

Modul 3

Regression

Data Science Program

What is Regression?

- Tool for finding existence of an association relationship between a dependent variable (Y) and one or more independent variables (X_1, X_2, \dots, X_n)
- Relationship can be linear or non-linear

Mathematical vs Statistical Relationship

- Mathematical is an exact relationship

$$Y = \beta_0 + \beta_1 X$$

- Statistical is NOT an exact relationship

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Regression Dictionary

- Dependent variable (*response variable*) measures an outcome of a study (can also be called *outcome variable*)
- Independent variables (*explanatory variables*) explain changes in a response variable
- Given set values of explanatory variable to see how it affects response variable
-> predict response variable

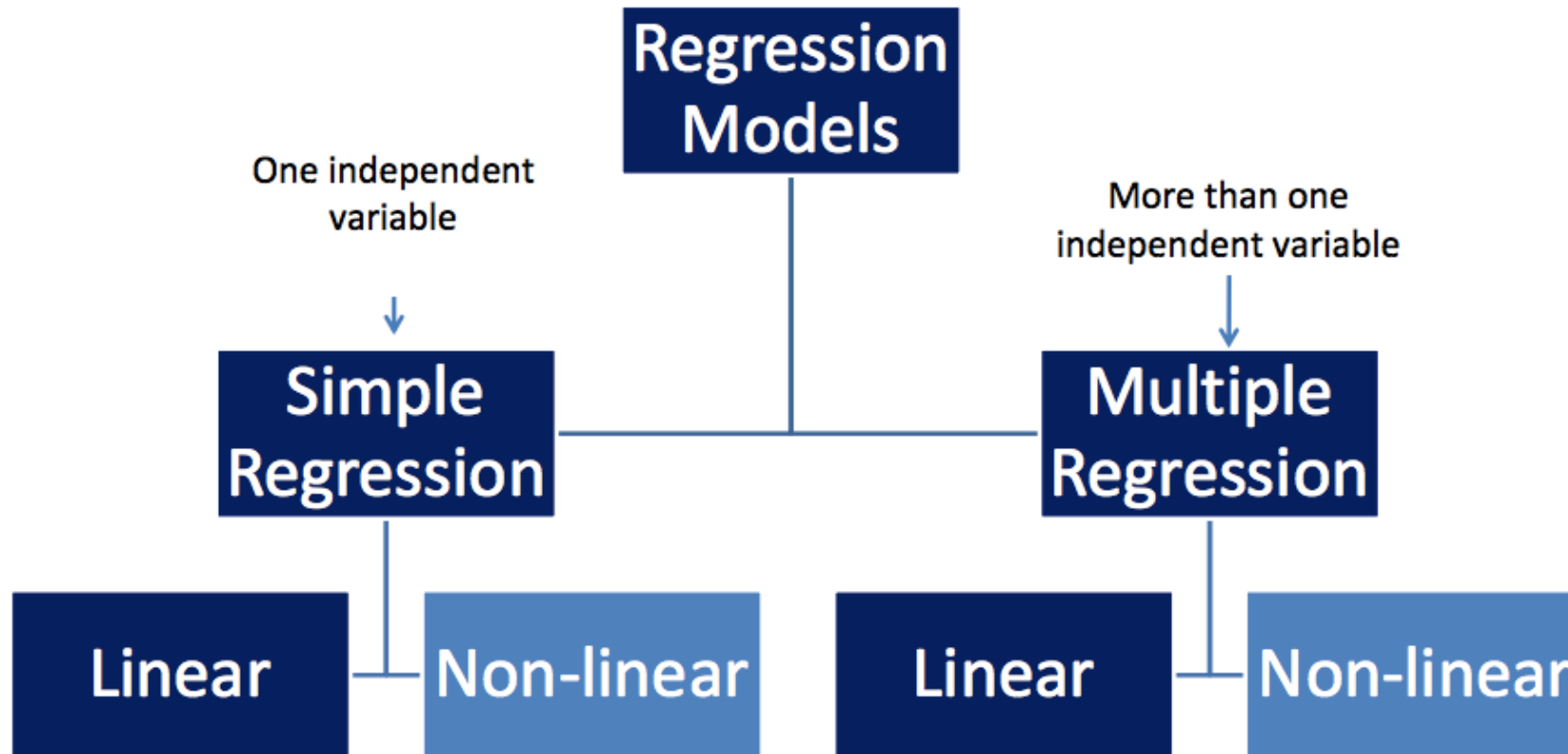
Dependent and Independent Variables

- Terms dependent and independent does not necessarily imply a causal relationship between two variables
- Regression is NOT to capture causality
- Purpose regression: predict the value of dependent variable given the values of independent variables

Why we need Regression?

