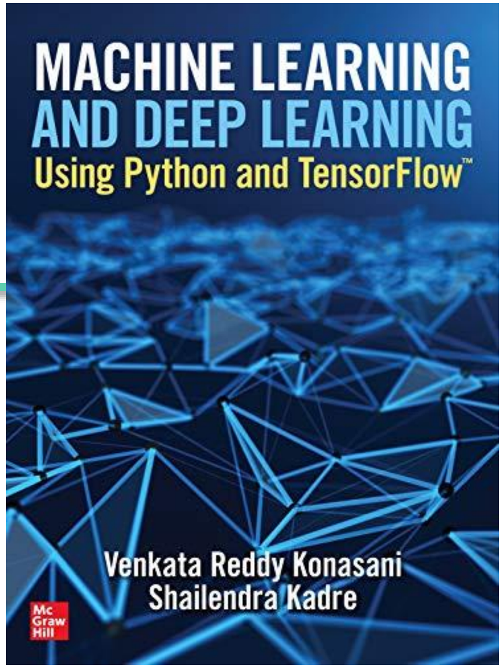
statinfer

Regression Analysis

Venkat Reddy

Chapter 3 in the book



Introduction

Contents

- Correlation
- Simple Regression
- R-Squared
- Multiple Regression
- Adj R-Squared
- P-value
- Multicollinearity
- Interaction terms

Correlation

Quantify Association

- Is there any association between hours of study and grades?
- What happens to sweater sales with increase in temperature? What is the strength of association between them?
- What happens to ice-cream sales v.s temperature? What is the strength of association between them?
- How to quantify the association?
- Which of the above examples has very strong association?

Correlation coefficient

- It is a measure of linear association
- r is the ratio of variance together vs product of individual variances.

$$\text{Correlation coefficient } r = \frac{\text{Covariance of XY}}{\text{sqrt(VarianceX * VarianceY)}} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n}}{\text{sqrt}\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \text{ X } \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}\right)}$$

- Correlation 0 No linear association
- Correlation 0 to 0.25 Negligible positive association
- Correlation 0.25-0.5 Weak positive association
- Correlation 0.5-0.75 Moderate positive association
- Correlation >0.75 Very Strong positive association

LAB –Correlation Calculation

- Dataset: AirPassengers\\AirPassengers.csv
- Find the correlation between number of passengers and promotional budget.

AirPassengers Data

```
#   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   Week_num                        80 non-null     int64
 1   Passengers                      80 non-null     int64
 2   Promotion_Budget                80 non-null     int64
 3   Service_Quality_Score           80 non-null     float64
 4   Holiday_week                    80 non-null     object
 5   Delayed_Cancelled_flight_ind    80 non-null     object
 6   Inter_metro_flight_ratio        80 non-null     float64
 7   Bad_Weather_Ind                 80 non-null     object
 8   Technical_issues_ind            80 non-null     object
```

Code –Correlation Calculation

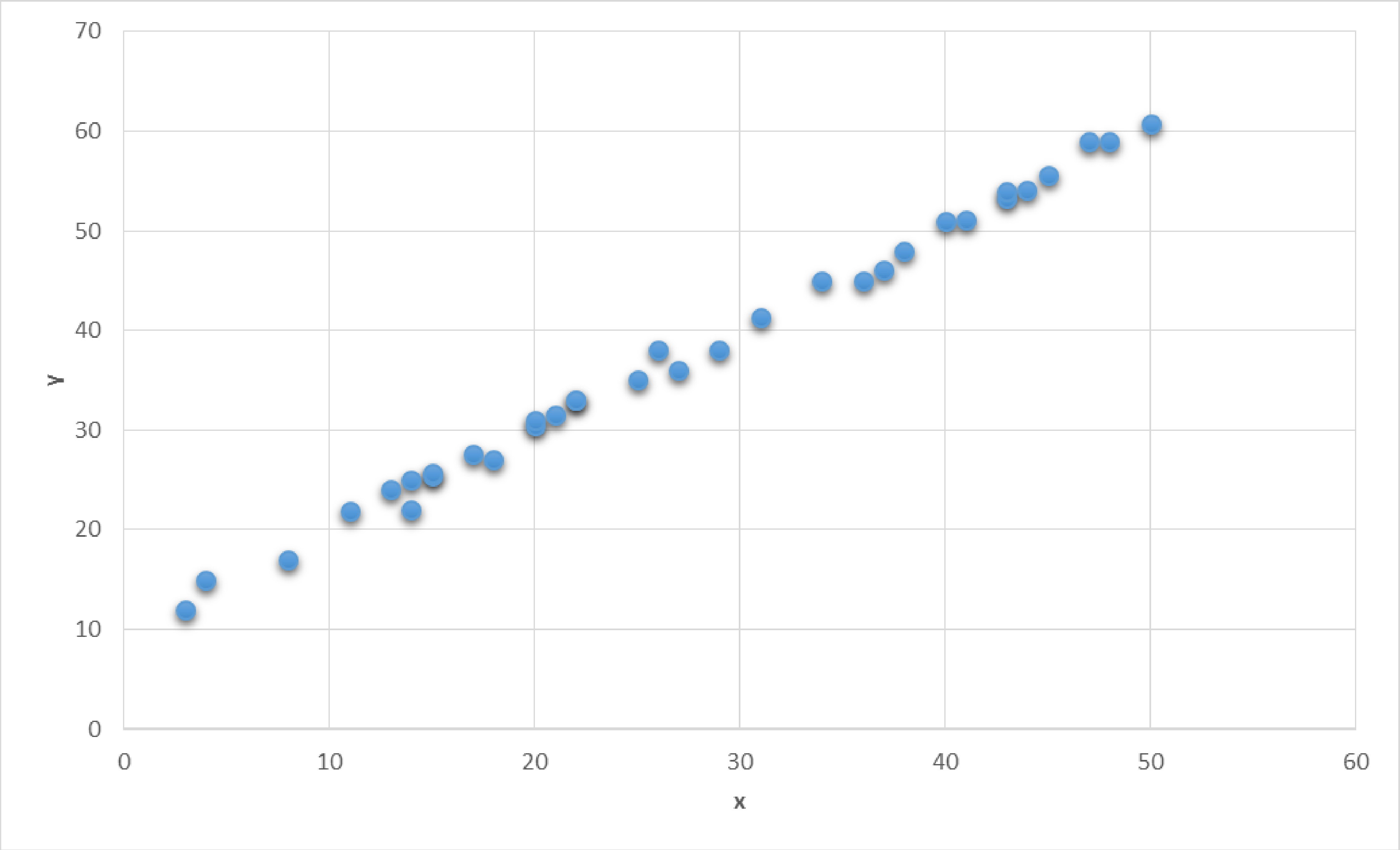
```
#Importing Air passengers data
air = pd.read_csv("D:\\Google Drive\\Training\\Datasets\\AirPassengers\\AirPassengers.csv")
air.shape
air.columns.values
air.head(10)
air.describe()

#Find the correlation between number of passengers and promotional budget.
np.corrcoef(air.Passengers,air.Promotion_Budget)
```

