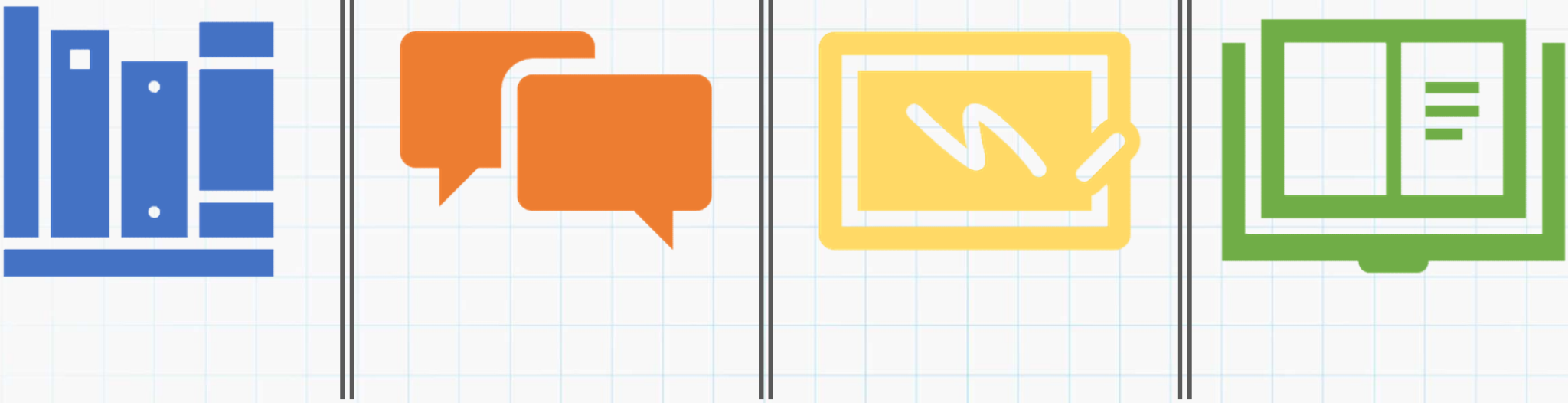
# COMPUTATIONAL STATISTICS

AKASH NAKASHE



# MULTIPLE LINEAR REGRESSION

# Regression Predictive Modelling

Regression predictive modelling is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y).

A continuous output variable is a real-value, such as an integer or floating point value. These are often quantities, such as amounts and sizes.

For example, a house may be predicted to sell for a specific dollar value, perhaps in the range of $100,000 to $200,000.

- A regression problem requires the prediction of a quantity. A regression can have real valued or discrete input variables.
- A problem with multiple input variables is often called a multivariate regression problem.
- A regression problem where input variables are ordered by time is called a time series forecasting problem.
- Because a regression predictive model predicts a quantity, the skill of the model must be reported as an error in those predictions.

# Classification Predictive Modelling

Classification predictive modelling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation.

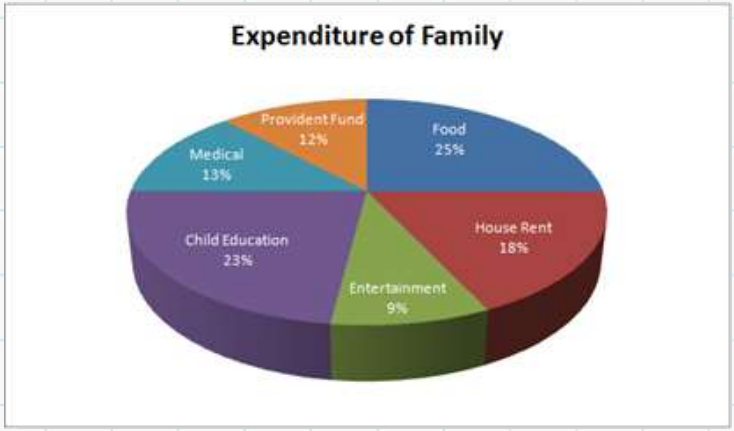
For example, an email of text can be classified as belonging to one of two classes: “spam“ and “*not spam*“.

- A classification problem requires that examples be classified into one of two or more classes.
- A classification can have real-valued or discrete input variables.
- A problem with two classes is often called a two-class or binary classification problem.
- A problem with more than two classes is often called a multi-class classification problem.
- A problem where an example is assigned multiple classes is called a multi-label classification problem.

## Example 1

I have data of my monthly spending, monthly income and the number of trips per month for the last three years. Now I need to answer the following questions:

- What will be my monthly spending for next year?
- Which factor(monthly income or number of trips per month) is more important in deciding my monthly spending?
- How monthly income and trips per month are correlated with monthly spending?

## Example 2

In the credit card industry, a financial company may be interested in minimizing the risk portfolio and wants to understand the top five factors that cause a customer to default. Based on the results the company could implement specific EMI options so as to minimize default among risky customers.

## Example 3

A company wanted to be able to estimate or predict how much fuel they needed to transport building materials to their oil wells so that they could line them with concrete. The data provided was:

- Number of wells
- Depth of wells
- Distance to wells
- Weight of materials
- Ton kilometers
- Fuel costs

## Example 4

An ecommerce company wants to measure

- the impact of product price,
- product promotions, and
- holiday seasonality  on product sales.


- A product sales manager can discover which predictors included in the analysis will have significant impact on *product sales*.
- For the predictors with the most impact, the team can make important strategic decisions to meet product sales targets.
- For instance, if promotions and holiday seasons are significant factors, these factors should be given more focus when devising a marketing strategy.

## Example 5

Predicting Gross Movie Revenue

- Success or failure of a movie can depend on many factors: star-power, release date, Critics review, budget, rating, plot and the highly unpredictable human reactions.
- Predicted revenues can be used for planning both the production and distribution stages.

## Example 6

A TV industry analyst wants to build a statistical model for predicting the number of subscribers that a cable station can expect

*Y = Number of cable subscribers*

$X_1$ = Advertising rate which the station charges local advertisers for one minute of prime time space,

$X_2$ = Kilowatt power of the station's non-cable signal,

$X_3$ = Number of families living in the station's area of dominant influence,

$X_4$ = Number of competing stations

## Purpose

Multiple regression analysis has three main uses:

- You can look at the strength of the effect of the independent variables on the dependent variable.
- You can use it to ask how much the dependent variable will change if the independent variables are changed.
- You can also use it to predict trends and future values.

## Correlation

**Definition:**

- Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.
- Two variables are said to be correlated if change in one variable affects the change in other variable, and the relation between them is known as correlation.
- If two variables vary together they are said to be correlated.
- For example, the fuel consumed by a car is correlated to the number of miles travelled. The exact relationship is not easily found because so many other factors are involved such as speed of travel, condition of the engine, tyre pressures and so on. What is important is that as the mileage increases so does the fuel used – there is a correlation.

## Correlation

**Uses:**

- Prediction: If there is a relationship between two variables, we can make predictions about one from another.
- Validity: Concurrent validity (correlation between a new measure and an established measure).
- Reliability: Test-retest reliability (are measures consistent).
- Inter-rater reliability (are observers consistent).
- Theory verification
- Predictive validity.

## Correlation

### Types of correlations:

- Positive Correlation: Two variables are said to be positively correlated if they deviates in the same direction. e.g. height & weight, income & expenditure

- Negative Correlation: Two variables are said to be negatively correlated if they deviates in the opposite directions. e.g. volume and pressure of a perfect gas, price and demand

- No Correlation: Two variables are said to be uncorrelated or statistically independent if there is no relation between them.

## Simple Linear Regression

- Regression analysis is a statistical technique for investigating and modeling the relationship between variables.
- Regression analysis is the process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable or other variables.
- The most elementary regression model is called simple regression or bivariate regression involving two variables in which one variable is predicted by another variable.
- In simple regression, the variable to be predicted is called the dependent variable and is designated as **y**. The predictor is called the independent variable, or explanatory variable, and is designated as **x**. In simple regression analysis, only a straight-line relationship between two variables is examined.
- In math courses, the slope-intercept form of the equation of a line often takes the form

$$y=mx+b$$

Where

m = slope of the line

b = y intercept of the line

## Simple Linear Regression

In statistics, the slope-intercept form of the equation of the regression line through the population points is

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

Where

$\hat{y}$ = the predicted value of y

$\beta_0$ = the population y intercept

$\beta_1$ = the population slope

It is known as Deterministic models or mathematical models that produce an “exact” output for a given input.

## Multiple Linear Regression

- By extending the simple regression model to a multiple regression model with two independent variables, it is possible to determine the multiple regression equation for any number of unknowns.
- Multiple regression analysis is similar in principle to simple regression analysis.
- However, it is more complex conceptually and computationally.
- Extending this notion to multiple regression gives the general equation for the probabilistic multiple regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots\ldots\ldots \beta_k x_k + \in$$

## Multiple Linear Regression

Where

- k = the number of independent variables
- $\beta_k$ = the partial regression coefficient for independent variable k
- $\beta_3$ = the partial regression coefficient for independent variable 3
- $\beta_2$ = the partial regression coefficient for independent variable 2
- $\beta_1$ = the partial regression coefficient for independent variable 1
- $\beta_0$ = the regression constant
- y = the value of the dependent variable

- In virtually all research, these values are estimated by using sample information.
- Shown here is the form of the equation for estimating y with sample information.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots\ldots\ldots b_k x_k + \in$$

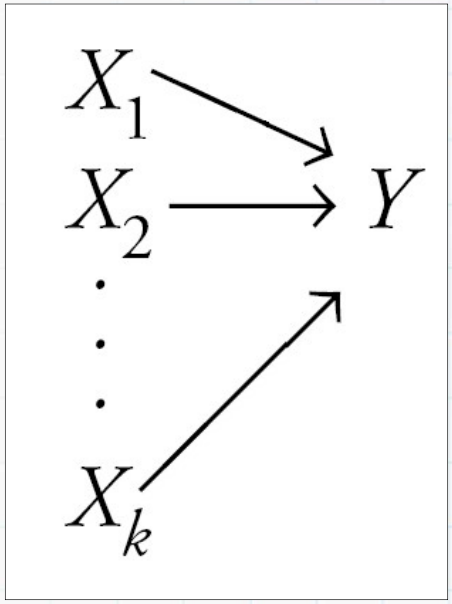
## Multiple Linear Regression

Where

- k = the number of independent variables
- $b_k$ = the estimate of the partial regression coefficient for independent variable k
- $b_3$ = the estimate of the partial regression coefficient for independent variable 3
- $b_2$ = the estimate of the partial regression coefficient for independent variable 2
- $b_1$ = the estimate of the partial regression coefficient for independent variable 1
- $b_0$ = the estimate of the regression constant
- y = the predicted value of y

## Multiple Linear Regression

Multiple regression simultaneously considers the influence of multiple explanatory variables on a response variable Y




The intent is to look at the independent effect of each variable while “adjusting out” the influence of potential confounders

## Case 1

- Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.
- The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.
- It is not possible for our client to directly increase sales of the product. On the other hand, they can control the advertising expenditure in each of the three media.
- Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales.
- In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

## Case 1


- In this setting, the advertising budgets are input variables while sales input is an output variable. So **X1** might be the TV budget, **X2** the radio budget, and **X3** the newspaper budget. The inputs go by different names, such as predictors, independent variables, features, predictor independent variable feature or sometimes just variables. The output variable—in this case, sales—is variable often called the response or dependent variable, and is typically denoted response dependent variable using the symbol **Y**.

## Case 1

Here are a few important questions that we might seek to address:

1. Is there a relationship between advertising budget and sales?
    - Our first goal should be to determine whether the data provide evidence of an association between advertising expenditure and sales.
    - If the evidence is weak, then one might argue that no money should be spent on advertising!

2. How strong is the relationship between advertising budget and sales?
    - Assuming that there is a relationship between advertising and sales, we would like to know the strength of this relationship.
    - In other words, given a certain advertising budget, can we predict sales with a high level of accuracy? This would be a strong relationship.
    - Or is a prediction of sales based on advertising expenditure only slightly better than a random guess? This would be a weak relationship.

## Case 1

Here are a few important questions that we might seek to address:

3. Which media contribute to sales?
    - Do all three media—TV, radio, and newspaper—contribute to sales, or do just one or two of the media contribute?
    - To answer this question, we must find a way to separate out the individual effects of each medium when we have spent money on all three media.