- Companies would like to know about factors that have significant impact on their key performance indicators.
- Helps to create new hypothesis that assist companies to improve their performance and hence better decision making

Types of Regression (1)



Types of Regression (2)

- Simple Linear Regression

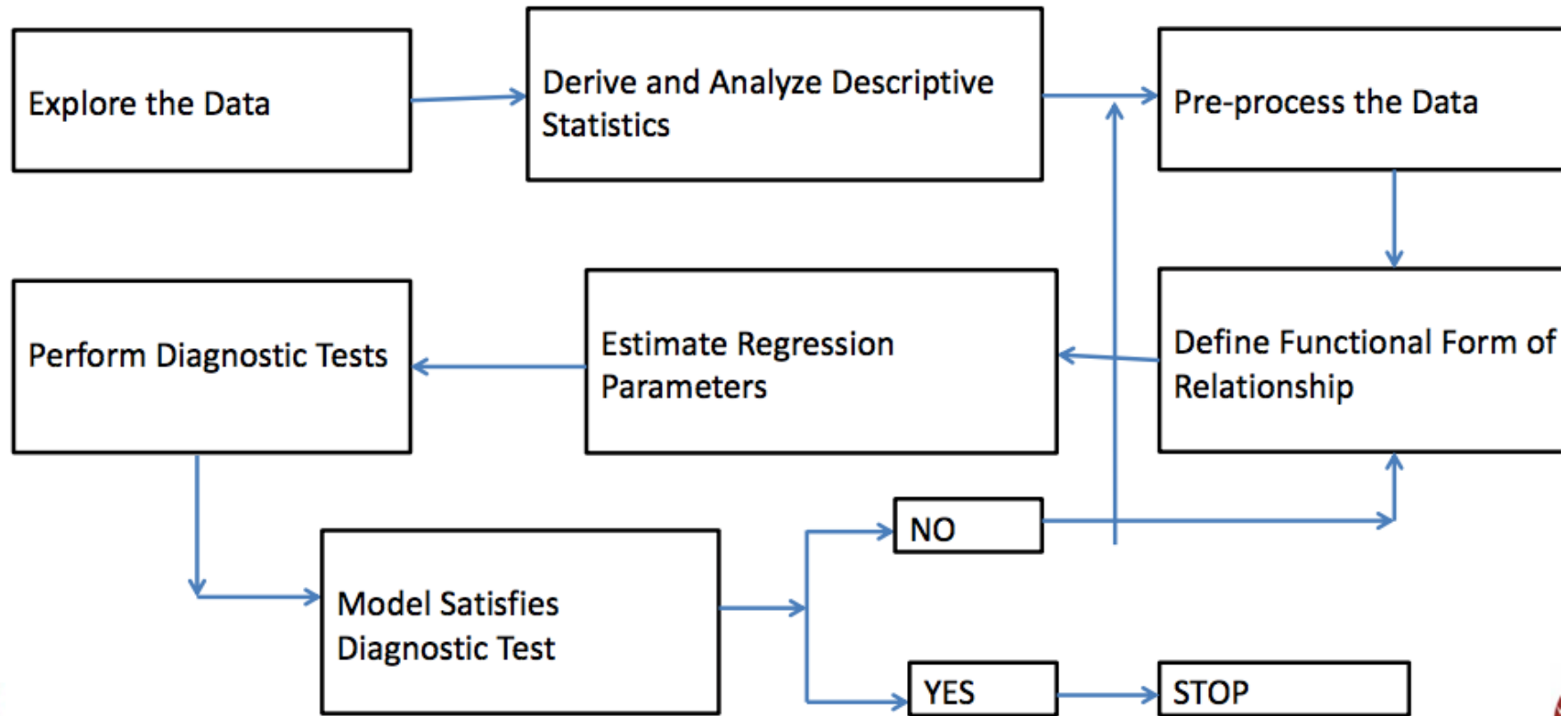
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

- Importa

Regression Model Development



Model Building

- Identify explanatory variable
- Specify the nature of relationship between dependent variable and explanatory variables

Linear Regression Model

- Relationship between variables is a linear function

The diagram illustrates the Linear Regression Model equation: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. The equation is centered on a dark blue background. Labels with yellow arrows point to each part of the equation: 'Population Y-Intercept' points to β_0 , 'Population Slope' points to β_1 , 'Random Error' points to ε_i (which is circled in red), 'Dependent (Response)' points to Y_i , and 'Independent (Explanatory)' points to X_i .

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Population Y-Intercept

Population Slope

Random Error

Dependent (Response)

Independent (Explanatory)

Linear Regression Model Assumption

Assumptions	Impacts on the Regression Model
Error follows normal distribution	This condition is necessary for reliability of statistical tests (<i>t</i> and <i>F</i>).
Homoscedasticity (constant variance)	It is necessary for statistical tests (<i>F</i> and <i>t</i>).
Multi-collinearity	Inflates standard error of estimate of regression coefficients (beta coefficients). May <i>reject</i> significant variables.
Auto-correlation	Underestimates standard error of estimates of regression coefficients. May <i>accept</i> insignificant variables.

