

Regression

Introduction

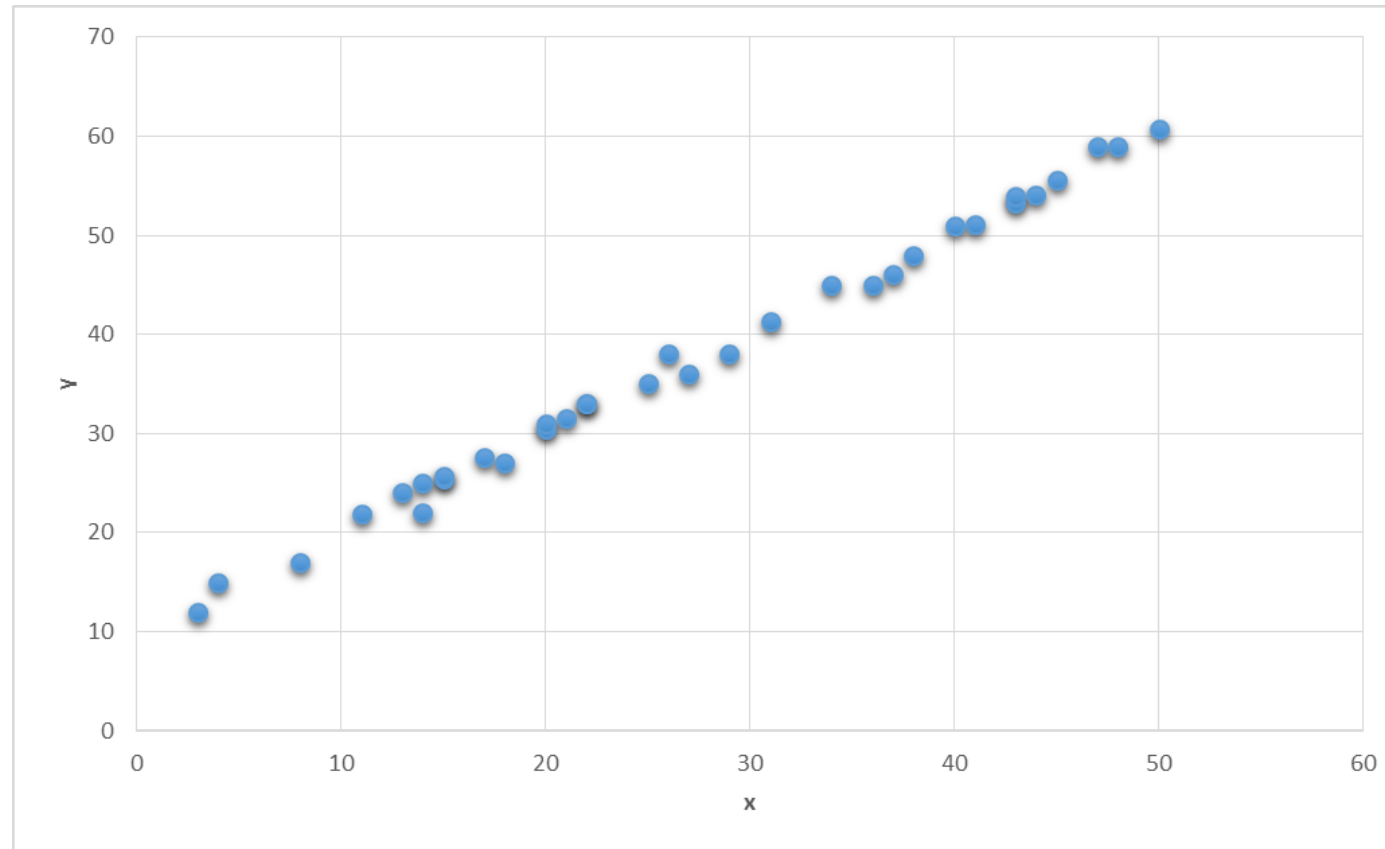
Contents

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

Correlation

What is Regression

Regression



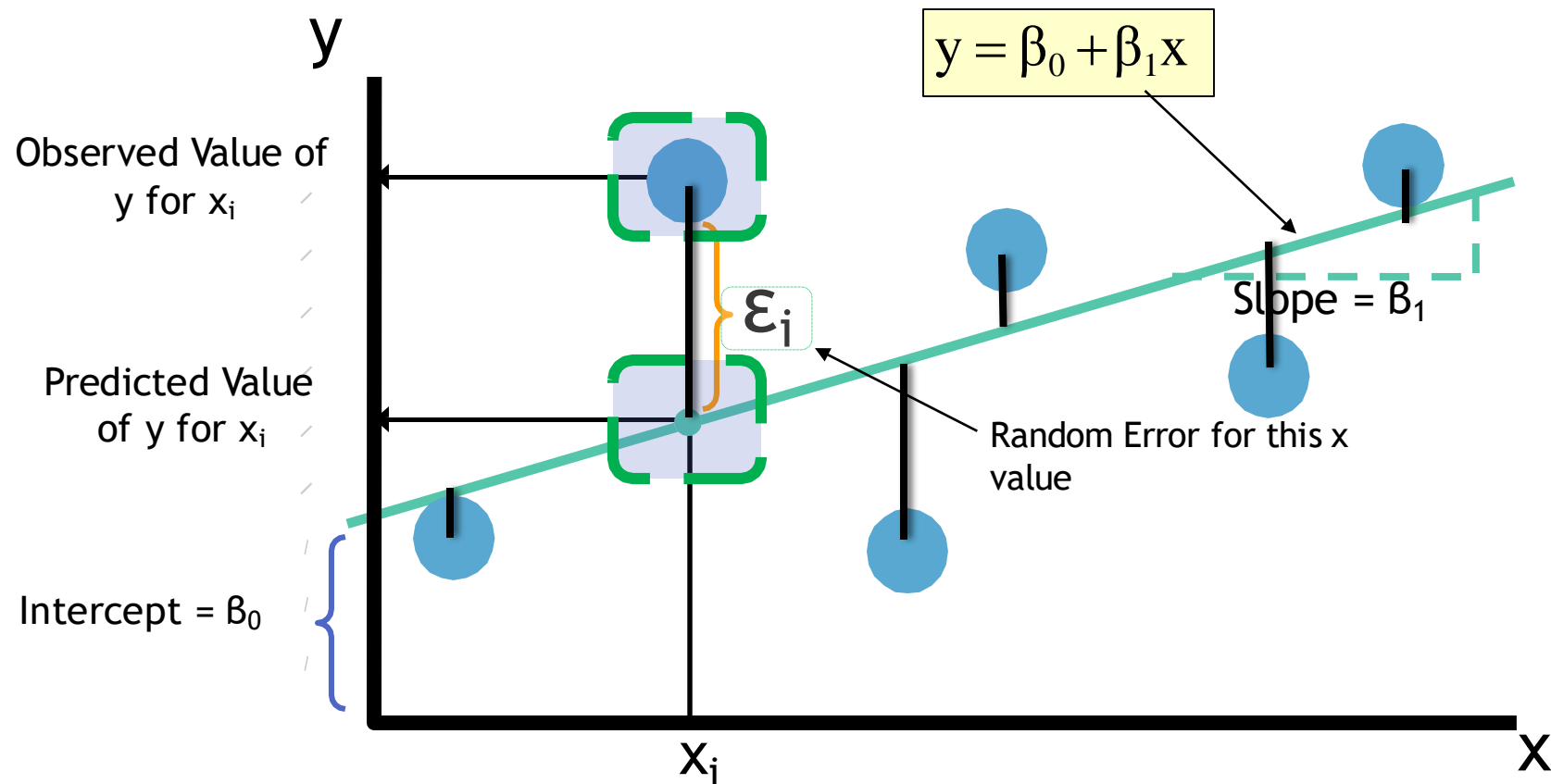
What is Regression

- 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

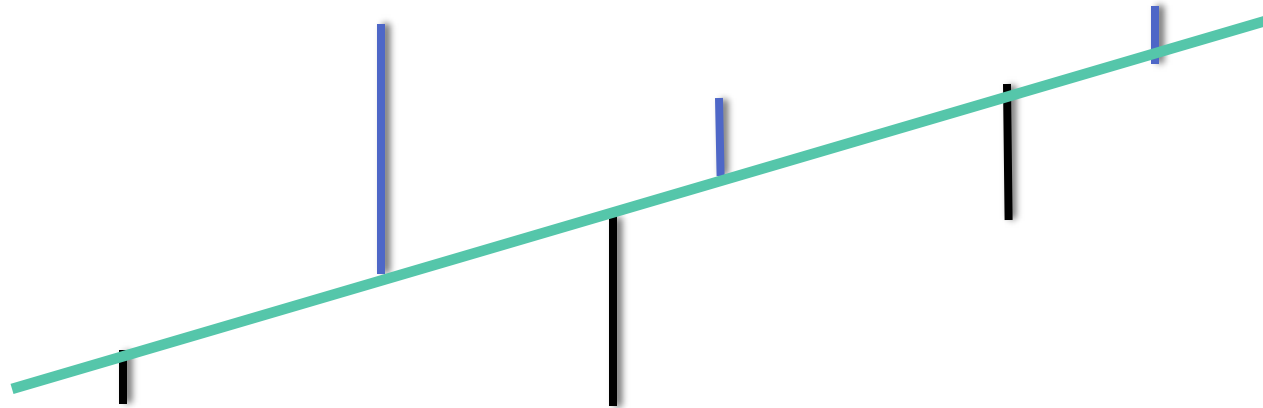
$$y = \beta_0 + \beta_1 x$$

Regression Line fitting-Least Squares Estimation

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: $x_1, x_2, x_3, x_4, x_5, x_6, x_7, \dots$
- Y: $y_1, y_2, y_3, y_4, y_5, y_6, y_7, \dots$
- Imagine a line through all the points
- Deviation from each point (residual or error)
- Square of the deviation
- Minimizing sum of squares of deviation

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

β_0 and β_1 are obtained by [minimize the sum of the squared residuals](#)

How good is my regression
line?

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 we expect $y_{\text{pred}}=y$ or $(y - y_{\text{pred}}) = 0$
- At every point of x , if we repeat the same, then we will get multiple error values $(y - y_{\text{pred}})$ values
- Some of them might be positive, some of them may be negative, so we can take the square of all such errors

