing card deals and enhance yearly income projections.
- A lodging establishment utilized relapse to recognize a profile for and anticipate potential customers who may default on a timeshare credit

Summary

manipalglobal
Academy of Data Science

- Regression is an approach for demonstrating the relationship between and one or more variables
- Function lm() is used in R to develop and model linear regression models in R
- Logistic regression is the suitable regression analysis to direct when the reliant variable is binary
- Ridge regression resembles least squares yet recoils the evaluated coefficients towards zero
- Ordinary Least Squares (OLS) is a strategy for assessing the obscure parameters in a linear regression model
- Applications based on Regression is sorted out

manipalglobal
Academy of Data Science

THANK YOU!