Estimation of Parameters in Regression

- Least squares function is given by:

$$SSE = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2$$

Coefficient Equations

- Prediction Equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sample Slope:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- Sample Y-intercept:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Why Least Squares Estimate?

- OLS beta estimates provided the error terms are uncorrelated (no auto regression) and have equal variance (homoscedasticity)

- it implies that $E\{\beta - \hat{\beta}\} = 0$ where β is the population parameter, $\hat{\beta}$ is the OLS estimate

Interpret Beta

- Interpretation depends on the functional form of the relationship between response and explanatory variables.
- Coefficients Interpretation:
 - The intercept, β_0 is the mean value of the dependent variable Y, when the independent variable $X = 0$
 - The slope, β_1 is the change in the value of dependent variable Y, for unit change in the independent variable X

Interpret Coefficient


- Regression parameter values are valid only in the range of data that we use for developing model
- β_0 is valid for a given range which we find in the sample itself

- $\ln(Y) = \beta_0 + \beta_1 \ln(X)$ %age change in X

β_1

Simple Linear Regression

Variable x and y has Linear relationship	Assumption of the world
$y = \beta_0 + \beta_1 x + \varepsilon$, Minimize SSE	Fitting a model
Is x really related to y ? Is β_1 statistically significant?	Validating the model
Predict y for a given x .	Using a model



Model validation

- Use of co-efficient of determination to check the goodness of fit of regression
- Analysis of Variance (ANOVA) and F-test to check the overall fitness of the regression model
- T-test to validate relationship between dependent and independent variables
- Residual analysis to check the model adequacies

Coefficient of Determination

- Measure of how well the regression line fits the data
- Coefficient of determination (R^2 square) lies between 0 and 1
- Percentage of variation that can be explained by the regression model

Variation in Y (1)

$$Y_i = \beta_0 + \beta_1 X + \varepsilon_i$$

$$\text{Variation in } Y_i = \text{Systemic Variation} + \text{Random Variation}$$

or

$$\text{Variation in } Y_i = \text{Explained Variation} + \text{Unexplained Variation}$$

Variation in Y (2)

$$\begin{array}{ccccc} Y_i - \bar{Y} & = & \hat{Y}_i - \bar{Y} & + & Y_i - \hat{Y}_i \\ \text{Total variation} & & \text{Explained variation} & & \text{Unexplained variation} \end{array}$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

SST *SSR* *SSE*

- SST (Total Sum of Squares):
 - How much error is there in predicting Y without the knowledge of X
- SSE (Sum of Squares Error):
 - How much error is there in predicting Y with the knowledge of X
- SSR (Sum of Squares Regression):
 - Amount of variation explained by the model
- Mathematically, $SST = SSR + SSE$