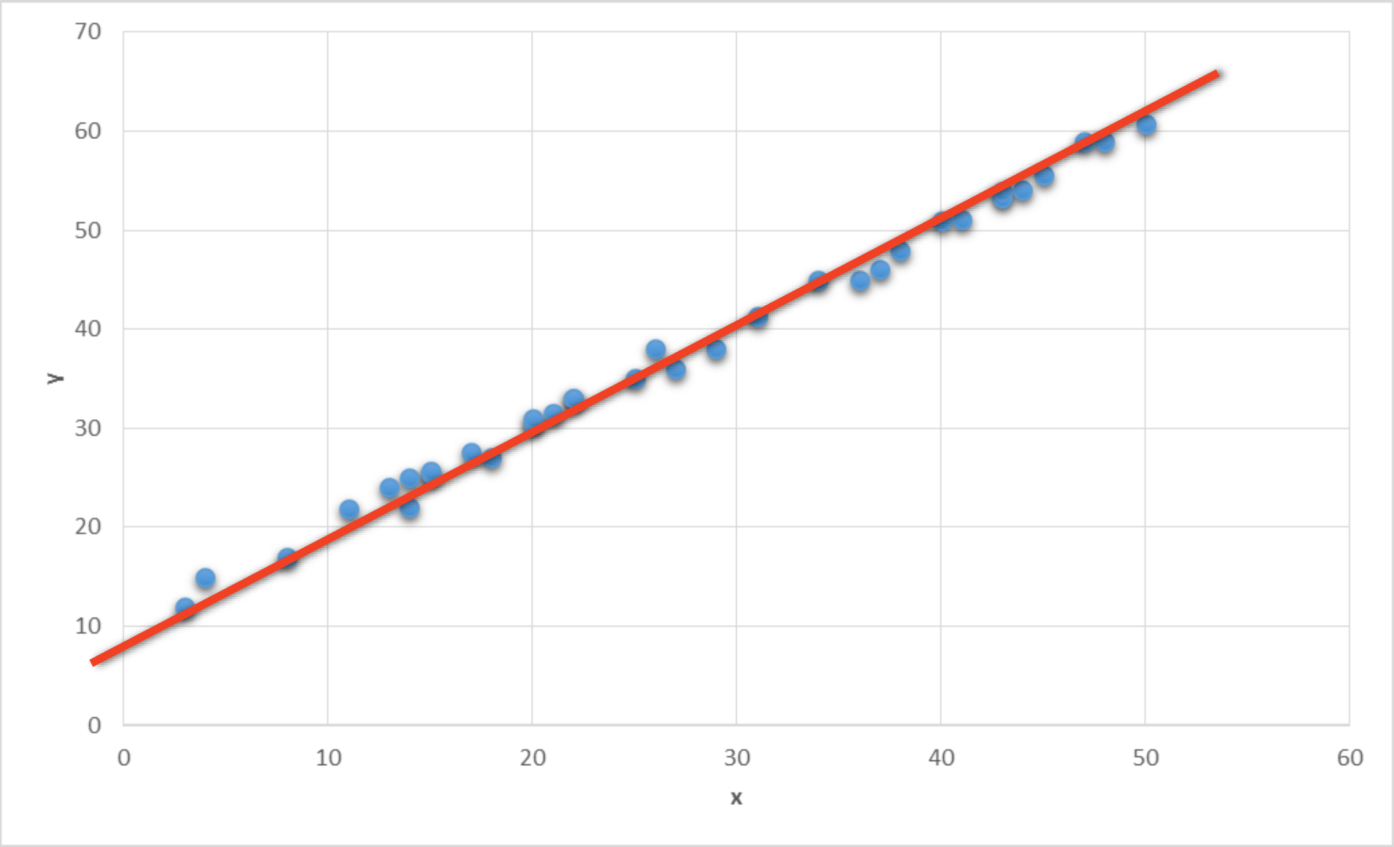
Correlation for Prediction

- Correlation is just a measure of association
- It can’t be used for prediction.
- Given the predictor variable, we can’t estimate the dependent variable.
- In the air passengers example, given the promotion budget, we can’t get an estimated value of passengers
- We need a model, an equation, a fit for the data.