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

SSE

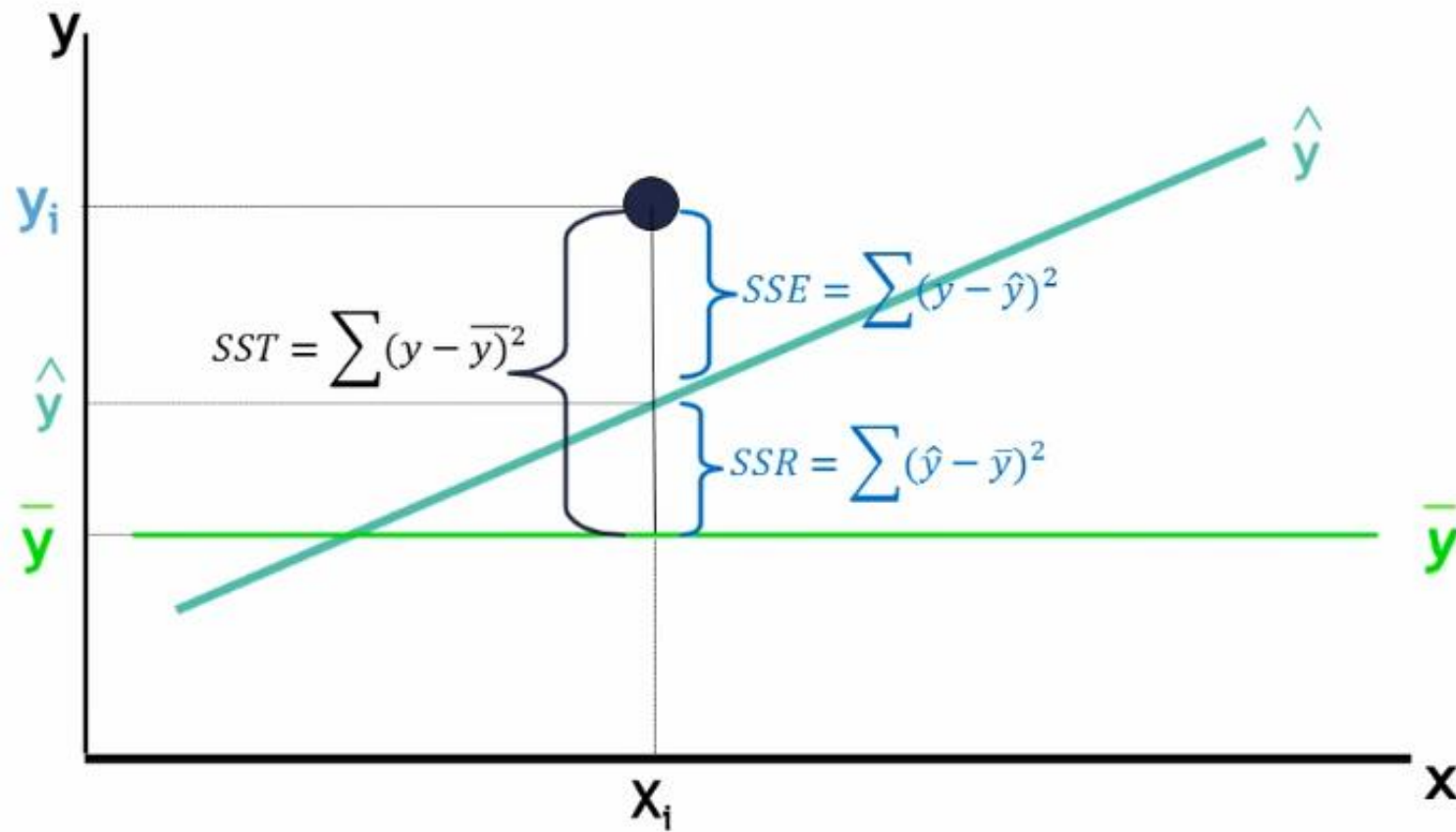
- For a good model we need SSE to be zero or near to zero
- Standalone SSE will not make any sense, For example SSE= 100, is very less when y is varying in terms of 1000's. Same value is is very high when y is varying in terms of decimals.
- We have to consider variance of y while calculating the regression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

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

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)
 - $SST = \sum (y - \bar{y})^2$
 - $SST = \sum (y - \hat{y} + \hat{y} - \bar{y})^2$
 - $SST = \sum (y - \hat{y} + \hat{y} - \bar{y})^2$
 - $SST = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2$
 - $SST = SSE + \sum (\hat{y} - \bar{y})^2$
 - $SST = SSE + SSR$
- 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	=	SSE	+	SSR
<div>• Total sum of Squares</div>		<div>Sum of Squares Error</div>		<div>Sum of Squares Regression</div>
$SST = \sum (y - \bar{y})^2$		$SSE = \sum (y - \hat{y})^2$		$SSR = \sum (\hat{y} - \bar{y})^2$