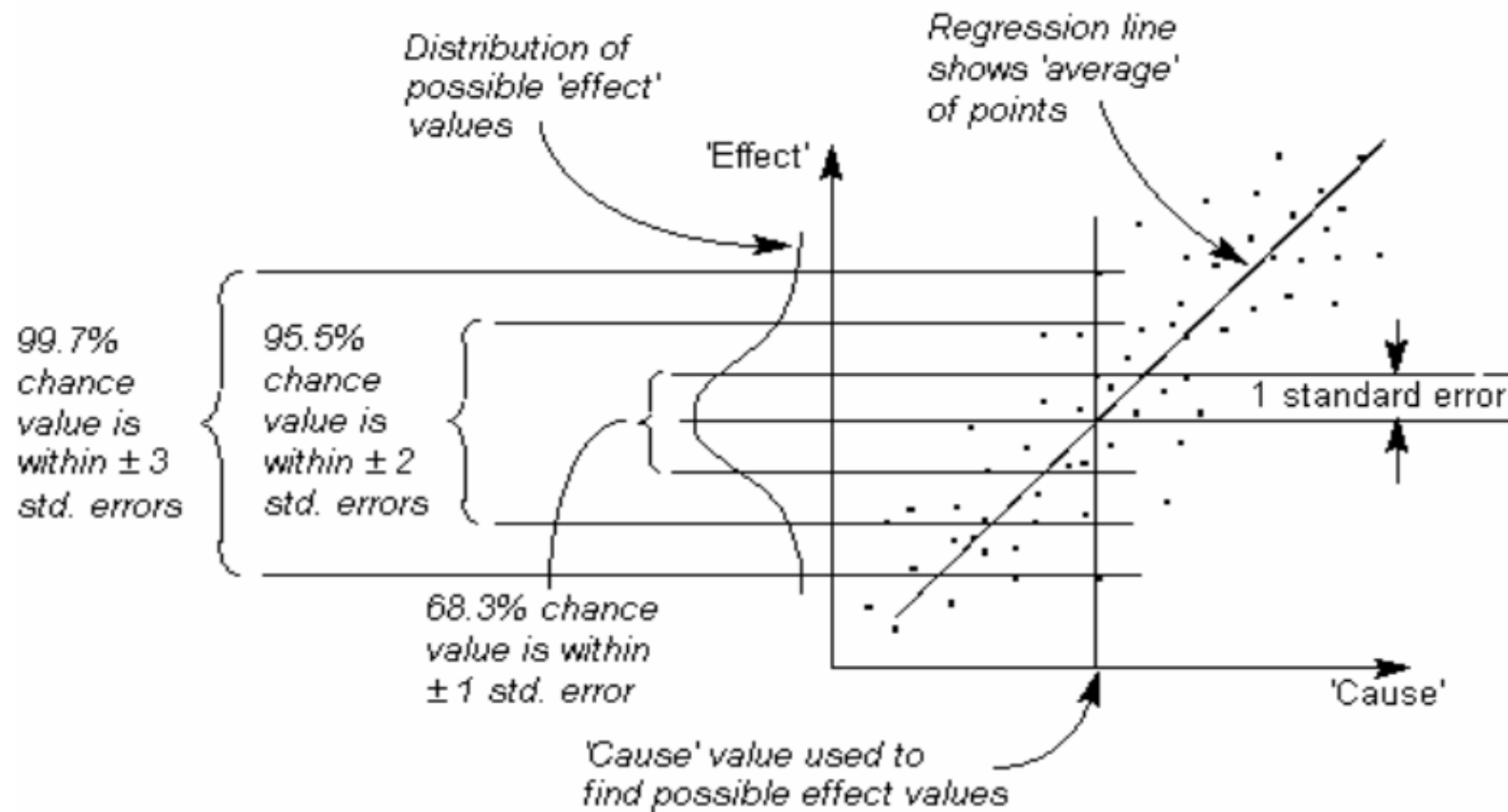
R-square

- What is explained by model over what is total variation

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Standard Error of Estimate

- Standard error is the estimate of the standard deviation of the regression errors
- Standard error of estimate, Se , measures the variability or scatter of the observed values around the regression line



Interpretation of SE Estimate

- Smaller SE of Estimate indicates better fit
- Larger SE of Estimate, the greater the scattering of points around the regression line
- Standard error of estimate for regression coefficient measures the amount of sampling error in a regression coefficient.

Standard Error of Estimate

$$S_e = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}}$$

$$S(\beta_0) = \frac{S_e \times \sqrt{\sum x^2}}{\sqrt{nSS_x}}$$

$$S(\beta_1) = \frac{S_e}{\sqrt{SS_x}}$$

$$SS_x = \sum_i (X_i - \bar{X})^2$$

T-test

- Beta co-efficient is a function of Y_i , since Y_i follows normal distribution, β_1 also follows normal distribution
- Y_i follows normal distribution since we assume that the error term follows normal distribution

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$$

T-test

- Check whether the real slope is zero or not,
- In simple linear regression, F-test and t-test check the same hypothesis ($H_0: \beta_1 = 0$)
- Use t-test instead of z-test since standard error is estimated from the sample

T-test Hypothesis

- Two-tailed test

Null hypothesis:

$$H_0: \beta_1 = 0$$

Alternative hypothesis:

$$H_1: \beta_1 \neq 0$$

Test Statistic

- Errors follow normal distribution, thus test statistic follows t-distribution with $n-2$ degrees of freedom
- Test statistic:

$$t_{(n-2)} = \frac{\text{Estimate value of parameter} - \text{hypothesis parameter}}{\text{Estimated standard error of estimate}} = \frac{\hat{\beta}_1 - \beta_1}{S_e(\hat{\beta}_1)}$$

$$t_{(n-2)} = \frac{\hat{\beta}_1 \sqrt{SS_x}}{S_e}, \text{ where } \beta = 0$$

Hypothesis testing decision rule

- If significance $\alpha = 0.05$, then we reject null hypothesis.
- $(1-\alpha)100\%$ is the confidence that we have that the null hypothesis may not be true.

P-value

- if less than 0.05, then we reject null hypothesis, accept alternative hypothesis -
> 95% confidence that null hypothesis may not be true

Multiple Linear Regression

- Several independent variables may influence the change in response variable we are trying to study
- Relationship between 1 dependent & 2 or more independent variables is a linear function

The diagram illustrates the Multiple Linear Regression equation:
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$
 Each term in the equation is labeled with a blue arrow pointing to it:

- Population Y-intercept** points to β_0 .
- Population slopes** points to the β coefficients ($\beta_1, \beta_2, \dots, \beta_k$).
- Random error** points to ε_i .
- Dependent (response) variable** points to Y_i .
- Independent (explanatory) variables** points to the X variables ($X_{1i}, X_{2i}, \dots, X_{ki}$).