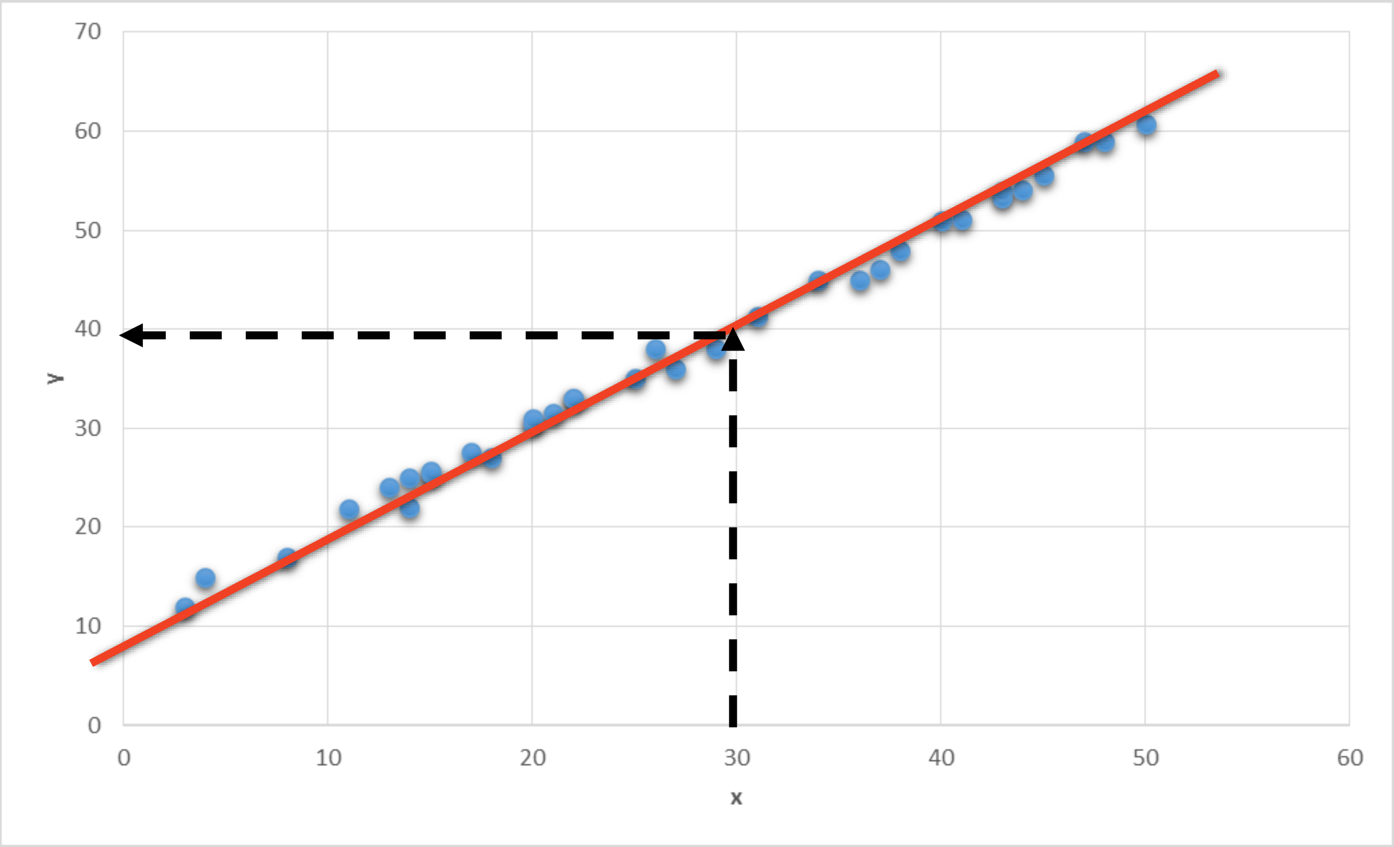
X Vs Y



Prediction



Prediction



Line Equation

Straight Line equation

$$\boxed{y = mx + c}$$

Regression terminology

$$\boxed{y = \beta_0 + \beta_1 x}$$

What is Regression

- A regression line is a mathematical formula that quantifies the general relation between a predictor/independent (or known variable x) and the target/dependent (or the unknown variable y)
- Below is the regression line. If we have the data of x and y then we can build a model to generalize their relation

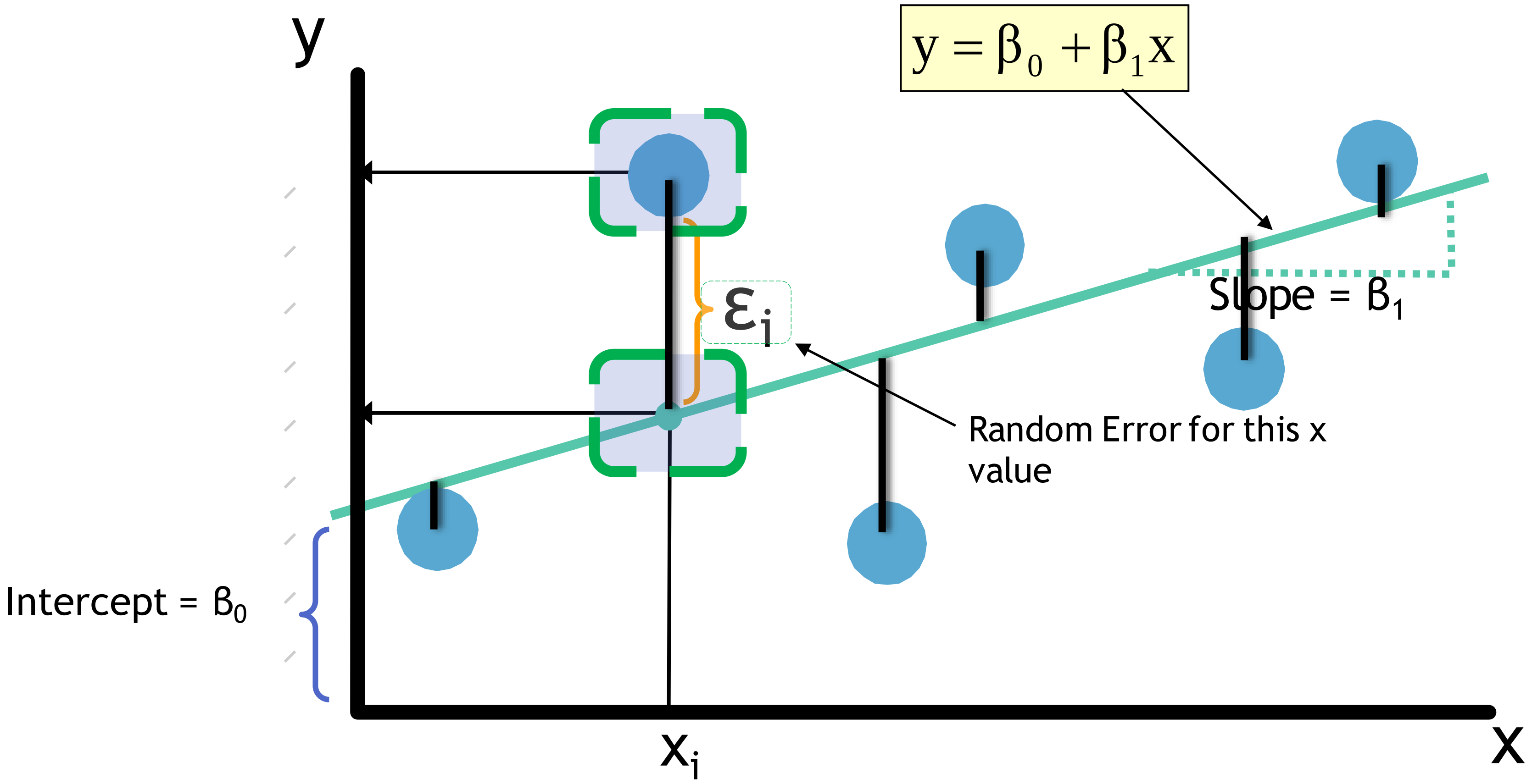
- What is the best fit for our data?
- The one which goes through the core of the data
- The one which minimizes the error