4. How accurately can we estimate the effect of each medium on sales?
    - For every dollar spent on advertising in a particular medium, by what amount will sales increase?
    - How accurately can we predict this amount of increase?

## Case 1

Here are a few important questions that we might seek to address:

5. How accurately can we predict future sales?
    - For any given level of television, radio, or newspaper advertising, what is our prediction for sales, and what is the accuracy of this prediction?

6. Is the relationship linear?
    - If there is approximately a straight-line relationship between advertising expenditure in the various media and sales, then linear regression is an appropriate tool.

    - If not, then it may still be possible to transform the predictor or the response so that linear regression can be used.

In our case the MLR equation will become:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

## Assumptions

### Assumption 1: Linearity

There must be a linear relationship between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.

### Assumption 2: Auto Correlation

Linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the ***residuals*** are not independent from each other. For instance, this typically occurs in stock prices, where the price is not independent from the previous price.

### Assumption 3: Homoscedasticity

The error terms must have constant variance. This phenomenon is known as ***homoscedasticity***. The presence of non-constant variance is referred to heteroscedasticity. The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line).

## Assumptions

### Assumption 4: Outliers/influential cases:

As with simple linear regression, it is important to look out for cases which may have a disproportionate influence over your regression model.

### Assumption 5: Multicollinearity:

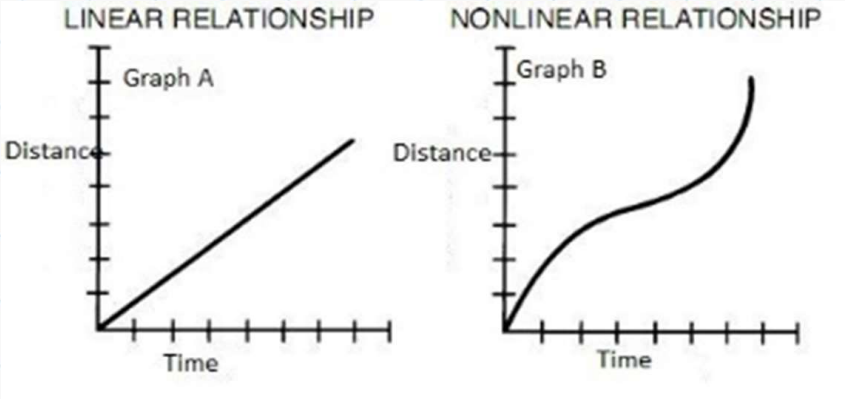
***Multicollinearity*** exists when two or more of the explanatory variables are highly correlated. This is a problem as it can be hard to disentangle which of them best explains any shared variance with the outcome. It also suggests that the two variables may actually represent the same underlying factor. We can check Multicollinearity using VIF(variance inflation factor).

### Assumption 6: Independence of Error

Residuals should be normally distributed. This can be checked by visualizing Q-Q Normal plot.

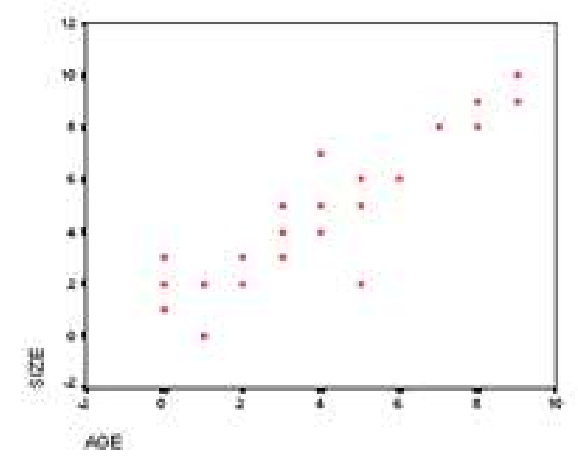
## Assumption : Linearity

- First, multiple linear regression requires the relationship between the independent and dependent variables to be linear.
- The linearity assumption can best be tested with *scatterplots*.




- If data is given in pairs then the scatter diagram of the data is just the points plotted on the xy-plane.
- The scatter plot is used to visually identify relationships between the first and the second entries of paired data.

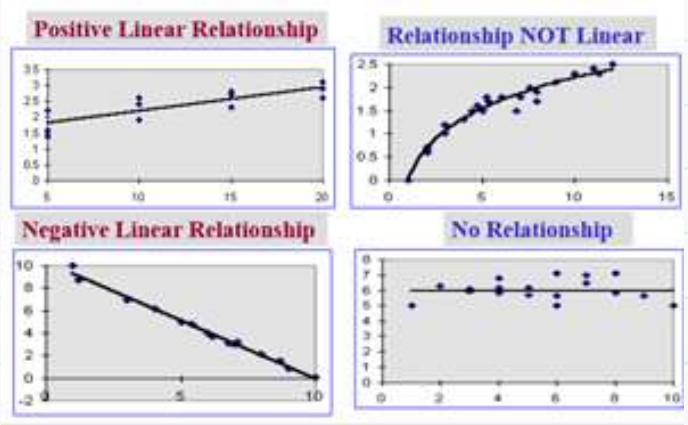
## Assumption : Linearity




- The scatter plot above represents the age vs. size of a plant. It is clear from the scatter plot that as the plant ages, its size tends to increase. If it seems to be the case that the points follow a linear pattern well, then we say that there is a high linear correlation, while if it seems that the data do not follow a linear pattern, we say that there is no linear correlation. If the data somewhat follow a linear path, then we say that there is a moderate linear correlation.

- Given a scatter plot, we can draw the line that best fits the data

## Assumption 1: Linearity



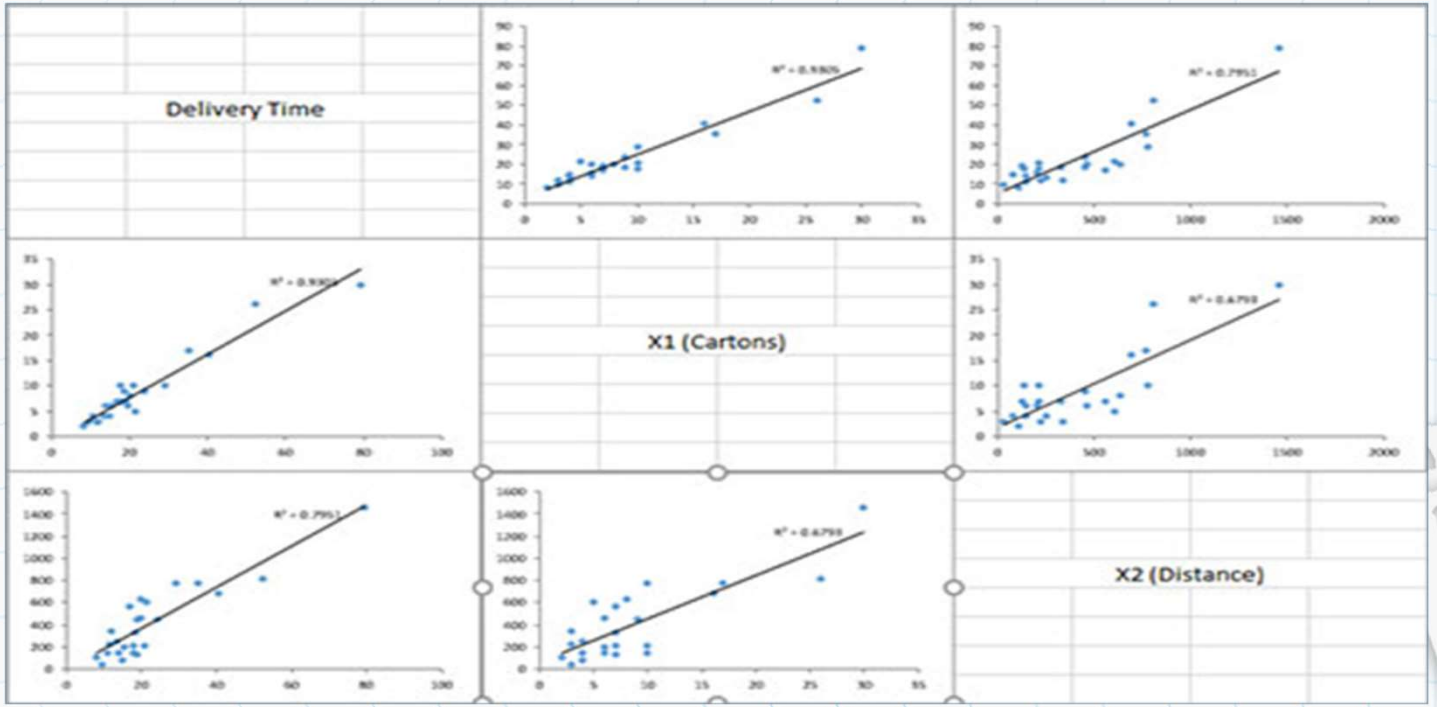

- A scatter plot matrix is a grid (or matrix) of scatter plots used to visualize bivariate relationships between combinations of variables.

- Each scatter plot in the matrix visualizes the relationship between a pair of variables, allowing many relationships to be explored in one chart.

## Assumption 1: Linearity

Example: A soft drink bottler is trying to predict delivery times for a driver. He has collected data on the delivery time, the number of cartons delivered and the distance the driver walked. He wants to see the relationship between these three variables. We will use the scatter plot matrix to do this.

| Delivery Time | X1 (Cartons) | X2 (Distance) |
|---|---|---|
| 16.68 | 7 | 560 |
| 11.5 | 3 | 220 |
| 12.03 | 3 | 340 |
| 14.88 | 4 | 80 |
| 13.75 | 6 | 150 |
| 18.11 | 7 | 330 |
| 8 | 2 | 110 |
| 17.83 | 7 | 210 |
| 79.24 | 30 | 1460 |
| 21.5 | 5 | 605 |
| 40.33 | 16 | 688 |
| 21 | 10 | 215 |
| 13.5 | 4 | 255 |
| 19.75 | 6 | 462 |
| 24 | 9 | 448 |
| 29 | 10 | 776 |
| 15.35 | 6 | 200 |
| 19 | 7 | 132 |
| 9.5 | 3 | 36 |
| 35.1 | 17 | 770 |
| 17.9 | 10 | 140 |
| 52.32 | 26 | 810 |
| 18.75 | 9 | 450 |
| 19.83 | 8 | 635 |
| 10.75 | 4 | 150 |



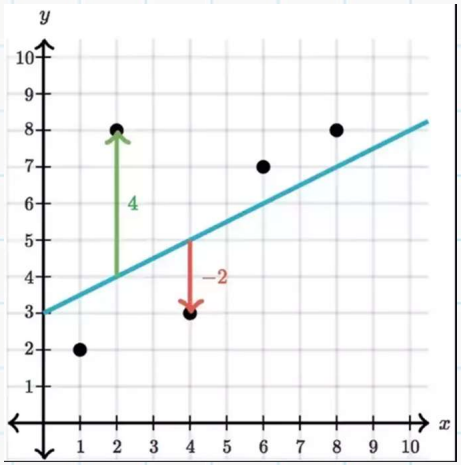
## Residuals

- The residual, or error, of the regression model is the difference between the y value and the predicted value.  Each data point has one residual.

Residual = Observed value - Predicted value

$$\text{Residual} = y - \hat{y}$$

- Because a linear regression model is not always appropriate for the data, you should assess the appropriateness of the model by defining residuals and examining residual plots.
- Both the sum and the mean of the residuals are equal to zero.