Multiple Regression Modeling Steps

- 1. Start with a hypothesis or belief
- 2. Estimate unknown model parameters (Beta coefficients)
- 3. Probability distribution of random error term -> assumed to be a normal distribution
- 4. Check the assumptions of regression (normality, heteroscedasticity and multi-collinearity)
- 5. Evaluate Model
- 6. Use Model for Prediction and Estimation

Prediction Model

- Prediction equation obtained from sample data

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}$$

Model Diagnostics

- Test for overall model fitness (R-square and adjusted R-square)
- Test for overall model statistical significance (F test test)
- Test for portions of the model (Partial F-test)
- Test for statistical significance of individual explanatory variables (t test)
- Test for Normality and Homoscedasticity of residuals
- Test for Multi-collinearity and Auto Correlation

Co-efficient of determination in Multiple Regression

- Coefficient of determination increases as the number of explanatory variables increases.
- In SSR/SST , the numerator, SSR , increases as the number of explanatory variables increases, whereas the denominator, SST , remains constant.
- Increase in R^2 can be deceptive, since more number of explanatory variables may over-fit the data.

Adjusted R-square

- Inclusion of additional explanatory variable will increase R^2 value.
- By introducing an additional explanatory variable, we increase the numerator of the expression for R^2 while the denominator remains the same.
- To correct this defect, we adjust the R^2 by taking into account the degrees of freedom.

Adjusted R-Square

$$R_A^2 = 1 - \frac{SSE/(n - k - 1)}{SST/(n - 1)}$$

R_A^2 = Adjusted R - Square

n = number of observations

k = number of explanatory variables

Test for overall significance of model – F Test

- Test for overall significance of multiple regression model.
- Checks if there is a statistically significant relationship between Y and any of the explanatory variables

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

H_A : Not all β values are zero

F statistic

- Mean square regression over mean square error
- Relationship between F and R^2

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

Testing for Significance of Individual Parameters

- T-test: by rejecting null hypothesis, there is a statistically significant relationship between the response variable Y and explanatory variable Xi.

$$H_0 : \beta_i = 0$$

$$H_A : \beta_i \neq 0$$

$$t = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

Dummy Variable (1)

- Categorical (qualitative) variables in Regression as explanatory variable
- Qualitative variables (categorical variables) in regression are replaced with dummy variables (or indicator variables) in regression model
- A categorical variable with n levels are replaced with $(n-1)$ dummy variables. The category for which no dummy variable assigned is known as “Base Category”

Dummy variable (2)

- When there are more than one qualitative variable, it is advisable to use $(n-1)$ dummy variables for both qualitative variables along with the intercept.
- Use of n dummy variables along with intercept will result in multi- collinearity, known as dummy variable trap.

Dummy variables in Regression

- The intercept, β_0 the mean value of the base category.
- The coefficients attached to dummy variables are called differential intercept coefficients -> measure deviation from the base category for that specific dummy variable


Derived Variables

- Derived variables: explain the variation in the response variable
- Example: Credit rating problem, approve loan. Bank may use factors such as loan amount and value of the property, etc.
 - Bank use ratio $\text{Loan} / \text{Value}$

Interaction Variables

- A regression model of type:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$



**Interaction
variable**

- Usui

Interaction Variable Example

- Predict Gender Discrimination
- Consider a regression model with salary as response variable Y:

$$Y = \beta_0 + \beta_1 \text{ Gender} + \beta_2 \text{ Work Experience} + \beta_3 \text{ Gender x Work Experience}$$

Let Gender = 1 implies Female:

Then Y for Female is:

$$Y = \beta_0 + \beta_1 + (\beta_2 + \beta_3) \text{ Work Experience}$$

Y for Male is:

$$Y = \beta_0 + \beta_2 \times \text{Work Experience}$$

Multicollinearity

- High correlation between explanatory variables is called multi-collinearity.
- Multi-collinearity leads to unstable coefficients.
- Always exists; matter of degree.

Things to check of Multi-collinearity

- High R^2 but few significant t ratios.
- F-test rejects the null hypothesis, but none of the individual t-tests are rejected.
- Correlations between pairs of X variables are more than with Y variables.

Effects of Multi-collinearity

- The variances of regression coefficient estimators are inflated.
- Magnitudes of regression coefficient estimates may be different
- Adding and removing variables produce large changes in the coefficient estimates.
- Regression coefficient may have opposite sign.

Identify Multi-collinearity Variance Inflation Factor (VIF)

- The variance inflation factor (VIF) is a relative measure of the increase in the variance in standard error of beta coefficient because of collinearity.
- A VIF greater than 10 indicates that collinearity is very high. A VIF value of more than 4 is not acceptable.