$$\boxed{y = \beta_0 + \beta_1 x}$$

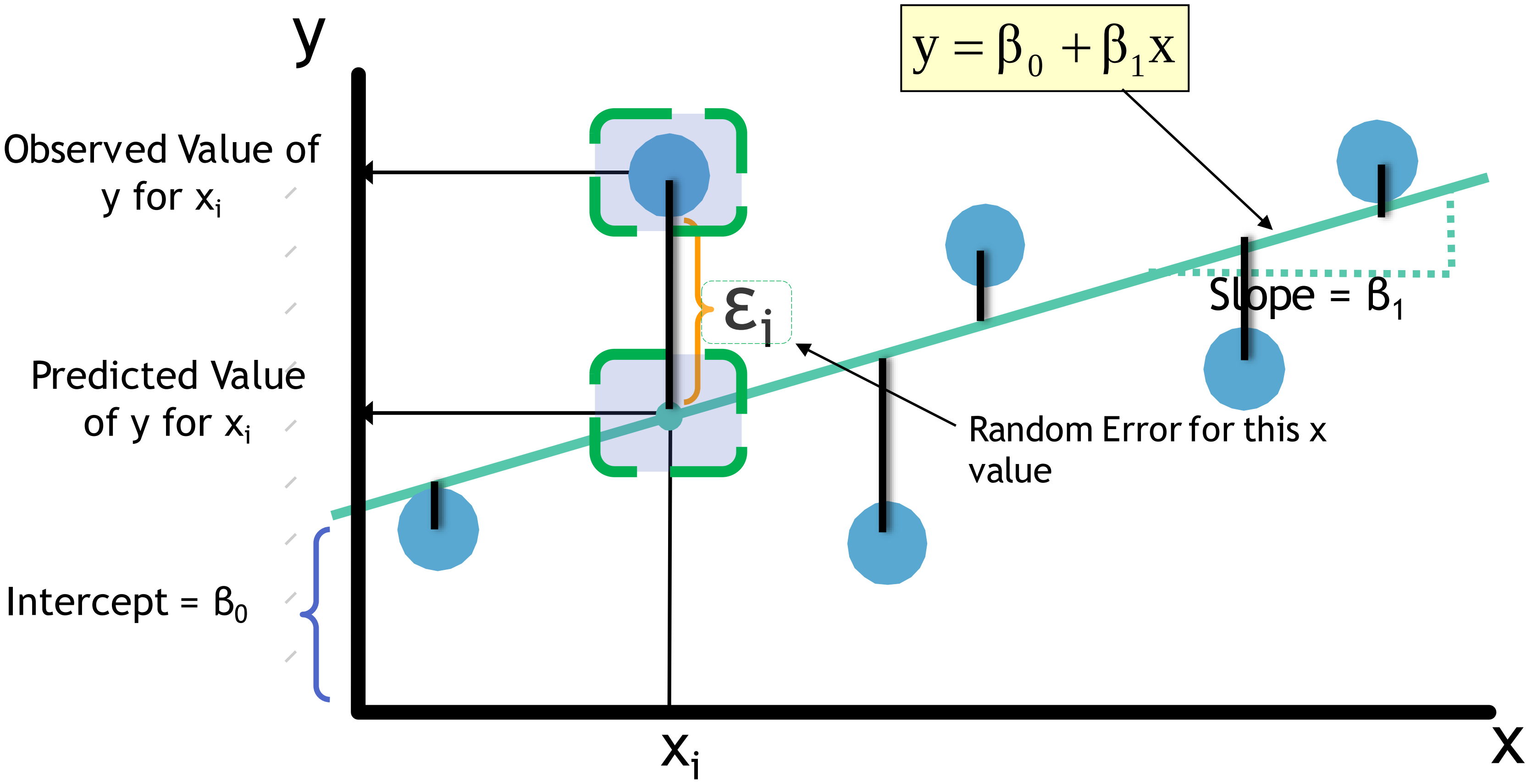
statinfer

Regression Line fitting-Least Squares Estimation

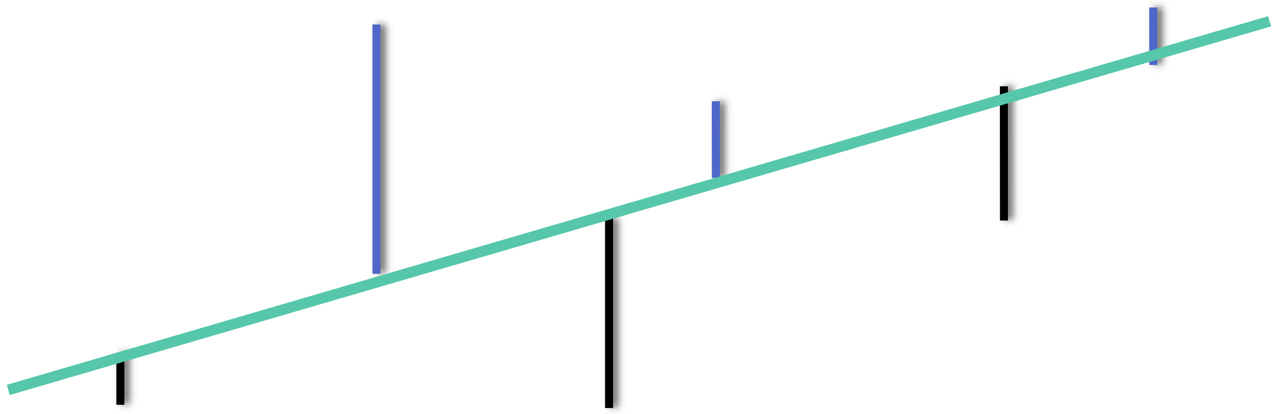
Regression Line fitting



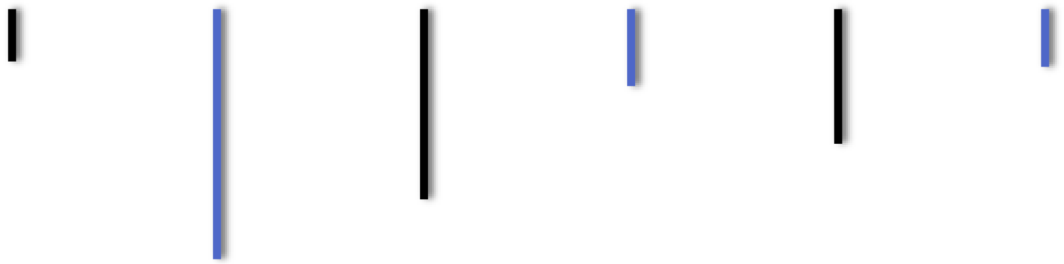
Regression Line fitting



Regression Line fitting



Minimizing the error



- The best line will have the minimum error
- Some errors are positive and some errors are negative. Taking their sum is not a good idea
- We can either minimize the squared sum of errors Or we can minimize the absolute sum of errors
- Squared sum of errors is mathematically convenient to minimize
- The method of minimizing squared sum of errors is called least squared method of regression

Least Squares Estimation

- X: x1, x2, x3, x4, x5, x6, x7,……..
- Y:y1, y2, y3, y4, y5, y6, y7…….
- Imagine a line through all the points
- Deviation from each point (residual or error)
- Square of the deviation
- Minimizing sum of squares of deviation

$$\sum e^2 = \sum (y - \hat{y})^2 = \sum (y - (\beta_0 + \beta_1 x))^2$$

β_0 and β_1 are obtained by minimize the sum of the squared residuals

LAB: Regression Line Fitting

- Dataset: Air Travel Data\Air_travel.csv
- Find the correlation between Promotion_Budget and Passengers
- Draw a scatter plot between Promotion_Budget and Passengers. Is there any pattern between Promotion_Budget and Passengers?

statinfer

Final Model – Predictive Model

$$\boxed{y = \beta_0 + \beta_1 x}$$

Code: Regression Line Fitting

```
import statsmodels.formula.api as sm
model = sm.ols(formula='Passengers ~ Promotion_Budget', data=air)
fitted1 = model.fit()
fitted1.summary()
```

Regression Line Equation – ML Model

statinfer

How good is my regression line?