## Residuals

- We define the residual sum of squares (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

- The least squares approach chooses $\beta\hat{}0$ and $\beta\hat{}1$ to minimize the RSS.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

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

## Residuals

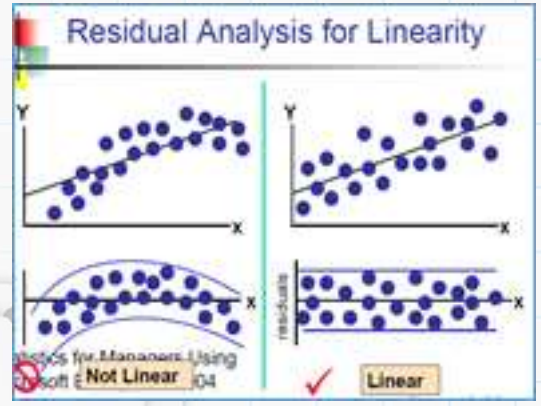
Using Residuals to Test the Assumptions of the Regression Model

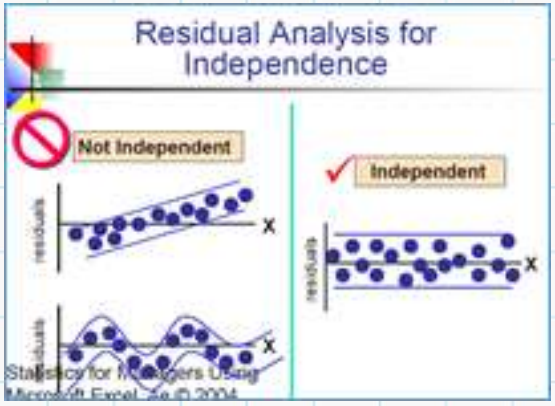
- One of the major uses of residual analysis is to test some of the assumptions underlying regression. The following are the assumptions of simple regression analysis.

    1. The model is linear.(Assumption 01)
    2. The error terms have constant variances.(Assumption 03)
    3. The error terms are independent.
    4. The error terms are normally distributed.(Assumption 06)

## Residuals Plot

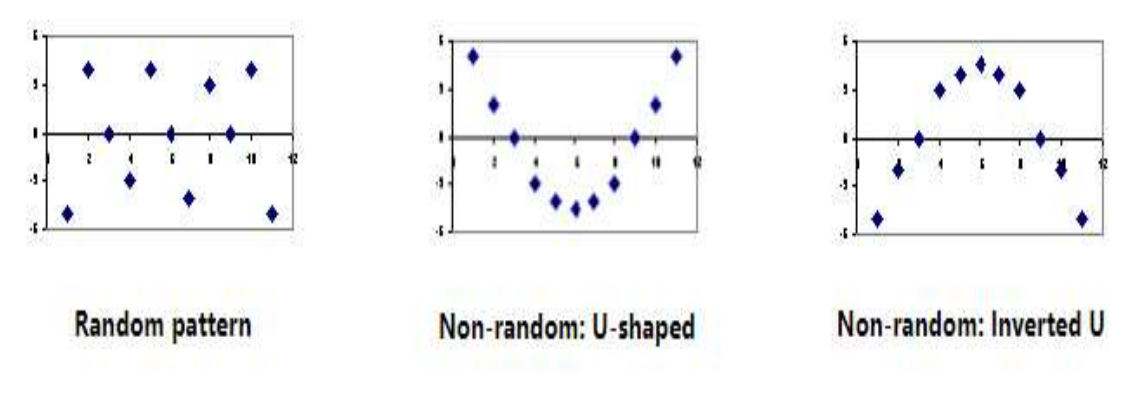
- A particular method for studying the behavior of residuals is the residual plot.
- The residual plot is a type of graph in which the residuals for a particular regression model are plotted along with their associated value of x as an ordered pair(x, Residual).
- A residual plot is a graph in which residuals are on the vertical axis and the independent variable is on the horizontal axis.
- If the dots are randomly dispersed around the horizontal axis then a linear regression model is appropriate for the data; otherwise, choose a non-linear model.



## Residuals Plot

Following example shows few patterns in residual plots.




- In first case, dots are randomly dispersed. So linear regression model is preferred.
- In Second and third case, dots are non-randomly dispersed and suggests that a non-linear regression method is preferred

**Problem 1:**

Suppose a study is conducted using only Boeing 737s traveling 500 miles on comparable routes during the same season of the year. Can the number of passengers predict the cost of flying such routes? Suppose the data displayed in Table are the costs and associated number of passengers for twelve 500-mile commercial airline flights using Boeing 737s during the same season of the year. Use these data to develop a regression model to predict cost by number of passengers. Analyze the residuals linearity by using graphic diagnostics.

Airline Cost Data

| Number of Passengers | Cost ($1,000) |
|---|---|
| 61 | 4.280 |
| 63 | 4.080 |
| 67 | 4.420 |
| 69 | 4.170 |
| 70 | 4.480 |
| 74 | 4.300 |
| 76 | 4.820 |
| 81 | 4.700 |
| 86 | 5.110 |
| 91 | 5.130 |
| 95 | 5.640 |
| 97 | 5.560 |

**Problem 1 Solution:**

N.E. $\Sigma y = Na + b\Sigma x \longrightarrow (i)$

$\Sigma xy = a\Sigma x + \Sigma x^2 \longrightarrow (ii)$

Assumption 01: Linearity

The first step in simple regression analysis is to construct a scatter plot. The scatter plot gives some idea of how well a regression line fits the data.

| Number of Passengers | Cost ($1,000) | | |
|---|---|---|---|
| $x$ | $y$ | $x^2$ | $xy$ |
| 61 | 4.280 | 3,721 | 261.080 |
| 63 | 4.080 | 3,969 | 257.040 |
| 67 | 4.420 | 4,489 | 296.140 |
| 69 | 4.170 | 4,761 | 287.730 |
| 70 | 4.480 | 4,900 | 313.600 |
| 74 | 4.300 | 5,476 | 318.200 |
| 76 | 4.820 | 5,776 | 366.320 |
| 81 | 4.700 | 6,561 | 380.700 |
| 86 | 5.110 | 7,396 | 439.460 |
| 91 | 5.130 | 8,281 | 466.830 |
| 95 | 5.640 | 9,025 | 535.800 |
| 97 | 5.560 | 9,409 | 539.320 |
| $\Sigma x = 930$ | $\Sigma y = 56.690$ | $\Sigma x^2 = 73,764$ | $\Sigma xy = 4462.220$ |

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 4462.22 - \frac{(930)(56.69)}{12} = 68.745$$

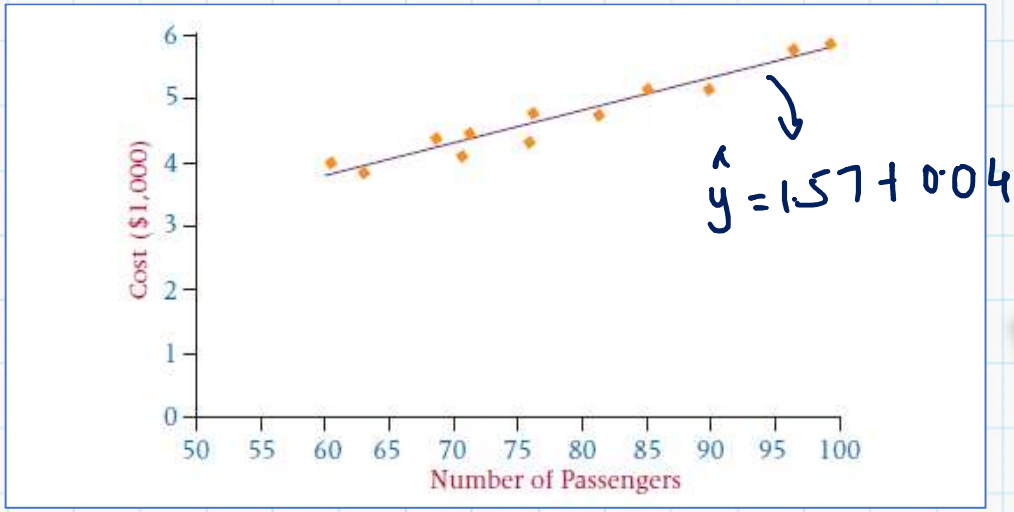
$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x^2)}{n} = 73,764 - \frac{(930)^2}{12} = 1689$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{68.745}{1689} = .0407$$

$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{n} = \frac{56.19}{12} - (.0407)\frac{930}{12} = 1.57$$

$$\hat{y} = 1.57 + .0407x$$

Superimposing the line representing the least squares equation for this problem on the scatter plot indicates how well the regression line fits the data points.




$\hat{y} = 1.57 + 0.0407x$

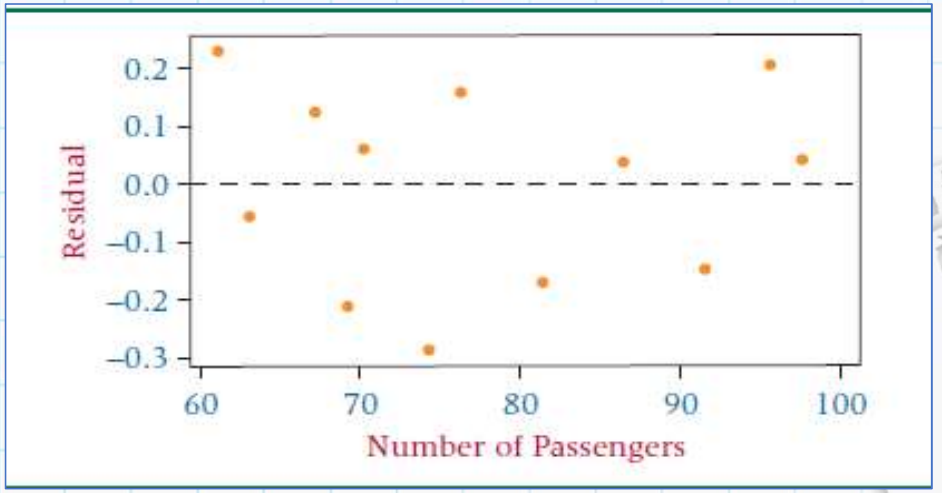
## Problem 1 Solution:

Assumption 02 Residual Analysis

$\hat{y} = 1.37 + 0.040x$

| Number of Passengers $x$ | Cost ($1,000) $y$ | Predicted Value $\hat{y}$ | Residual $y - \hat{y}$ |
|---|---|---|---|
| 61 | 4.280 | 4.053 | .227 |
| 63 | 4.080 | 4.134 | −.054 |
| 67 | 4.420 | 4.297 | .123 |
| 69 | 4.170 | 4.378 | −.208 |
| 70 | 4.480 | 4.419 | .061 |
| 74 | 4.300 | 4.582 | −.282 |
| 76 | 4.820 | 4.663 | .157 |
| 81 | 4.700 | 4.867 | −.167 |
| 86 | 5.110 | 5.070 | .040 |
| 91 | 5.130 | 5.274 | −.144 |
| 95 | 5.640 | 5.436 | .204 |
| 97 | 5.560 | 5.518 | .042 |
| | | | $\Sigma(y - \hat{y}) = -.001$ |

Note that the sum of the residuals is approximately zero. Except for rounding error, the sum of the residuals is *always zero.* Residuals are usually plotted against the $x$-axis, which reveals a view of the residuals as $x$ increases.



## Problem 2:

A specialist in hospital administration stated that the number of FTEs (full-time employees) in a hospital can be estimated by counting the number of beds in the hospital (a common measure of hospital size). A healthcare business researcher decided to develop a regression model in an attempt to predict the number of FTEs of a hospital by the number of beds. She surveyed 12 hospitals and obtained the following data. The data are presented in sequence, according to the number of beds:

| Number of Beds | FTEs | Number of Beds | FTEs |
|---|---|---|---|
| 23 | 69 | 50 | 138 |
| 29 | 95 | 54 | 178 |
| 29 | 102 | 64 | 156 |
| 35 | 118 | 66 | 184 |
| 42 | 126 | 76 | 176 |
| 46 | 125 | 78 | 225 |

Compute the residuals for Demonstration Problem in which a regression model was developed to predict the number of full-time equivalent workers (FTEs) by the number of beds in a hospital. Analyze the residuals by using graphic diagnostics.