Variance inflation factor

Variance inflation factor associated with introducing a new variable X_j is given by:

$$VIF(X_j) = \frac{1}{1 - R_j^2}$$

R_j^2 is the coefficient of determination for the regression of X_j as dependent variable

The standard error of the corresponding Beta is inflated by \sqrt{VIF}

VIF method

- Take particular X as response variable and all other explanatory variables as explanatory variables.
- Run a regression between one of those explanatory variables with remaining explanatory variables.
- Standard error of estimate is inflated by a quantity which is square root of VIF

Regression Model Building

- In *Forward selection* method, the entry variable is the one with smallest p-value based on F-test
- In *Backward elimination* method, all variables are entered into the equation and then sequentially removed starting with the most insignificant variable. At each step, the largest probability of F is removed
- In *Step-wise Regression*, the entry variable is the one with smallest p-value based on F-test. At each step, the independent variable not in the equation that has the smallest probability of F is entered.

Residual Plot

- Residual plot is a plot of error (or standardized error) against one of the following variables:
 - The dependent variable Y.
 - The independent variable X.
 - The standardized independent or dependent variable.

Residual Analysis

- Analysis of residuals reveal whether the assumption of normally distributed errors hold.
- Residual plots are used to check if there is heteroscedasticity problem (non constant variance for the error term).
- Residual analysis could also indicate if there are any missing variables.
- Residual plot can also reveal if the actual relationship is non-linear.

Normality of error terms

- Probability plot is a graphical technique for checking whether or not a data set follows a given distribution.
- The data is plotted against a theoretical distribution in such a way that the points should form a straight line.
- In Regression, we create a probability plot of error against normal distribution.
- If residual do not follow normal distribution, t-test and F-test are not valid

Check for heteroscedasticity

- A graph of the residuals versus independent variable Y or dependent variable X will reveal whether the variance of the errors are constant.
- If the width of the scatter plot of the residuals either increases or decreases as X (or Y) increases, then the assumption of constant variance is not met.

Check for non-linearity

- If the residual plot exhibits a curve when plotted, then the actual relationship is non-linear.

Prediction Error

- Irreducible Error: cannot be reduced
- Reducible Error: can be reduced
 - Bias Error
 - Variance Error

Prediction Error

$$y = f(x) + \epsilon$$

$$\hat{y} = \hat{f}(x)$$

↓
error term average to zero

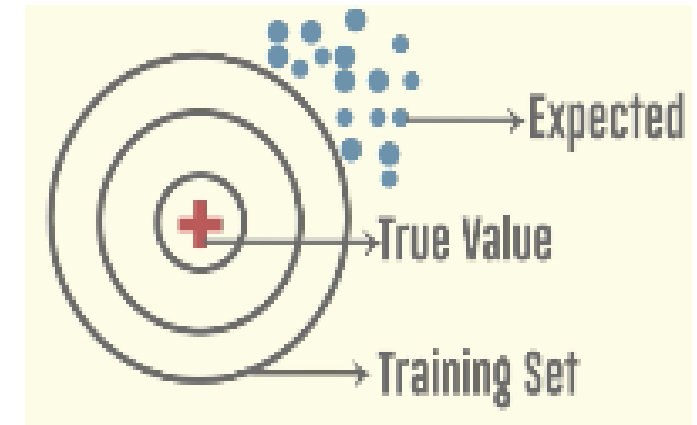
$$\text{Expected Squared Prediction Error } (Err(x)) = E[(Y - \hat{f}(x))^2]$$

Prediction Error

$$\text{Err}(x) = \underbrace{\left(\overset{\text{Predicted}}{E[\hat{f}(x)]} - \overset{\text{True}}{f(x)} \right)^2}_{\text{Bias}^2} + \underbrace{E\left[\left(\overset{\text{Predicted}}{\hat{f}(x)} - \overset{\text{Average Predicted}}{E[\hat{f}(x)]} \right)^2 \right]}_{\text{Variance}} + \underbrace{\sigma_e^2}_{\text{Irreducible Error}}$$

Error due to Bias

- Because of simplifying assumptions
- Less flexible
- Lower predictive performance
- Example – Linear Algorithms