Two models

- Model-1 : Passengers vs. Promo budget
- Model-2: Passengers vs. inter metro flight ratio

- Model-1 vs Model-2 to predict the same target. Which model to pick?

How good is my regression line?

Model-1

X1	Y Actual	Y Pred
	30K	31K
	40K	39K
	35K	35K
	27K	26K
	32K	32K
	33K	35K
	28K	26K

Model-2

X2	Y Actual	Y Pred
	30K	42K
	40K	49K
	35K	15K
	27K	20K
	32K	32K
	33K	38K
	28K	20K

SSE

X1	Y Actual	Y Pred	Error
	30K	31K	
	40K	39K	
	35K	35K	
	27K	26K	1K
	32K	32K	
	33K	35K	
	28K	26K	

SSE

X1	Y Actual	Y Pred	Error
	30K	31K	-1K
	40K	39K	1K
	35K	35K	0K
	27K	26K	1K
	32K	32K	0K
	33K	35K	-2K
	28K	26K	2K

SSE

X1	Y Actual	Y Pred	Error	Squared Error
	30K	31K	-1K	
	40K	39K	1K	
	35K	35K	0K	
	27K	26K	1K	
	32K	32K	0K	
	33K	35K	-2K	
	28K	26K	2K	
				SSE

SSE, SSR and SST

X1	Y Actual	Y Pred	Error	Squared Error
	30K	31K	-1K	
	40K	39K	1K	
	35K	35K	0K	
	27K	26K	1K	
	32K	32K	0K	
	33K	35K	-2K	
	28K	26K	2K	
	SST	SSR		SSE

How good is my regression line?

- Take an (x,y) point from data.
- Imagine that we submitted x in the regression line, we got a prediction as y_{pred}
- If the regression line is a good fit then the we expect y_{pred}=y or ($y-y_{pred}$) =0
- At every point of x, if we repeat the same, then we will get multiple error values ($y-y_{pred}$) values
- Some of them might be positive, some of them may be negative, so we can take the square of all such errors

$$\boxed{SSE = \sum (y - \hat{y})^2}$$

SSE

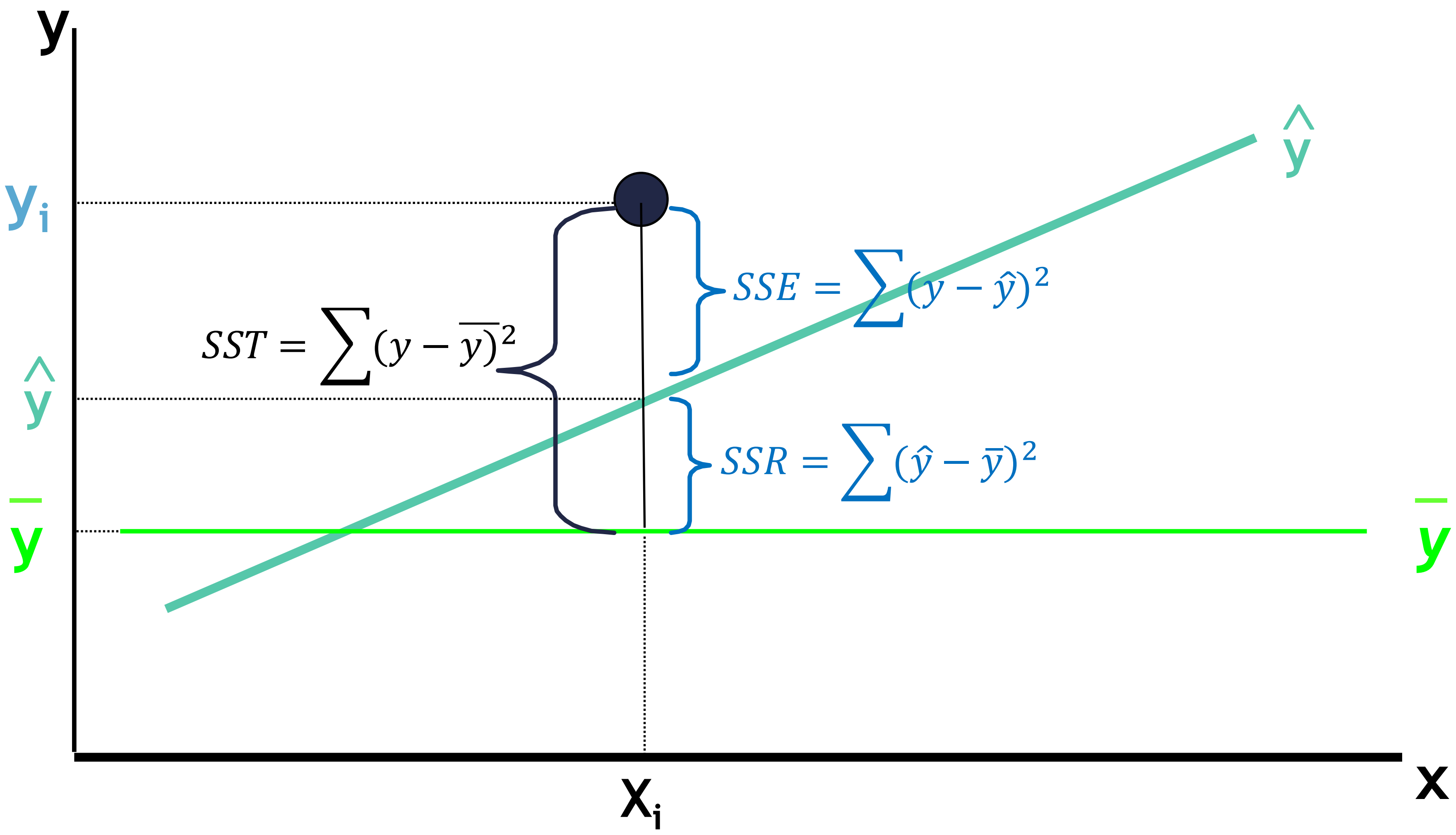
- For a good model we need SSE to be zero or near to zero
- Standalone SSE will not make any sense, For example SSE= 100, is very less when y is varying in terms of 1000's. Same value is is very high when y is varying in terms of decimals.
- We have to consider variance of y while calculating the regression line accuracy

$$\boxed{SSE = \sum (y - \hat{y})^2}$$

How good is my regression line?