# Problem 2 Solutions:

| Hospital | Number of Beds $x$ | FTEs $y$ | $x^2$ | $xy$ |
|---|---|---|---|---|
| 1 | 23 | 69 | 529 | 1,587 |
| 2 | 29 | 95 | 841 | 2,755 |
| 3 | 29 | 102 | 841 | 2,958 |
| 4 | 35 | 118 | 1,225 | 4,130 |
| 5 | 42 | 126 | 1,764 | 5,292 |
| 6 | 46 | 125 | 2,116 | 5,750 |
| 7 | 50 | 138 | 2,500 | 6,900 |
| 8 | 54 | 178 | 2,916 | 9,612 |
| 9 | 64 | 156 | 4,096 | 9,984 |
| 10 | 66 | 184 | 4,356 | 12,144 |
| 11 | 76 | 176 | 5,776 | 13,376 |
| 12 | 78 | 225 | 6,084 | 17,550 |
| | $\Sigma x = 592$ | $\Sigma y = 1{,}692$ | $\Sigma x^2 = 33{,}044$ | $\Sigma xy = 92{,}038$ |

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 92{,}038 - \frac{(592)(1692)}{12} = 8566$$

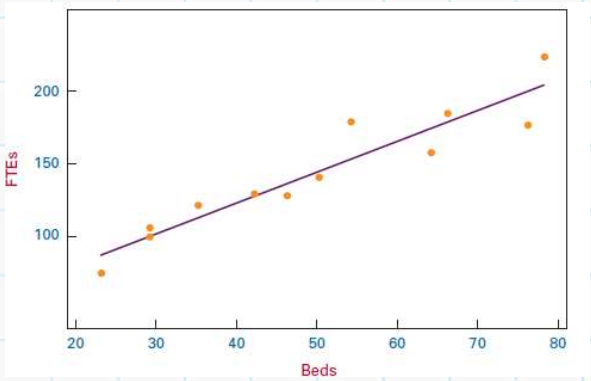
$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 33{,}044 - \frac{(592)^2}{12} = 3838.667$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{8566}{3838.667} = 2.232$$

$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{12} = \frac{1692}{12} - (2.232)\frac{592}{12} = 30.888$$

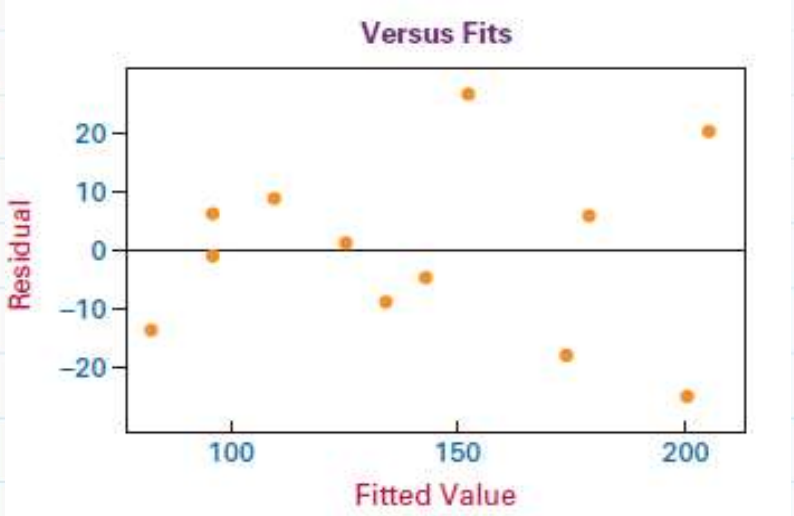
The least squares equation of the regression line is

$$\hat{y} = 30.888 + 2.232x$$



## Problem 2 Solutions:

| Hospital | Number of Beds $x$ | FTES $y$ | Predicted Value $\hat{y}$ | Residuals $y - \hat{y}$ |
|---|---|---|---|---|
| 1 | 23 | 69 | 82.22 | −13.22 |
| 2 | 29 | 95 | 95.62 | −.62 |
| 3 | 29 | 102 | 95.62 | 6.38 |
| 4 | 35 | 118 | 109.01 | 8.99 |
| 5 | 42 | 126 | 124.63 | 1.37 |
| 6 | 46 | 125 | 133.56 | −8.56 |
| 7 | 50 | 138 | 142.49 | −4.49 |
| 8 | 54 | 178 | 151.42 | 26.58 |
| 9 | 64 | 156 | 173.74 | −17.74 |
| 10 | 66 | 184 | 178.20 | 5.80 |
| 11 | 76 | 176 | 200.52 | −24.52 |
| 12 | 78 | 225 | 204.98 | 20.02 |
| | | | | $\Sigma(y - \hat{y}) = -.01$ |




Note that the regression model fits these particular data well for hospitals 2 and 5, as indicated by residuals of -.62 and 1.37 FTEs, respectively. For hospitals 1, 8, 9, 11, and 12, the residuals are relatively large, indicating that the regression model does not fit the data for these hospitals well.

## SSE and Standard Error of the Estimate

- Residuals represent errors of estimation for individual points.
- With large samples of data, residual computations become laborious.
- Even with computers, a researcher sometimes has difficulty working through pages of residuals in an effort to understand the error of the regression model.
- An alternative way of examining the error of the model is the standard error of the estimate, which provides a single measurement of the regression error.
- Because the sum of the residuals is zero, attempting to determine the total amount of error by summing the residuals is fruitless.

- The total of the residuals squared column is called the sum of squares of error (SSE).

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

- Computational formula for SSE:

$$SSE = \sum y^2 - b_0 \sum y - b_1 \sum xy$$

## SSE and Standard Error of the Estimate (Accessing the accuracy)

In theory, infinitely many lines can be fit to a sample of points. Line of best fit is for which the SSE is the smallest. For this reason, the regression process is called least squares regression. A value closer to 0 indicates that the model has a smaller random error component, and that the fit will be more useful for prediction.

### The standard error of the estimate

- A more useful measurement of error is the standard error of the estimate.
- The standard error of the estimate, denoted se, is a standard deviation of the error of the regression model and has more practical use than SSE.
- An assumption underlying regression analysis is that the error terms are approximately normally distributed with a mean of zero. With this information and by the empirical rule, approximately 68% of the residuals should be within 1se and 95% should be within 2se. This property makes the standard error of the estimate a useful tool in estimating how accurately a regression model is fitting the data.
- The standard error of the estimate is a standard deviation of error. The standard error of the regression provides the absolute measure of the typical distance that the data points fall from the regression line.

## SSE and Standard Error of the Estimate

The standard error of the estimate is computed by dividing SSE by the degrees of freedom of error for the model and taking the square root. The standard error of the estimate follows.

$$S_e = \sqrt{\frac{SSE}{n-k-1}}$$

where

n = number of observations

k = number of independent variables

SSE/(n-k-1)= mean squared errors or (MSE).

n – k – 1 =degrees of freedom

## SSE and Standard Error of the Estimate

**Problem 1: (Slide 37)**

Determining SSE and Standard Error of the Estimate for the Airline Cost Example 01

| Number of Passengers $x$ | Cost ($1,000) $y$ | Residual $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|
| 61 | 4.280 | .227 | .05153 |
| 63 | 4.080 | −.054 | .00292 |
| 67 | 4.420 | .123 | .01513 |
| 69 | 4.170 | −.208 | .04326 |
| 70 | 4.480 | .061 | .00372 |
| 74 | 4.300 | −.282 | .07952 |
| 76 | 4.820 | .157 | .02465 |
| 81 | 4.700 | −.167 | .02789 |
| 86 | 5.110 | .040 | .00160 |
| 91 | 5.130 | −.144 | .02074 |
| 95 | 5.640 | .204 | .04162 |
| 97 | 5.560 | .042 | .00176 |
| | | $\Sigma(y - \hat{y}) = -.001$ | $\Sigma(y - \hat{y})^2 = .31434$ |

## SSE and Standard Error of the Estimate

### Problem 1: Another Method

Determining SSE and Standard Error of the Estimate for the Airline Cost Example 01

$$b_1 = .0407016^*$$

$$\Sigma y = 56.69$$

$$\Sigma xy = 4462.22$$

$$SSE = \Sigma y^2 - b_0\Sigma y - b_1\Sigma xy$$

$$= 270.9251 - (1.5697928)(56.69) - (.0407016)(4462.22) = .31405$$

The standard error of the estimate for the airline cost example is

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{.31434}{10}} = .1773$$

## SSE and Standard Error of the Estimate

**Problem 2: (Slide No. 40)**

Determining SSE and Standard Error of the Estimate for the Hospital Example 02

| Hospital | Number of Beds $x$ | FTEs $y$ | $x^2$ | $xy$ |
|---|---|---|---|---|
| 1 | 23 | 69 | 529 | 1,587 |
| 2 | 29 | 95 | 841 | 2,755 |
| 3 | 29 | 102 | 841 | 2,958 |
| 4 | 35 | 118 | 1,225 | 4,130 |
| 5 | 42 | 126 | 1,764 | 5,292 |
| 6 | 46 | 125 | 2,116 | 5,750 |
| 7 | 50 | 138 | 2,500 | 6,900 |
| 8 | 54 | 178 | 2,916 | 9,612 |
| 9 | 64 | 156 | 4,096 | 9,984 |
| 10 | 66 | 184 | 4,356 | 12,144 |
| 11 | 76 | 176 | 5,776 | 13,376 |
| 12 | 78 | 225 | 6,084 | 17,550 |
| | $\Sigma x = 592$ | $\Sigma y = 1,692$ | $\Sigma x^2 = 33,044$ | $\Sigma xy = 92,038$ |

## SSE and Standard Error of the Estimate

**Problem 2: (Slide No. 40)**

Determining SSE and Standard Error of the Estimate for the Hospital Example 02

| Hospital | Number of Beds $x$ | FTES $y$ | Predicted Value $\hat{y}$ | Residuals $y - \hat{y}$ |
|---|---|---|---|---|
| 1 | 23 | 69 | 82.22 | −13.22 |
| 2 | 29 | 95 | 95.62 | −.62 |
| 3 | 29 | 102 | 95.62 | 6.38 |
| 4 | 35 | 118 | 109.01 | 8.99 |
| 5 | 42 | 126 | 124.63 | 1.37 |
| 6 | 46 | 125 | 133.56 | −8.56 |
| 7 | 50 | 138 | 142.49 | −4.49 |
| 8 | 54 | 178 | 151.42 | 26.58 |
| 9 | 64 | 156 | 173.74 | −17.74 |
| 10 | 66 | 184 | 178.20 | 5.80 |
| 11 | 76 | 176 | 200.52 | −24.52 |
| 12 | 78 | 225 | 204.98 | 20.02 |
| | | | | $\Sigma(y - \hat{y}) = -.01$ |

## SSE and Standard Error of the Estimate

**Problem 2: (Slide No. 40)**

Determining SSE and Standard Error of the Estimate for the Hospital Example 02

| Hospital | Number of Beds $x$ | FTES $y$ | Residuals $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|---|
| 1 | 23 | 69 | −13.22 | 174.77 |
| 2 | 29 | 95 | −.62 | −0.38 |
| 3 | 29 | 102 | 6.38 | 40.70 |
| 4 | 35 | 118 | 8.99 | 80.82 |
| 5 | 42 | 126 | 1.37 | 1.88 |
| 6 | 46 | 125 | −8.56 | 73.27 |
| 7 | 50 | 138 | −4.49 | 20.16 |
| 8 | 54 | 178 | 26.58 | 706.50 |
| 9 | 64 | 156 | −17.74 | 314.71 |
| 10 | 66 | 184 | 5.80 | 33.64 |
| 11 | 76 | 176 | −24.52 | 601.23 |
| 12 | 78 | 225 | 20.02 | 400.80 |
| | $\Sigma x = 592$ | $\Sigma y = 1692$ | $\Sigma(y - \hat{y}) = -.01$ | $\Sigma(y - \hat{y})^2 = 2448.86$ |

SSE = 2448.86

$$S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{2448.86}{10}} = 15.65$$

## Practice Problems

Q.02 Check where a linear regression model is appropriate for the following data.