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

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

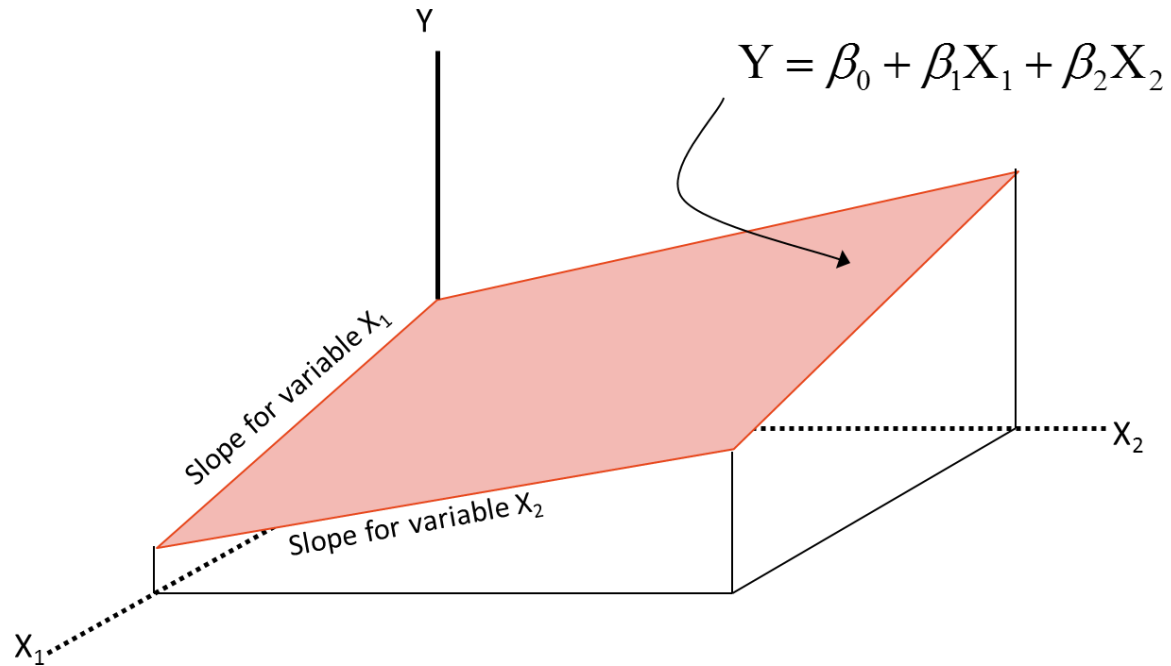
where

$$0 \leq R^2 \leq 1$$

Multiple Regression

Multiple Regression

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



Code-Multiple Regression

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

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

Part-12: 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

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

$$t < -t\left(\frac{\alpha}{2}; n - k - 1\right)$$

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 x_1, x_2 and x_3 . Note down R-Square and Adj R-Square values
- Build a model to predict y using x_1, x_2, x_3, x_4, x_5 and x_6 . Note down R-Square and Adj R-Square values
- Build a model to predict y using $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ and x_8 . 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  
R-Square\\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()
```

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

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

n-number of observations, k-number of parameters

Multiple Regression- issues

LAB: Multiple Regression- issues

- Import Final Exam Score data
- Build a model to predict final score using the rest of the variables.
- How are Sem2_Math & Final score related? As Sem2_Math score increases, what happens to Final score?
- Remove “Sem1_Math” variable from the model and rebuild the model
- Is there any change in R square or Adj R square
- How are Sem2_Math & Final score related now? As Sem2_Math score increases, what happens to Final score?
- Draw a scatter plot between Sem1_Math & Sem2_Math
- Find the correlation between Sem1_Math & Sem2_Math

Code: Multiple Regression- issues

```
import statsmodels.formula.api as sm
model1 = sm.ols(formula='Final_exam_marks ~
Sem1_Science+Sem2_Science+Sem1_Math+Sem2_Math', data=final_exam)
fitted1 = model1.fit()
fitted1.summary()
fitted1.rsquared
```

```
import statsmodels.formula.api as sm
model2 = sm.ols(formula='Final_exam_marks ~
Sem1_Science+Sem2_Science+Sem2_Math', data=final_exam)
fitted2 = model2.fit()
fitted2.summary()
```

Part-15: 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 multicollinearity
- 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 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

Multiple Regression model building

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.
- Outliers can influence the regression line, we need to take care of data sanitization before building the regression line.

Thank you