- Error Sum of squares (SSE- Sum of Squares of error)
  - $SSE = \sum(y - \hat{y})^2$
- Total Variance in Y (SST- Sum of Squares of Total)
  - $SST = \sum(y - \overline{y})^2$
  - $SST = \sum(y - \hat{y} + \hat{y} - \overline{y})^2$
  - $SST = \sum(y - \hat{y} + \hat{y} - \overline{y})^2$
  - $SST = \sum(y - \hat{y})^2 + \sum(\hat{y} - \bar{y})^2$
  - $\mathrm{SST} = \mathrm{SSE} + \sum(\hat{y} - \bar{y})^2$
  - $\mathrm{SST} = \mathrm{SSE} + SSR$
- So, total variance in Y is divided into two parts,
  - Variance that can’t be explained by x (error)
  - Variance that can be explained by x, using regression

Explained and Unexplained Variation



How good is my regression line?

- So, total variance in Y is divided into two parts,
  - Variance that can be explained by x, using regression
  - Variance that can't be explained by x

$$SST = SSR + SSE$$

Total sum of Squares	Sum of Squares Regression	Sum of Squares Error
$SST = \sum (y - \bar{y})^2$	$SSR = \sum (\hat{y} - \bar{y})^2$	$SSE = \sum (y - \hat{y})^2$

R-Squared

R-Squared

- A good fit will have
  - SSE (Minimum or Maximum?)
  - SSR (Minimum or Maximum?)
  - And we know SST= SSE + SSR
  - SSE/SST(Minimum or Maximum?)
  - SSR/SST(Minimum or Maximum?)
- The coefficient of determination is 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

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

where

$$\boxed{0 \le R^2 \le 1}$$

Lab: R- Square

- What is the R-square value of Passengers vs Promotion_Budget model?
- What is the R-square value of Passengers vs Inter_metro_flight_ratio

Code: R- Square

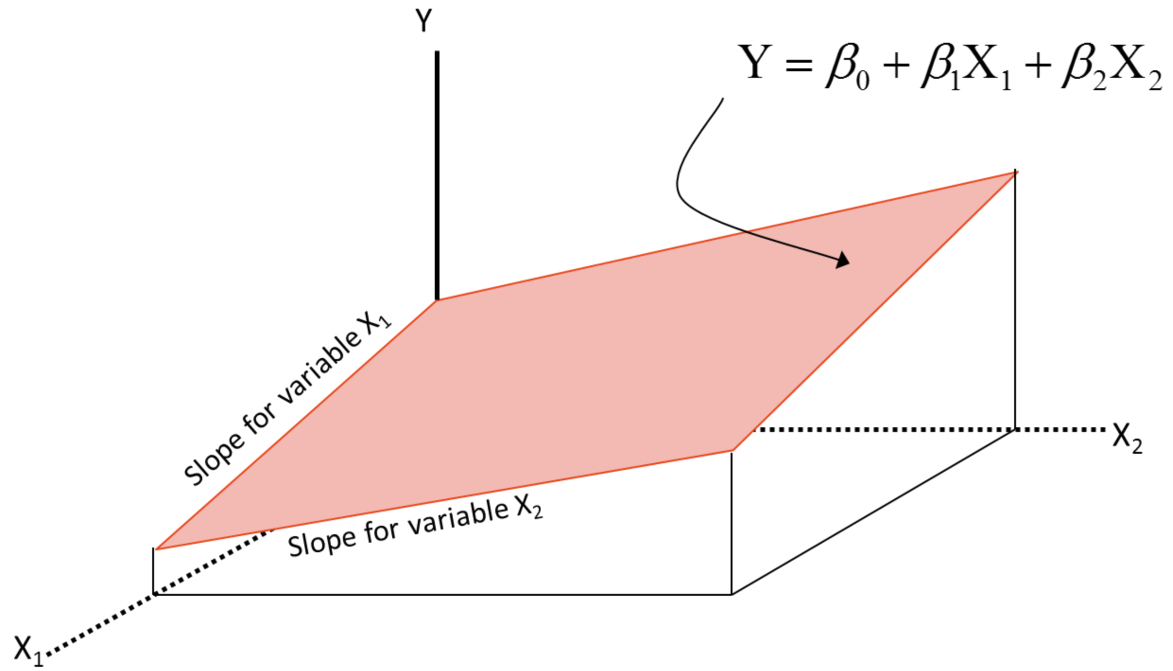
```
#What is the R-square value of Passengers vs Promotion_Budget model?
fitted1.summary()

#What is the R-square value of Passengers vs Inter_metro_flight_ratio
fitted2.summary()
```

Multiple Regression

Multiple Regression

- Using multiple predictor variables instead of single variable
- We need to find a perfect plane here



Code-Multiple Regression

```
import statsmodels.formula.api as sm

model = sm.ols(formula='Passengers ~ Promotion_Budget +
Inter_metro_flight_ratio + Service_Quality_Score ', data=air)

fitted = model.fit()
fitted.summary()
```

Individual Impact of variables

Individual Impact of variables

- Look at the P-value
- Probability of the hypothesis being right.
- Individual variable coefficient is tested for significance
- Beta coefficients follow t distribution.
- Individual P values tell us about the significance of each variable
- A variable is significant if P value is less than 5%. Lesser the P-value, better the variable
- Note it is possible all the variables in a regression to produce great individual fits, and yet very few of the variables be individually significant.

To test

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

Test statistic:

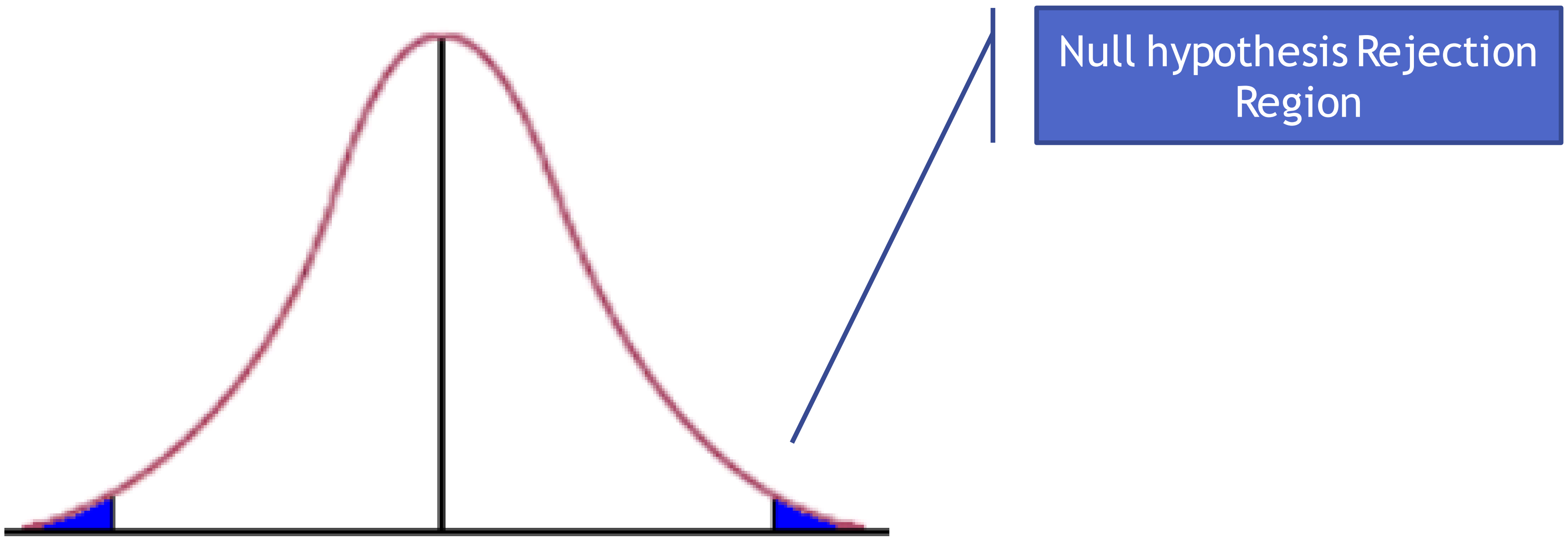
$$t = \frac{b_i}{s(b_i)}$$

Reject H_0 if

$$t > t(\frac{\alpha}{2}; n-k-1) \quad or$$
$$t < -t(\frac{\alpha}{2}; n-k-1)$$

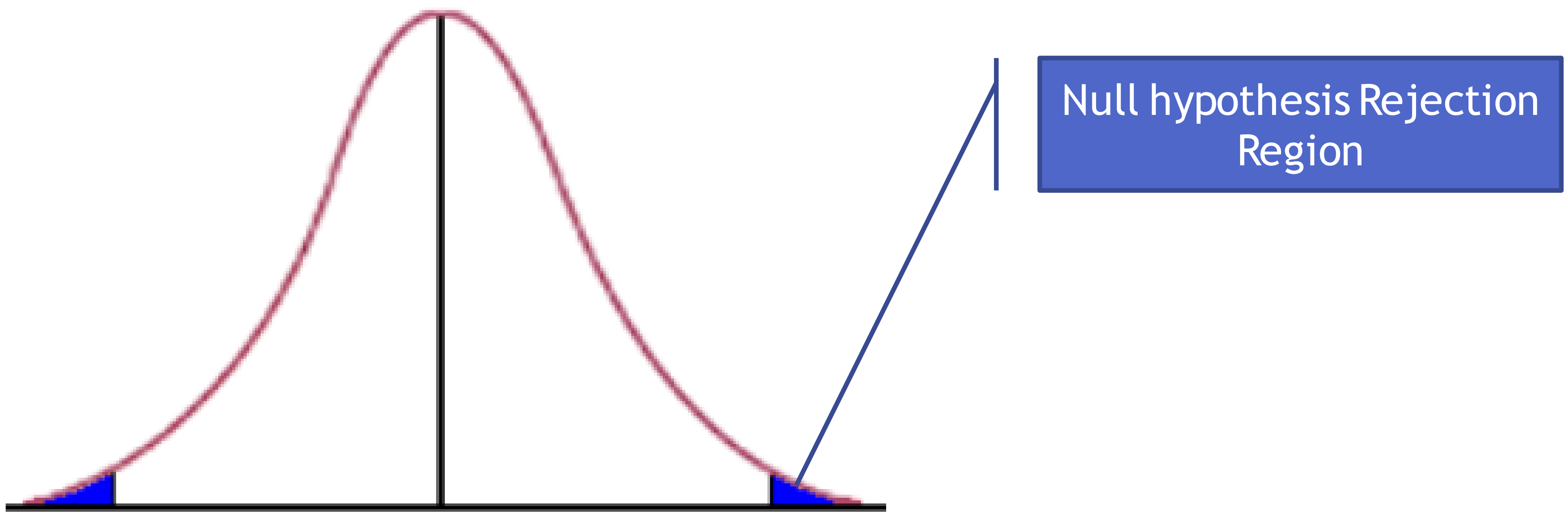
What is testing?

- Population – 1 million soaps
- Sample – 100 soaps
- Null Hypothesis – Population avg weight of soap is 250 grams
- Test statistic – avg weight of soaps in sample



Individual Impact of variables

- Beta coefficients follow t-distribution under null hypothesis.



LAB: Multiple Regression

- Build a multiple regression model to predict the number of passengers use three predictor variables
  - Promotion_Budget
  - Service_Quality_Score
  - Inter_metro_flight_ratio
- What is R-square value
- Are there any predictor variables that are not impacting the dependent variable
- Drop least impacting variable and rebuild the model. What is the

Code: Multiple Regression

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr.fit(air[["Promotion_Budget"]+["Inter_metro_flight_ratio"]+["Service_Quality_Score"]], air[["Passengers"]])
predictions = lr.predict(air[["Promotion_Budget"]+["Inter_metro_flight_ratio"]+["Service_Quality_Score"]])
predictions

import statsmodels.formula.api as sm
model = sm.ols(formula='Passengers ~ Promotion_Budget+Service_Quality_Score+Inter_metro_flight_ratio', data=air)
fitted = model.fit()
fitted.summary()
```

Adjusted R-Squared

LAB: Adjusted R-Square

- Dataset: “Adjusted Rsquare/ Adj_Sample.csv”
- Build a model to predict y using x1,x2 and x3. Note down R-Square and Adj R-Square values
- Build a model to predict y using x1,x2,x3,x4,x5 and x6. Note down R-Square and Adj R-Square values
- Build a model to predict y using x1,x2,x3,x4,x5,x6,x7 and x8. Note down R-Square and Adj R-Square values

Code: Adjusted R-Square

```
##Adjusted R-Square

adj_sample=pd.read_csv("D:\\Google Drive\\Training\\Datasets\\Adjusted RSquare\\Adj_Sample.csv")
#Build a model to predict y using x1,x2 and x3. Note down R-Square and Adj R-Square values
model = sm.ols(formula='Y ~ x1+x2+x3', data=adj_sample)
fitted = model.fit()
fitted.summary()
#R-Squared