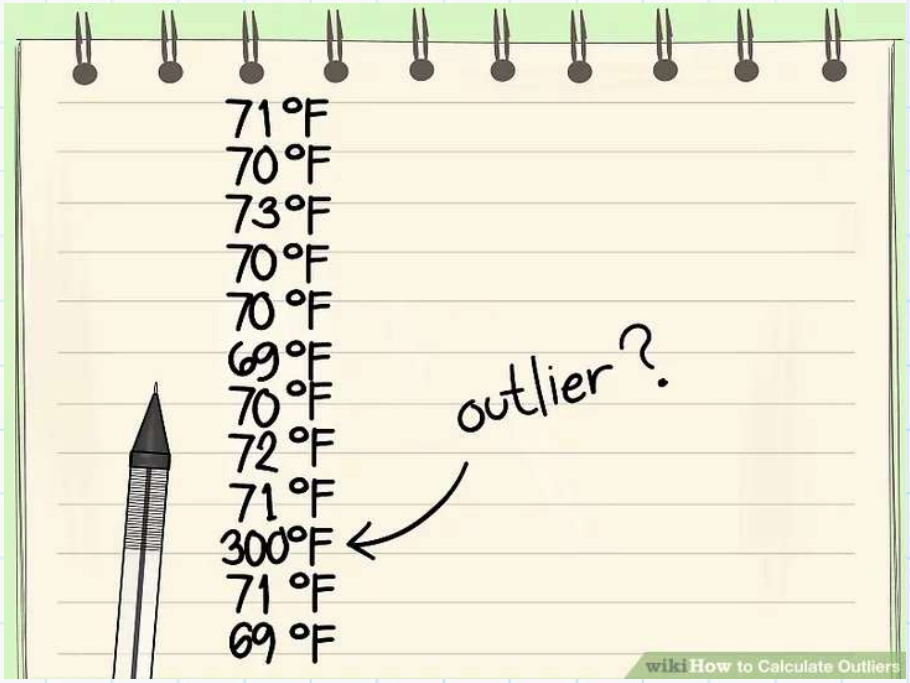
| | | | | | |
|---|---|---|---|---|---|
| x | 60 | 70 | 80 | 85 | 95 |
| y (Actual Value) | 70 | 65 | 70 | 95 | 85 |
| y^ (Predicted Value) | 65.41 | 71.84 | 78.28 | 81.50 | 87.94 |

Q.03 The equation of a regression line is and the data are as follows:

| | | | | | |
|---|---|---|---|---|---|
| x | 57 | 11 | 12 | 19 | 25 |
| y | 47 | 38 | 32 | 24 | 22 |

Solve for the residuals and graph a residual plot. Do these data seem to violate any of the assumptions of regression?
Determine SSE and SE of the Estimate.

# Assumption : Outliers / Influential Cases

## Assumption : Outliers / Influential Cases

- An outlier is a data point which is very far, somehow, from the rest of the data.
- An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an extreme value.
- When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. Outlier is a value that lies in a data series on its extremes, which is either very small or large and thus can affect the overall observation made from the data series. Outliers are also termed as extremes because they lie on the either end of a data series.
- Outliers are usually treated as abnormal values that can affect the overall observation due to its very high or low extreme values and hence should be discarded from the data series.

## Assumption : Outliers / Influential Cases

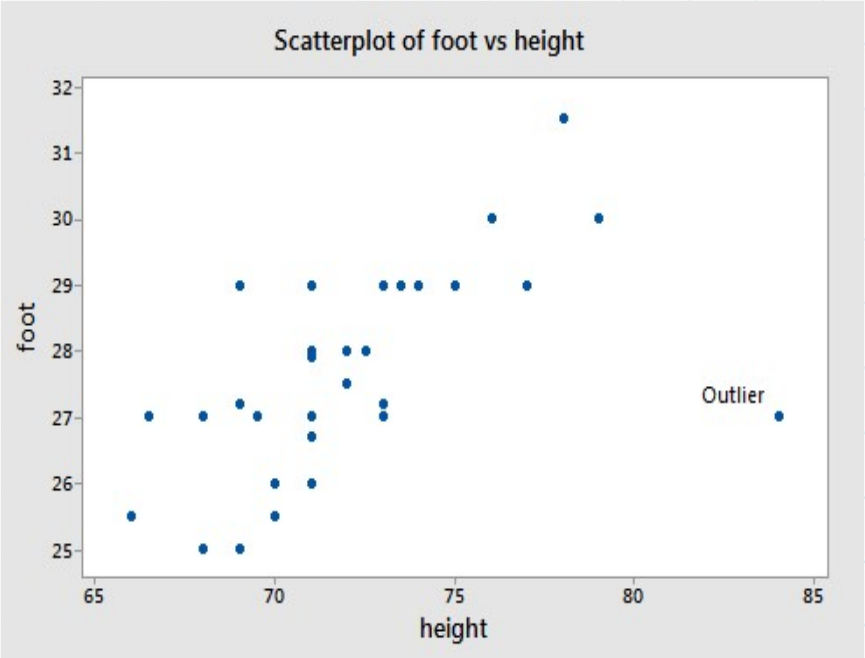
- Most common causes of outliers on a data set:
    1. Data entry errors (human errors)
    2. Measurement errors (instrument errors)
    3. Experimental errors (data extraction or experiment planning/executing errors)
    4. Intentional (dummy outliers made to test detection methods)
    5. Data processing errors (data manipulation or data set unintended mutations)
    6. Sampling errors (extracting or mixing data from wrong or various sources)
    7. Natural (not an error, novelties in data)

- Univariate outliers can simply be identified by considering the distributions of individual variables say by using boxplots. Multivariate outliers can be detected from residual scatterplots.

## Assumption : Outliers / Influential Cases

Example:

Let us consider a dataset where y = foot length (cm) and x = height (in) for n = 33 male students in a class. A scatterplot of the male foot length and height data shows one point labeled as an outlier.

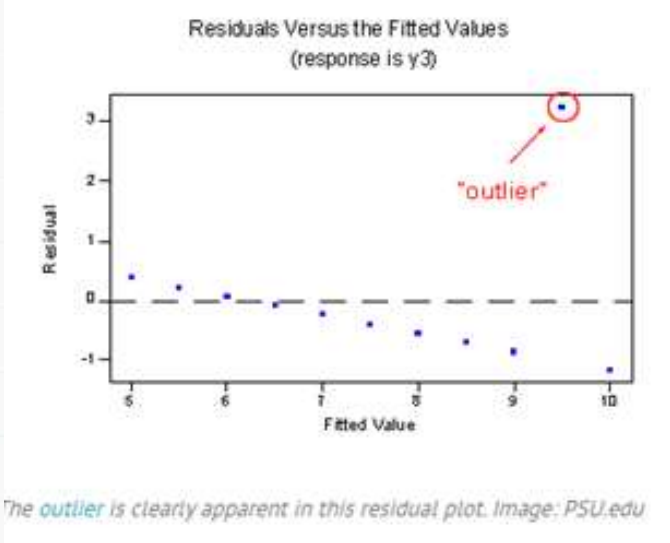



- There is a clear outlier with values $(x_i , y_i) = (84, 27)$.
- If that data point is *deleted* from the dataset, the estimated equation, using the other 32 data points, is y^i=0.253+0.384xi.
- For the deleted observation, $x_i = 84$, so
  ^yi=0.253+0.384(84)=32.5093
- The (unstandardized) deleted residual is

  di= 27−32.5093=−5.5093

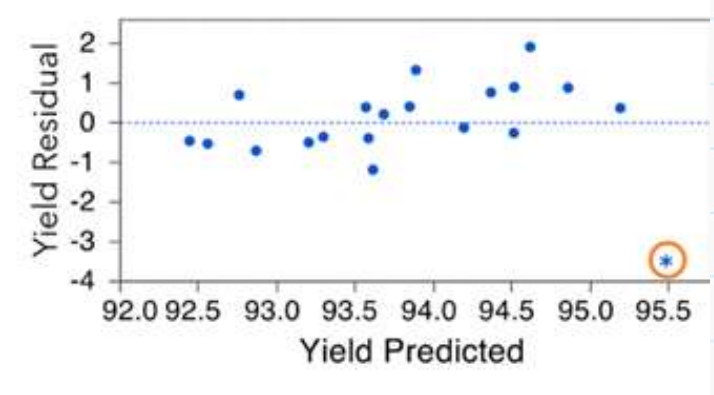
## Assumption : Outliers / Influential Cases

Residual Plots to deduct Multivariate Outliers:




The outlier is clearly apparent in this residual plot. Image: PSU.edu

Another example is the regression model for **Yield** as a function of **Concentration** is significant, but note that the line of fit appears to be tilted towards the outlier. We can see the effect of this outlier in the residual by predicted plot. The center line of zero does not appear to pass through the points.



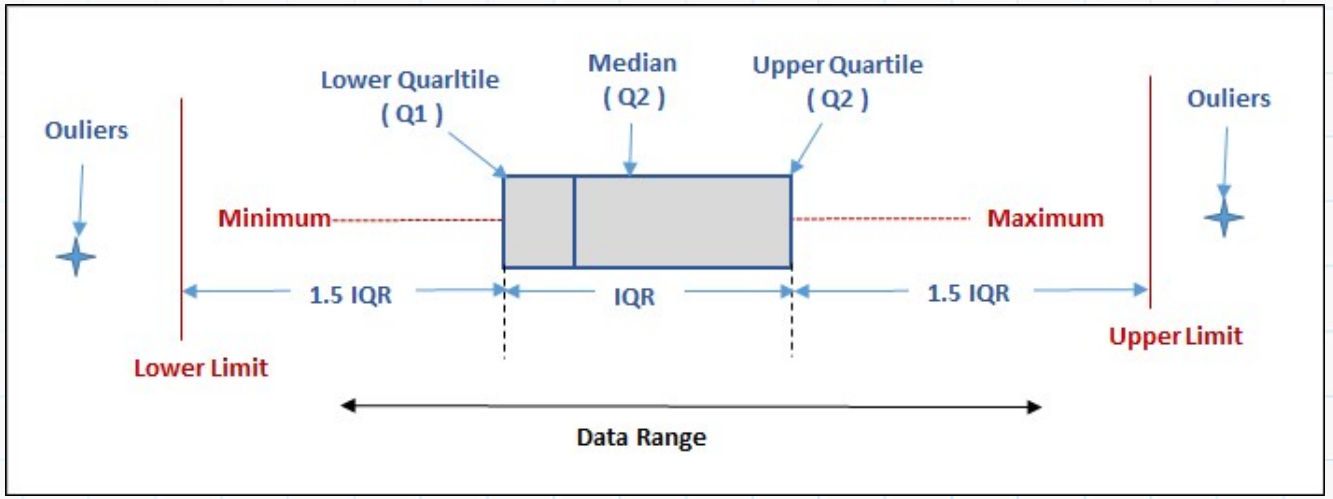
## Assumption : Outliers / Influential Cases

Box Plot Diagram to identify Outliers Lower

- Box plot diagram also termed as Whisker's plot is a graphical method typically depicted by quartiles and inter quartiles that helps in defining the upper limit and lower limit beyond which any data lying will be considered as outliers.
- The very purpose of this diagram is to identify outliers and discard it from the data series before making any further observation so that the conclusion made from the study gives more accurate results not influenced by any extremes or abnormal values.
- Box plots can be used as an initial screening tool for outliers as they provide a graphical depiction of data distribution and extreme values