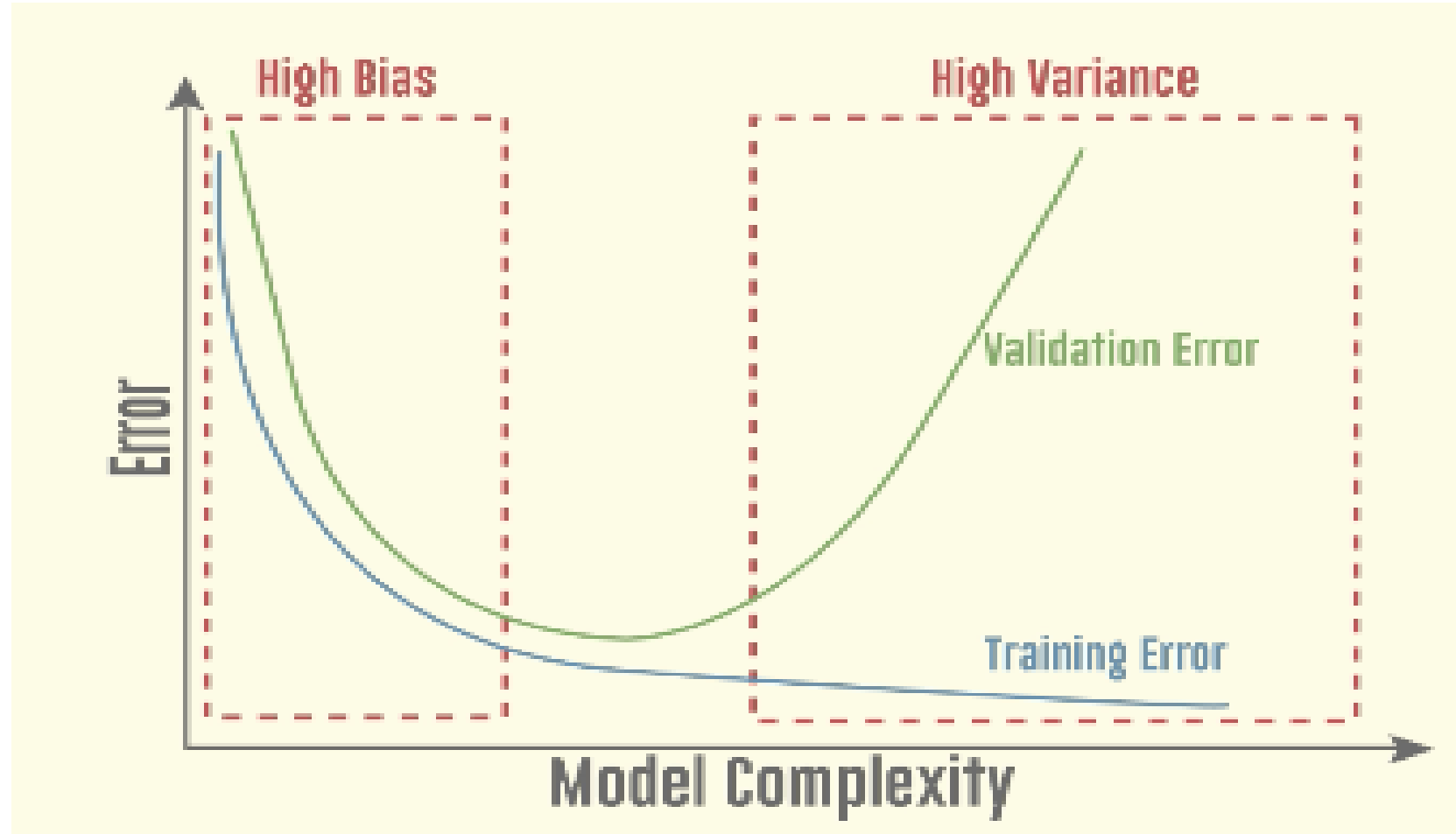
Error due to Variance

- Because of complex algorithms
- Lot of Flexibility
- Example: Decision Trees, Support Vector Machine (SVM)



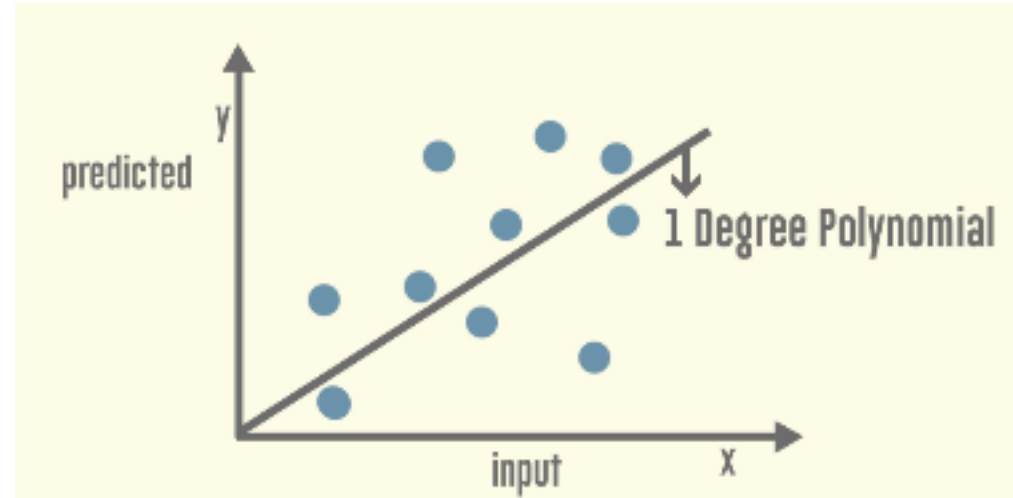
How can you know whether model has High Bias or High Variance ?