#Model2
model = sm.ols(formula='Y ~ x1+x2+x3+x4+x5+x6', data=adj_sample)
fitted = model.fit()
fitted.summary()

#Model3
model = sm.ols(formula='Y ~ x1+x2+x3+x4+x5+x6+x7+x8', data=adj_sample)
fitted = model.fit()
fitted.summary()
```

	R Squared	Adj R-Squared
Y vs x1, x2, x3	68%	56%
Y vs x1, x2…x6	71%	37%
Y vs x1,x2 …x8	80%	28%

Adjusted R-Squared

- Is it good to have as many independent variables as possible? Nope
- R-square is deceptive. R-squared never decreases when a new X variable is added to the model – True?
- We need a better measure or an adjustment to the original R-squared formula.
- Adjusted R squared
  - Its value depends on the number of explanatory variables
  - Imposes a penalty for adding additional explanatory variables
  - It is usually written as (R-bar squared)
  - Very different from R when there are too many predictors and n is less

$$\bar{R}^2 = R^2 - \frac{k-1}{n-k}(1-R^2)$$

n-number of observations, k-number of parameters

statinfer

Multiple Regression

LAB: Multiple Regression

- Import Regional Sales Data
- Build a model to predict the sales
- Write down your observations
- What is the relation between Avg expenses and Regional sales?

Code: Multiple Regression Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Regional_Sales   R-squared:                       0.845
Model:                              OLS   Adj. R-squared:                  0.838
Method:                   Least Squares   F-statistic:                     124.0
Date:                  Thu, 29 Oct 2020   Prob (F-statistic):           5.71e-36
Time:                          05:14:03   Log-Likelihood:                -955.00
No. Observations:                    96   AIC:                             1920.
Df Residuals:                        91   BIC:                             1933.
Df Model:                             4
Covariance Type:              nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept      -4.318e+04   1.95e+04     -2.211      0.030    -8.2e+04   -4386.455
Avg Income        27.4480     13.252      2.071      0.041       1.125      53.771
Avg_Expenses     -26.2249     20.191     -1.299      0.197     -66.332      13.883
Percent_Male     414.4477    205.405      2.018      0.047       6.436     822.459
Percent_Female   429.5928    181.552      2.366      0.020      68.962     790.223
==============================================================================
```

Code: Multiple Regression Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Regional_Sales   R-squared:                       0.838
Model:                             OLS   Adj. R-squared:                  0.832
Method:                  Least Squares   F-statistic:                     158.3
Date:                 Thu, 29 Oct 2020   Prob (F-statistic):           3.35e-36
Time:                         05:42:24   Log-Likelihood:                -957.21
No. Observations:                   96   AIC:                             1922.
Df Residuals:                       92   BIC:                             1933.
Df Model:                            3
Covariance Type:             nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept      -3.879e+04   1.98e+04     -1.964      0.053    -7.8e+04     444.025
Avg_Expenses      15.5707      0.723     21.549      0.000      14.136      17.006
Percent_Male     395.6785    208.842      1.895      0.061     -19.100     810.457
Percent_Female   399.9817    184.196      2.172      0.032      34.153     765.811
==============================================================================
```

Multicollinearity

Multicollinearity

- Multiple regression is wonderful - In that it allows you to consider the effect of multiple variables simultaneously.
- Multiple regression is extremely unpleasant -Because it allows you to consider the effect of multiple variables simultaneously.
- The relationships between the explanatory variables are the key to understanding multiple regression.
- Multicollinearity (or inter correlation) exists when at least some of the predictor variables are correlated among themselves.
- The parameter estimates will have inflated variance in presence of multicollineraity
- Sometimes the signs of the parameter estimates tend to change
- If the relation between the independent variables grows really strong then the variance of parameter estimates tends to be infinity – Can you prove it?

Multicollinearity - Example

- **Y=X1 + 2X2 -X3**

X1=2X3

Y=X1 + 2X2 +X3 -2X3
Y=X1+2X2 + X3 -X1
Y=2X2+X3

Y=-X1 +2X1 +2X2 -X3
Y=-X1 + 4X3 + 2X2 -X3
Y=-X1+ 3X3 +2X2

Multicollinearity detection

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

- Build a model X1 vs X2 X3 X4 find R square, say R1
- Build a model X2 vs X1 X3 X4 find R square, say R2
- Build a model X3 vs X1 X2 X4 find R square, say R3
- Build a model X4 vs X1 X2 X3 find R square, say R4

- For example if R3 is 95% then we don’t really need X3 in the model
- Since it can be explained as liner combination of other three
- For each variable we find individual R square.
- $1/(1-R^2)$ is called VIF.
- VIF option in SAS automatically calculates VIF values for each of the predictor variables

R Square	40%	50%	60%	70%	75%	80%	90%
VIF	1.67	2.00	2.50	3.33	4.00	5.00	10.00

LAB: Multicollinearity

- Identify the Multicollinearity in the Regional Sales Data
- Drop the variable one by one to reduce the multicollinearity

Code: Multicollinearity

```
def vif_cal(input_data):
    x_vars = input_data
    xvar_names=x_vars.columns
    for i in range(0,xvar_names.shape[0]):
        y=x_vars[xvar_names[i]]
        x=x_vars[xvar_names.drop(xvar_names[i])]
        rsq=sm.ols(formula="y~x", data=x_vars).fit().rsquared
        vif=round(1/(1-rsq),2)
        print (xvar_names[i], " VIF = " , vif)
```

Code: Multicollinearity

```
X_Data=regional_sales.drop(["Region_id","Regional_Sales"],axis=1)
vif_cal(input_data=X_Data)
```

statinfer

Multiple Regression model building

Steps in Building Regression Model

1. Build benchmark model with all the variables
2. Check for R-Squared value (>80%)
3. Adj- Rsquare (should be near to R-Square)
4. P-Value for variable impact
    1. If p<0.05 Impactful - then keep variable
    2. If p>=0.05 Not Impactful - then drop variable
5. VIF for variable independence
    1. If vif < 5 Independence - then keep variable
    2. If vif >=5 Dependent -then drop variables

Lab: Multiple Regression

- Dataset: Webpage_Product_Sales/Webpage_Product_Sales.csv
- Build a model to predict sales using rest of the variables
- Drop the less impacting variables based on p-values.
- Is there any multicollinearity?
- How many variables are there in the final model?
- What is the R-squared of the final model?
- Can you improve the model using same data and variables?

statinfer

Conclusion - Regression

Conclusion - Regression

- We discussed the basic concepts of correlation, regression
- Adjusted R-squared is a good measure of training/in time sample error. We can't be sure about the final model performance based on this. We may have to perform cross-validation to get an idea on testing error.
- Outlies can influence the regression line, we need to take care of data sanitization before building the regression line.

statinfer

Thank you