## Assumption : Outliers / Influential Cases

Box Plot Diagram to identify Outliers Lower

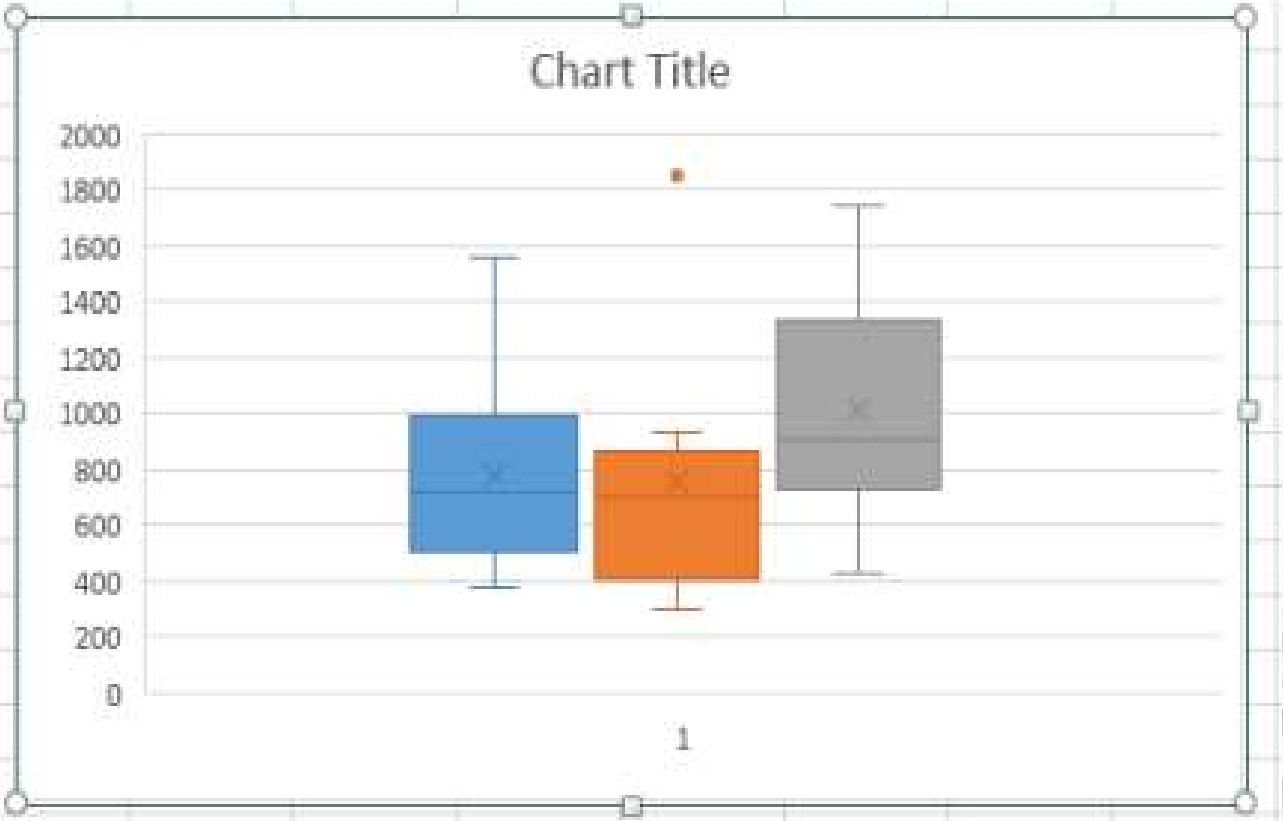



Where

- n be the number of data values in the data set.
- The Median (Q2) is the middle value of the data set.
- The Lower quartile (Q1) is the median of the lower half of the data set=$\frac{1}{4}(n+1)th\ term$
- The Upper quartile (Q3) is the median of the upper half of the data set=$\frac{3}{4}(n+1)th\ term$
- The Interquartile range (IQR) is the spread of the middle 50% of the data values.
- Interquartile Range (IQR) = Upper Quartile (Q3) – Lower Quartile (Q1) = Q3 – Q1
- Lower Limit = Q1 – 1.5 IQR.
- Upper Limit = Q3 + 1.5 IQR

# Assumption : Outliers / Influential Cases

| | A | B | C |
|---|---|---|---|
| 1 | Brand A | Brand B | Brand C |
| 2 | 1020 | 840 | 1430 |
| 3 | 1560 | 940 | 1750 |
| 4 | 560 | 780 | 870 |
| 5 | 780 | 650 | 920 |
| 6 | 990 | 720 | 1300 |
| 7 | 670 | 430 | 890 |
| 8 | 510 | 1850 | 740 |
| 9 | 490 | 300 | 720 |
| 10 | 380 | 360 | 430 |
| 11 | 880 | 690 | 1050 |
| 12 | | | |
| 13 | | | |
| 14 | | | |
| 15 | | | |



## Assumption : Autocorrelation

- Autocorrelation refers to the degree of correlation between the values of the same variables across different observations in the data.
- The concept of autocorrelation is most often discussed in the context of time series data in which observations occur at different points in time (e.g., air temperature measured on different days of the month).
- For example, one might expect the air temperature on the 1st day of the month to be more similar to the temperature on the 2nd day compared to the 31st day. If the temperature values that occurred closer together in time are, in fact, more similar than the temperature values that occurred farther apart in time, the data would be auto correlated.
- In a regression analysis, autocorrelation of the regression residuals can also occur if the model is incorrectly specified.
- For example, if you are attempting to model a simple linear relationship but the observed relationship is non-linear (i.e., it follows a curved or U-shaped function), then the residuals will be auto correlated.

## Assumption : Autocorrelation

How to Detect Autocorrelation

- A common method of testing for autocorrelation is the Durbin-Watson test.
- The Durbin-Watson tests produces a test statistic that ranges from 0 to 4.
- Values close to 2 (the middle of the range) suggest less autocorrelation, and values closer to 0 or 4 indicate greater positive or negative autocorrelation respectively.

## Coefficient of Determination – $R^2$ Statistics (Accessing the accuracy)

- The coefficient of determination is the square of the coefficient of correlation.
- Or The coefficient of determination is the proportion of variability of the dependent variable (y) accounted for or explained by the independent variable (x).
- The coefficient of determination ranges from 0 to 1. An $r^2$ of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of y by x.
- An $r^2$ of 1 means perfect prediction of y by x and that 100% of the variability of y is accounted for by x.
- Of course, most $r^2$ values are between the extremes.
- The researcher must interpret whether a particular $r^2$ is high or low, depending on the use of the model and the context within which the model was developed.

## Coefficient of Determination – $R^2$ Statistics

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{\sum y^2 - \frac{(\sum y)^2}{n}}$$

- The coefficient of multiple determination ($R^2$) is analogous to the coefficient of determination ($r^2$).
- $R^2$ represents the proportion of variation of the dependent variable, y, accounted for by the independent variables in the regression model.
- As with $r^2$, the range of possible values for $R^2$ is from 0 to 1.

## Coefficient of Determination – $R^2$ Statistics

- R-squared is always between 0 and 100%:
- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.
- In general, the higher the R-squared, the better the model fits your data.
- Of course, it is desirable for R2 to be high, indicating the strong predictability of a regression model.
- The coefficient of multiple determination can be calculated by the following formula..

## Coefficient of Determination – $R^2$ Statistics

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

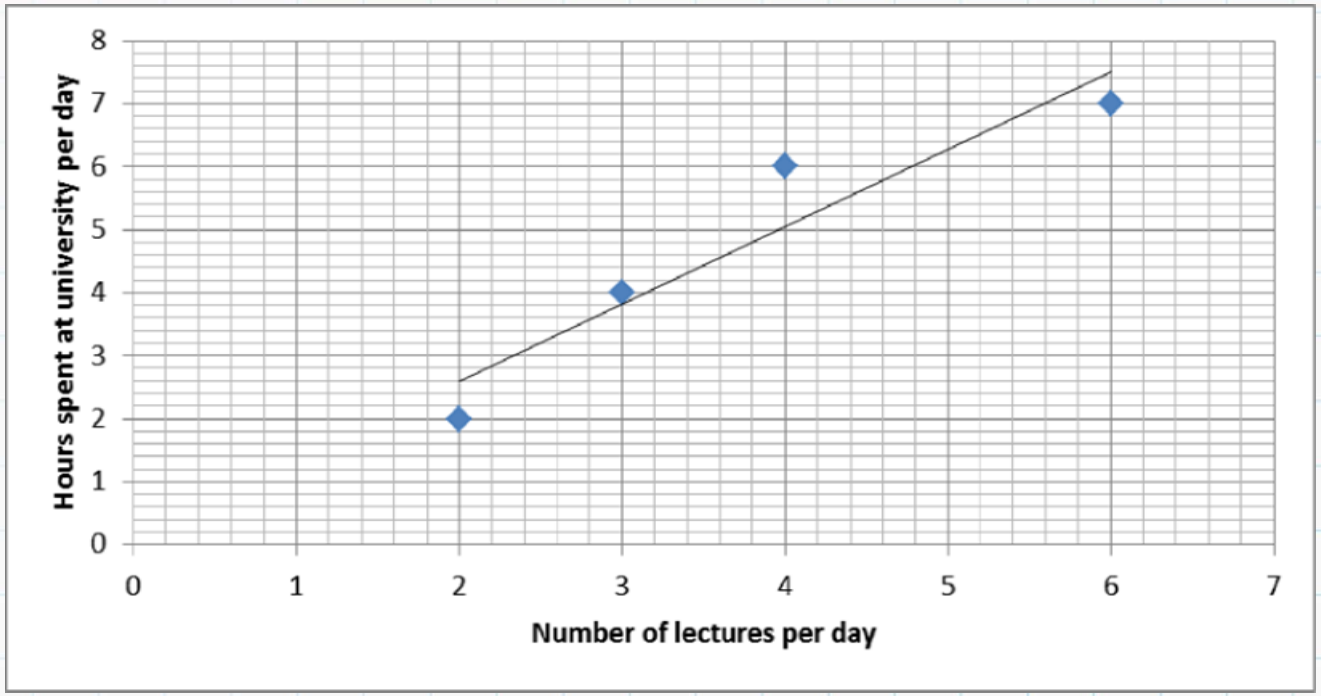
*Key Limitations of R-squared:*

- R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots.

- R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!

## Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation $y\hat{}=0.143+1.229x$. Calculate $R^2$.



## Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation $\hat{y}=0.143+1.229x$. Calculate $R^2$.

To calculate $R^2$ you need to find the sum of the residuals squared and the total sum of squares.

Start off by finding the residuals, which is the distance from regression line to each data point. Work out the predicted $y$ value by plugging in the corresponding $x$ value into the regression line equation.

- For the point $(2, 2)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x \\ &= 0.143 + (1.229 \times 2) \\ &= 0.143 + 2.458 \\ &= 2.601\end{aligned}$$

The actual value for $y$ is $2$.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value} \\ r_1 &= y_i - \hat{y}_i \\ &= 2 - 2.601 \\ &= -0.601\end{aligned}$$

## Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation y^=0.143+1.229x. Calculate $R^2$.

As you can see from the graph the actual point is below the regression line, so it makes sense that the residual is negative.

- For the point $(3, 4)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x \\ &= 0.143 + (1.229 \times 3) \\ &= 0.143 + 3.687 \\ &= 3.83\end{aligned}$$

The actual value for $y$ is 4.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value} \\ r_2 &= y_i - \hat{y}_i \\ &= 4 - 0.3.83 \\ &= 0.17\end{aligned}$$

As you can see from the graph the actual point is above the regression line, so it makes sense that the residual is positive.

## Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation y^=0.143+1.229x. Calculate $R^2$.

- For the point $(4, 6)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x \\ &= 0.143 + (1.229 \times 4) \\ &= 0.143 + 4.916 \\ &= 5.059\end{aligned}$$

The actual value for $y$ is $6$.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value} \\ r_3 &= y_i - \hat{y}_i \\ &= 6 - 5.059 \\ &= 0.941\end{aligned}$$

- For the point $(6, 7)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x \\ &= 0.143 + (1.229 \times 6) \\ &= 0.143 + 7.374 \\ &= 7.517\end{aligned}$$

## Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation y^=0.143+1.229x. Calculate $R^2$.

The actual value for $y$ is $7$.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value}\\ r_4 &= y_i - \hat{y}_i\\ &= 7 - 7.517\\ &= -0.517\end{aligned}$$

To find the residuals squared we need to square each of $r_1$ to $r_4$ and sum them.

$$\begin{aligned}\sum (y_i - \hat{y}_i)^2 &= \sum r_i\\ &= {r_1}^2 + {r_2}^2 + {r_3}^2 + {r_4}^2\\ &= (-0.601)^2 + (0.17)^2 + (0.941)^2 - (-0.517)^2\\ &= 1.542871\end{aligned}$$

To find $\sum (y_i - \bar{y})^2$ you first need to find the mean of the $y$ values.

$$\begin{aligned}\bar{y} &= \frac{\sum y}{n}\\ &= \frac{2+4+6+7}{4}\\ &= \frac{19}{4}\\ &= 4.75\end{aligned}$$

## Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation $\hat{y}=0.143+1.229x$. Calculate $R^2$.

Now we can calculate $\sum(y_i - \bar{y})^2$.

$$\begin{aligned}\sum(y_i - \bar{y})^2 &= (2-4.75)^2 + (4-4.75)^2 + (6-4.75)^2 + (7-4.75)^2 \\ &= (-2.75)^2 + (-0.75)^2 + (1.25)^2 + (2.25)^2 \\ &= 14.75\end{aligned}$$

Therefore;

$$\begin{aligned}R^2 &= 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}} \\ &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \\ &= 1 - \frac{1.542871}{14.75} \\ &= 1 - 0.105 \text{ (3.s.f)} \\ &= 0.895 \text{ (3.s.f)}\end{aligned}$$

This means that the number of lectures per day account for 89.5% of the variation in the hours people spend at university per day.