- High Training Set Error and High Validation Set Error ->
 - High Bias
- Low Training Set Error and High Validation Set Error ->
 - High Variance



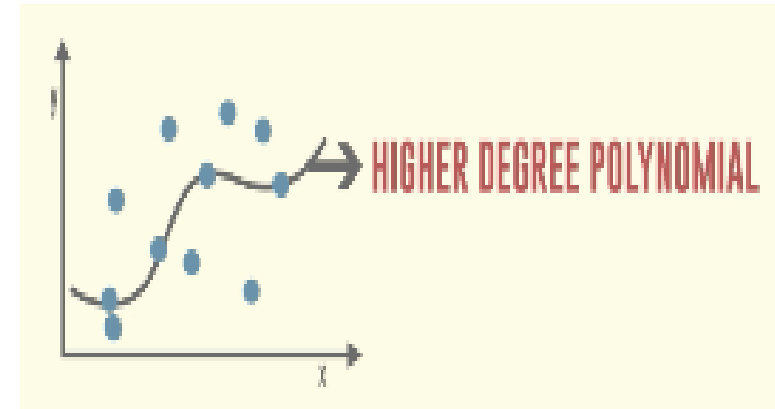
Underfitting

- High Bias and Low Variance
- Simple Model
- Does not capture the underlying trend of the data



Underfitting: How to fix a high bias and underfitting problem

- Train a more complex model
- Obtain more features



Overfitting

- High Variance and Low Bias
- Too complex
- Fits the data too well
- Learns the noise in the training data which impacts the performance on test data

Overfitting: How to fix high variance and overfitting problem

- Decrease the number of features
- Increasing the number of training examples

Python Application

- Import
- `% matplotlib inline` -> this allows plots to appear directly in the notebook

Case 1: Advertising Data

- `'http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv'`
- #read data into a DataFrame
 - `df = pd.read_csv('dataset directory/website link', index_col = 0)`
 - `index_col=0` we're explicitly stating to treat the first column as the index

Exploratory

- Descriptive stats
- Scatter plot

Build Model using statsmodels

- # this is the standard import if you're using "formula notation" (similar to R)
 - `import statsmodels.formula.api as smf`
- Fitted model
 - `lm = smf.ols(formula= 'Y variable ~ X variable', data=df).fit()`

- Print coefficient: `lm.params`
- Make predictions: `lm.predict`
- Print confidence interval: `lm.conf_int()`
- Print p-values: `lm.pvalues`
- Print R-squared: `lm.rsquared`
- Print summary of fitted model: `lm.summary()`

Build model using scikit-learn

- `from sklearn.linear_model import LinearRegression`
- `lm = LinearRegression()`
- Model fit: `lm.fit`
- Print intercept: `lm.intercept_`
- Print coefficients: `lm.coef_`
- Calculate r-squared: `lm.score`

Case 2: Housing Boston

- Train Test Split
- `X_train, X_test, Y_train, Y_test = sklearn.cross_validation.train_test_split(X, Y, test_size=, random_state =)`

- Using statsmodel: `import statsmodels.api as sm`
- Model fit -> `sm.OLS(Y, X).fit()`

- Test for normality -> Jarque-Bera normality test
 - `sm.stats.stattools.jarque_bera(residual data)`

Test to check if the observed skewness and kurtosis matching a normal distribution

- Normal probability plot
 - `sm.qqplot(residual data)`

Graphical technique based on comparison between observed distribution and the theoretical distribution under normal assumption

Null hypothesis (normal dist.) is rejected if the points are not aligned on a straight line

Detection of outliers and influential points

- Detection of outliers and influential points
 - Create object for the analysis of influential point: `get_influence()`
- Leverage: `hat_matrix_diag`
- Internally studentized residual: `resid_studentized_internal`

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_\varepsilon \sqrt{1 - h_i}}$$

- Externally studentized residual: resid_studentized_external

$$t_i^* = t_i \sqrt{\frac{n-p-2}{n-p-1-t_i^2}}$$

- Automatic detection: Correlation value
 - Define threshold values from which a point becomes suspect
- Leverage: threshold value
 - Observation is suspicious if leverage > threshold value

$$s_h = 2 \times \frac{p+1}{n}$$

- Externally studentized residual:

- Probability distribution of externally studentized residual is a student distribution with $(n-p-2)$ degrees of freedom.
- Threshold value at 5% level

$$s_i = t_{1-0.05/2}(n-p-2)$$

Where $t_{1-0.05/2}$ is the quantile of the t distribution for a probability 0.975.

An observation is suspicious if

$$|t_i^*| > s_i$$

- Other criterias:

- DFFITS

$$|DFFITS_i| > 2\sqrt{\frac{p+1}{n}} \text{ for DFFITS ;}$$

- Cook's distance

$$D_i > \frac{4}{n-p-1} \text{ for Cook's distance.}$$

Multicollinearity problem

- Disturbs statistical inference, inflates the estimated SE of coefficients.
- Build correlation matrix -> use `corrcoef()` from the `scipy` library
- Variance inflation factor (VIF) -> evaluate relationship of one predictor with all other explanatory variables