## Adjusted $R^2$

- The adjusted R-squared is a modified version of R-squared for the number of predictors in a model.
- R2 assumes that every single variable explains the variation in the dependent variable.
- The adjusted R2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.
- The value of R Squared never decreases.
- Adding new independent variables will result in an increased value of R Squared.
- This is a major flow as R Squared will suggest that adding new variables irrespective of whether they are really significant or not, will increase the value.
- For example, the person's Name for predicting the Salary, the value of R squared will increase suggesting that the model is better.
- This is where Adjusted R Squared comes to the rescue.

## Adjusted $R^2$

- Compared to R Squared which can only increase, Adjusted R Squared has the capability to decrease with the addition of less significant variables, thus resulting in a more reliable and accurate evaluation.

$$Adj\ R^2 = 1 - \frac{SSE/(n-k-1)}{SS_{yy}/n-1} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

Note:

- Adjusted R-squared value always be less than or equal to r-squared value.
- Adjusted R-square should be used while selecting important predictors (independent variables) for the regression model.
- If you add more and more useless variables to a model, adjusted R-squared will decrease.
- If you add more useful variables, adjusted R-squared will increase.

## Problems

A fund has a sample R-squared value close to 0.5 and it is doubtlessly offering higher risk adjusted returns with the sample size of 50 for 5 predictors. Find Adjusted R square value.

Solution: .
Sample size = 50 Number of predictor = 5 Sample R - square = 0.5.Substitute the qualities in the equation

$$R^2_{adj} = 1 - [\frac{(1-0.5^2)(50-1)}{50-5-1}]$$

$$= 1 - (0.75) \times \frac{49}{44},$$

$$= 1 - 0.8352,$$

$$= 0.1648$$

## Problems

A regression model has 9 independent variables, 47 observations, and R2 = 0.879. Calculate adjusted R squared.

## Understanding Adjusted R Square

Consider an example using data collected by a pizza owner, as shown below:

| Temperature (Celsius) X1 | Price of Dough X2 | Price of Pizza Y1 |
|---|---|---|
| 21 | 1 | 5 |
| 15 | 3 | 12 |
| 16 | 6 | 15 |
| 21 | 8 | 19 |
| 27 | 12 | 24 |
| 24 | 15 | 27 |
| 21 | 17 | 29 |
| 23 | 21 | 31 |
| 21 | 26 | 36 |

## Understanding Adjusted R Square

Assume the pizza owner runs two regressions:

Model 1: Price of Dough (input variable), Price of Pizza (output variable)

$\Rightarrow y_{PP} = \beta_0 + \beta_1 x_{PD} + \varepsilon$

R-square= 0.9557 and an adjusted R-square= 0.9493.

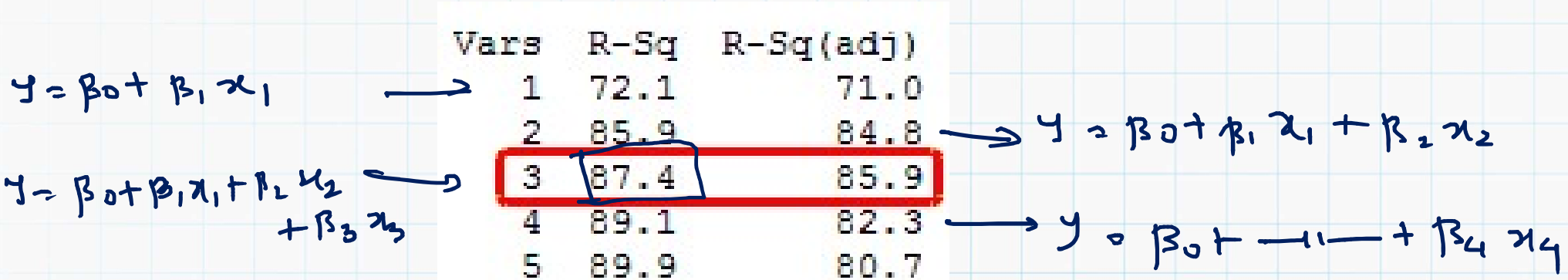
Model 2: Temperature (input variable 1), Price of Dough (input variable 2), Price of Pizza (output variable)

R-square= 0.9573 and an adjusted R-square=0.9431.

$y_{PP} = \beta_0 + \beta_1 x_T + \beta_2 x_{PD} + \varepsilon$

## Problems

Comment on following table:




**Solution:**

We can see where the adjusted R-squared peaks, and then declines. Meanwhile, the R-squared continues to increase. We might want to include only three predictors in this model.

## Problems

Which model should be used? Information regarding both models are provided below:

| | Model 1 | Model 2 |
|---|---|---|
| Variables Used | X1, X2, X3, Y1 | X1, X2, Y1 |
| R-squared | 0.5923 | 0.5612 |
| Adjusted R-squared | 0.4231 | 0.3512 |

**Solution:**

- Comparing the R-squared between Model 1 and Model 2, the R-squared predicts that Model 1 is a better model as it carries greater explanatory power (0.5923 in Model 1 vs. 0.5612 in Model 2).
- Comparing the R-squared between Model 1 and Model 2, the adjusted R-squared predicts that the input variable X3 contributes to explaining output variable Y1 (0.4231 in Model 1 vs. 0.3512 in Model 2).
- As such, Model 1 should be used, as the additional X3 input variable contributes to explaining the output variable Y1.

## Problems

Calculate R-Squared and Adjusted R-Squared with 3 independent variables and interpret the result:

| Y | Y^ |
|---|---|
| 21 | 21.5 |
| 21 | 21.14 |
| 22.8 | 26.1 |
| 21.4 | 20.2 |
| 18.7 | 17.5 |
| 18.1 | 19.7 |
| 14.3 | 14.9 |
| 24.4 | 22.5 |
| 22.8 | 25.1 |
| 19.2 | 18 |

Mean = Ym = 20.37,

k=3,

n=10

## Problems

Calculate R-Squared and Adjusted R-Squared with 3 independent variables and interpret the result:

| Y | Y^ | Y - Y^ | Y - Ym | $(Y - Y^{\wedge})^2$ | $(Y - Ym)^2$ |
|---|---|---|---|---|---|
| 21 | 21.5 | -0.5 | 0.63 | 0.25 | 0.40 |
| 21 | 21.14 | -0.14 | 0.63 | 0.0196 | 0.40 |
| 22.8 | 26.1 | -3.3 | 2.43 | 10.89 | 5.90 |
| 21.4 | 20.2 | 1.2 | 1.03 | 1.44 | 1.06 |
| 18.7 | 17.5 | 1.2 | -1.67 | 1.44 | 2.79 |
| 18.1 | 19.7 | -1.6 | -2.27 | 2.56 | 5.15 |
| 14.3 | 14.9 | -0.6 | -6.07 | 0.36 | 36.84 |
| 24.4 | 22.5 | 1.9 | 4.03 | 3.61 | 16.24 |
| 22.8 | 25.1 | -2.3 | 2.43 | 5.29 | 5.90 |
| 19.2 | 18 | 1.2 | -1.17 | 1.44 | 1.37 |

**Problems**

Calculate R-Squared and Adjusted R-Squared with 3 independent variables and interpret the result:

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} = 64.11\%$$

$$Adj\ \mathrm{R}^2 = 1 - \left[ \frac{(1 - R^2)(n-1)}{n-k-1} \right] = 46.16\%$$

## Assumption : Multicollinearity

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

$$(x_1 \neq x_2) \neq x_3 \neq \cdots \neq x_p)$$

- Multicollinearity occurs when independent variables in a regression model are correlated.
- This correlation is a problem because independent variables should be independent.
- If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.
- For example, height and weight, household income and water consumption, mileage and price of a car, study time and leisure time, etc.

## Assumption : Multicollinearity

- Multicollinearity is a statistical phenomenon in which there exists a perfect or exact relationship between the predictor variables.

- When there is a perfect or exact relationship between the predictor variables, it is difficult to come up with reliable estimates of their individual coefficients.

- It will result in incorrect conclusions about the relationship between outcome variable and predictor variables.

- Remedial Measures: To drop one or several predictor variables in order to lessen the multicollinearity

## Assumption : Multicollinearity

*How to detect Multicollinearity?*

- There is a very simple test to assess multicollinearity in your regression model. The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation.
- VIFs start at 1 and have no upper limit. A value of 1 indicates that there is no correlation between this independent variable and any others. VIFs between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures. VIFs greater than 5 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

The variance inflation factor for the $j^{th}$ predictor is:

$$VIF_j = \frac{1}{1 - R_j^2}$$

## Problem

This is a case of a Malaysian telecom operator that was interested in customer churn analysis. The table below shows a glimpse of a few of the important features of the dataset:

Sample data:

| Gender | AgeGroup | Dependents | contract_age(Mor | Contract_renewal_type | PaperlessBilling | MonthlyChar | TotalCharge: | Customer_status |
|---|---|---|---|---|---|---|---|---|
| Female | 25_to_30 | No | 1 | Monthly | Yes | 30 | 30 | 0 |
| Male | 30_to_50 | No | 34 | Yearly | No | 59 | 1900 | 0 |
| Male | below_25 | No | 2 | Monthly | Yes | 55 | 109 | 1 |
| Male | 25_to_30 | No | 45 | Yearly | No | 43 | 1841 | 0 |
| Female | 30_to_50 | No | 2 | Monthly | Yes | 72 | 152 | 1 |
| Female | 25_to_30 | No | 8 | Monthly | Yes | 101 | 826 | 1 |
| Male | below_25 | Yes | 22 | Monthly | Yes | 90 | 1950 | 0 |
| Female | 25_to_30 | No | 10 | Monthly | No | 30 | 302 | 0 |

*Contract_age (months) – for how long the customer has been with the operator*

*MonthlyCharges – charges per month*

*TotalCharges – total charges so far from the beginning of the contract*

*Customer_status = 0: active, 1: inactive (churn) – Target variable*

**Solution:**

Contract age VIF=5.7

Monthly Charges VIF=3.16

Total Charges VIF=9.27

Clearly, the VIF of TotalCharges is higher than the other 2 features. If 5 is the cut-off point to identify the variables causing multicollinearity, do we remove contract_age and TotalCharges features both at a time? No. We first remove the one with the highest VIF and calculate the VIFs again.

(After removing TC)

Contract age VIF=1.07

Monthly Charges VIF=1.07

Here we go! Just removing TotalCharges from the data frame brought down the VIF of other variables to the desired value ~ 1. Also, it is easy to guess that as the contract_age is in months, we also expect the charges to be monthly, and as expected, VIF did the job!

## Assumption : Homoscedasticity → Equality of variance

- The assumption of homoscedasticity is that the residuals are approximately equal for all predicted values. Homoscedasticity means “same variance” .
- In other words residuals are equal across regression line.
- The presence of non-constant variance is referred to heteroscedasticity.

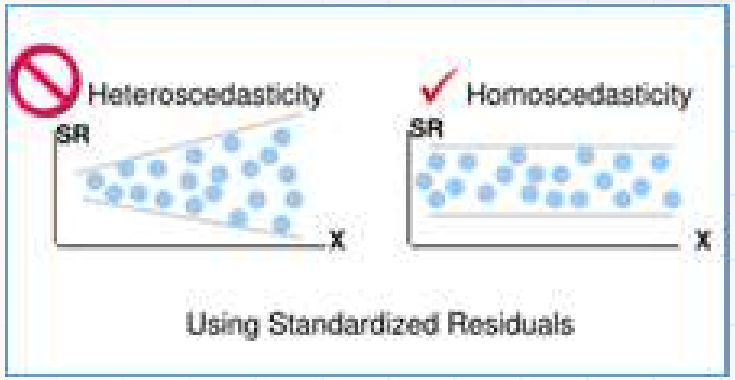
## Assumption : Homoscedasticity

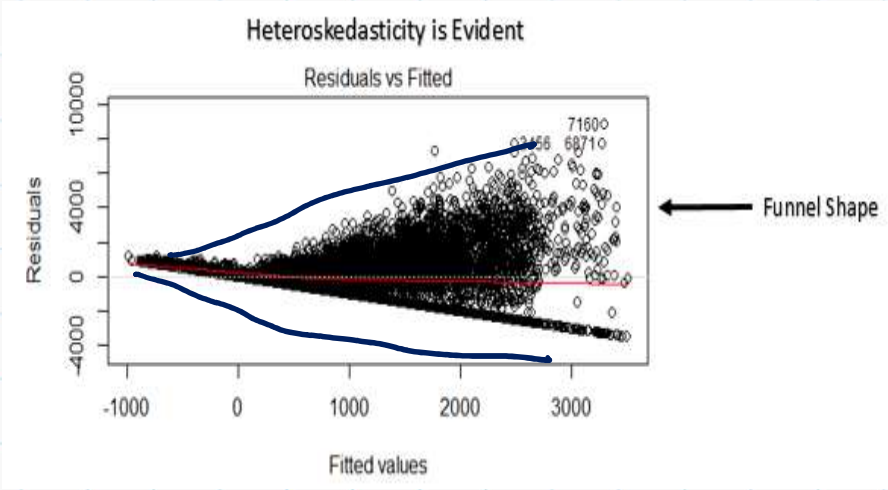
*Possible reasons of arising Heteroscedasticity:*

1. Often occurs in those data sets which have a large range between the largest and the smallest observed values i.e. when there are outliers.
2. When model is not correctly specified.
3. If observations are mixed with different measures of scale.
4. When incorrect transformation of data is used to perform the regression.
5. Skewness in the distribution of a regressor, and may be some other sources.

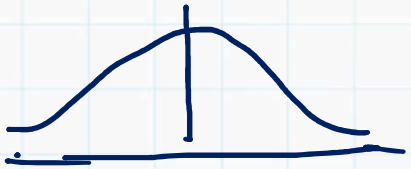
## Assumption : Homoscedasticity

- Homoscedasticity can also be tested using scatter plot of residual vs fitted values.
- If heteroscedasticity exists, the plot would exhibit a funnel shape pattern. For example:



## Assumption : Independence of Error

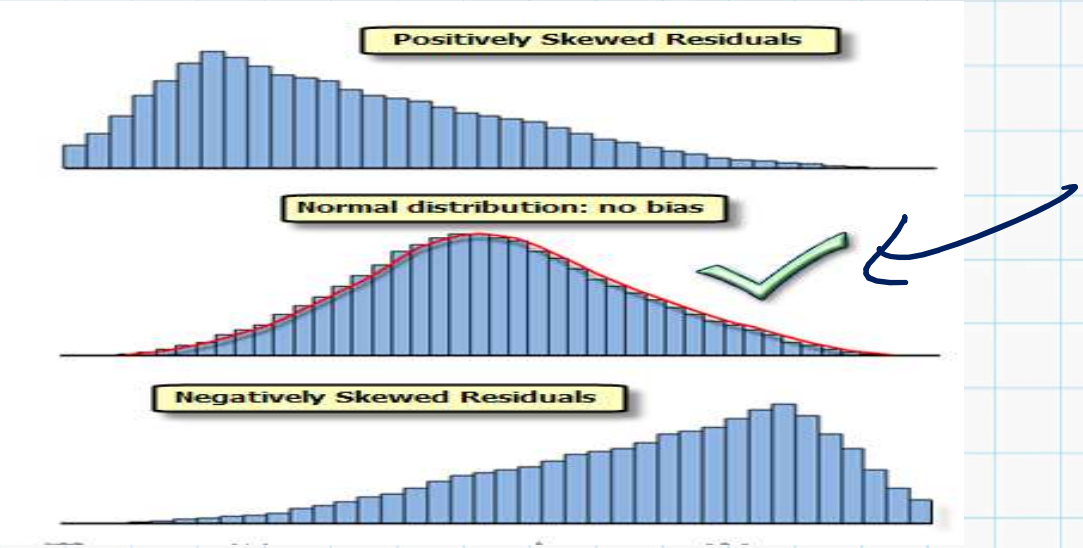


- The values of the residuals are normally distributed.
- This assumption can be tested by looking at the distribution of residuals.
- We can do this by CHECKING the Histogram and Normal probability plot.
- However, unless the residuals are far from normal or have an obvious pattern, we generally don't need to be overly concerned about normality.

| Pattern | What the pattern may indicate |
|---|---|
| A long tail in one direction | Skewness |
| A bar that is far away from the other bars | An outlier |

## Assumption : Independence of Error

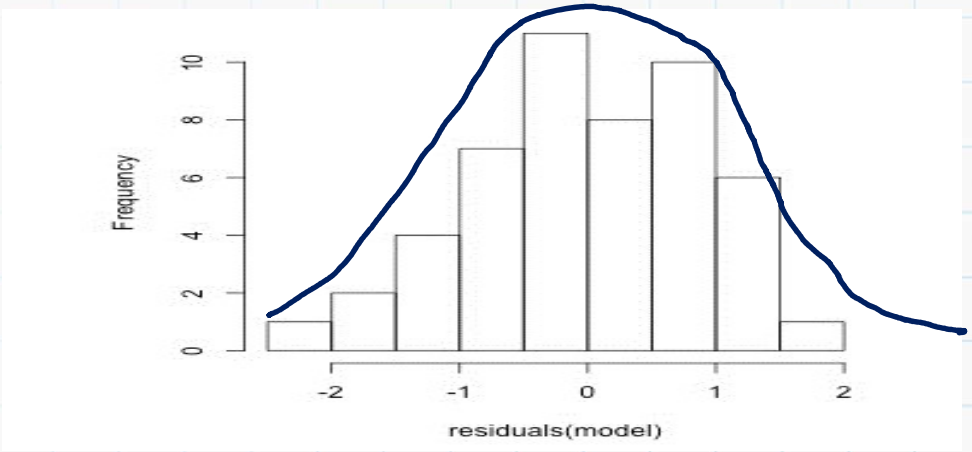
- Because the appearance of a histogram depends on the number of intervals used to group the data,
- A histogram is most effective when you have approximately 20 or more data points.
- If the sample is too small, then each bar on the histogram does not contain enough data points to reliably show skewness or outliers.



## Assumption : Independence of Error

Preprocessing
of data → 90 | 10 → Model

- The following histogram of residuals suggests that the residuals (and hence the error terms) are normally distributed




1) linearity
2) Outliers
3) Autocorrelation
4) Homoscadesticity
5) multicollinearity
6) Ind. of Error

- Sample sizes of residuals are generally small (<50) because experiments have limited treatment combinations, so a histogram is not be the best choice for judging the distribution of residuals.

- A more sensitive graph is the normal probability plot.

## Testing the Overall Model

- It is common in regression analysis to compute an F test to determine the overall significance of the model.
- Most computer software packages include the F test and its associated ANOVA table as standard regression output.
- This test determines whether at least one of the regression coefficients (from multiple predictors) is different from zero.
- Simple regression provides only one predictor and only one regression coefficient to test.

## Testing the Overall Model

Here are the five steps of the overall F-test for regression:

*Step I: State the null and alternative hypotheses:*

The null hypothesis states that the model with no independent variables fits the data as well as your model. Whereas the alternative hypothesis says that your model fits the data better than the intercept-only model.

The hypotheses being tested in simple regression by the F test for overall significance are:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

And for a multiple regression model with intercept, we want to test the following null hypothesis and alternative hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots\ldots\ldots = \beta_{k-1} = 0$$
$$H_1 : \beta_j \neq 0, \text{ For at least one value of j.}$$

## Testing the Overall Model

Here are the five steps of the overall F-test for regression:

Step II: Compute the test statistic F:

1. n is the number of observations, k is the number of independent variables.

2. Sum of Squares for Error = $SSE = \sum (y - \hat{y})^2$

3. Corrected Sum of Squares Total = $SST = \sum (y - \bar{y})^2$

4. Corrected Sum of Squares for Model = SSR = $SSM = \sum (\hat{y} - \bar{y})^2$

## Testing the Overall Model

Here are the five steps of the overall F-test for regression:

Step II: Compute the test statistic F:

5. SSM+SSE=SST
6. Mean of Squares for Model: MSM = SSM / k
7. Mean of Squares for Error: MSE = SSE / (n-k-1)
8. Mean of Squares Total: MST = SST /n-1
9. $F = \frac{MSM}{MSE} = \frac{SSM/k}{SSE/n-k-1}$

## Testing the Overall Model

Here are the five steps of the overall F-test for regression:

Step III: Find a (1 - α) 100% confidence interval for degrees of freedom using an F-table or statistical software.

Step IV: Decide whether to accept or reject the null hypothesis.

If Cal Value(P value)< Tabulated Value: Accept Null Hypothesis otherwise reject Null Hypothesis.

## Testing the Overall Model

**Degrees of freedom (df):**

- Regression df is the number of independent variables in our regression model=p=k-1
- Residual df is the total number of observations (rows) of the dataset subtracted by the number of variables being estimated=n-p=n-k-1
- Total df =n-1

p-value < los (0.05)
Reject Ho

**p Value:**

- To determine whether any of the differences between the means are statistically significant, compare the p-value to your significance level to assess the null hypothesis.
- If the p-value is less than or equal to the significance level, you reject the null hypothesis.
- If the p-value is greater than the significance level, you do not have enough evidence to reject the null hypothesis

p-value > los (0.05)
accept Ho

# Testing the Overall Model

## ANOVA for Significance of Regression

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | F Ratio | P Value |
|---|---|---|---|---|---|
| Regression /Model | SSR | (k) | $MSR = \frac{SSR}{k}$ | $F = \frac{MSR}{MSE}$ | |
| Error /Residual | SSE | (n-(k+1)) =(n-k-1) | $MSE = \frac{SSE}{(n-(k+1))}$ | | |
| Total | SST | (n-1) | $MST = \frac{SST}{(n-1)}$ | | |

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

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

$$\bar{R}^2 = 1 - \frac{\frac{SSE}{(n-(k+1))}}{\frac{SST}{(n-1)}} = \frac{MSE}{MST}$$

F-Tab

$\bar{R}^2$

$R^2$

F-cal

Adjusted $R^2$

## Interpreting Multiple Regression Computer Output

Many of the concepts discussed thus far are highlighted.

Note the following items:

1. The equation of the regression model
2. The ANOVA table with the F value for the overall test of the model
3. The t ratios, which test the significance of the regression coefficients
4. The value of SSE and MSE
5. The value of standard error
6. The value of $R^2$
7. The value of adjusted $R^2$

# Interpreting Multiple Regression Computer Output

# Interpreting Multiple Regression Computer Output

*1. The multiple linear regression equation :*

The multiple linear regression equation is just an extension of the simple linear regression equation – it has an “x” for each explanatory variable and a coefficient for each “x”.

*2. Interpretation of the coefficients:*

In the multiple linear regression equation The coefficients in the equation are the numbers in front of the x’s. Each “x” has a coefficient. We should understand what these values mean in the context of the problem.

## Interpreting Multiple Regression Computer Output

(4.9 , 5.6)

*3.Confidence intervals for the coefficients:*

In the multiple linear regression equation A confidence interval is of the form of best estimate ± margin of error In general: Formula for confidence interval for a coefficient (βi):

t* = t cal

**Formula for confidence interval for a coefficient ($\beta_i$):**

$$b_i \pm (t^*_{n-v-1})(SE(b_i))$$

↓ β - coefficient

Note 1:
The degrees of freedom for the t* critical value is the DFE in the Analysis of Variance table.
(Recall, DFE = n – k – 1 where k = the number of explanatory variables)

Note 2:
The subscript "i" in the formula are for the specific explanatory variable.

# Interpreting Multiple Regression Computer Output

*4. Using the Multiple Linear Regression equation for prediction:*

One of the uses of a regression analysis is for prediction. Predicting using a multiple linear regression equation is just an extension of predicting with a simple linear regression equation. We just have to make sure to put the right values in for the right x's.

*5. Determining a final model – how to choose "significant" predictors of the response variable:*

Another reason for performing a multiple linear regression analysis is to determine which (if any) of the explanatory variables are significant predictors of the response variable. It means to identify explanatory variables which are useful predictors of the response variable (i.e. help to "explain" the response variable That is, does each explanatory variable

# Case I – Understanding the Output

A health researcher wants to be able to predict "VO2max", an indicator of fitness and health. Normally, to perform this procedure requires expensive laboratory equipment and necessitates that an individual exercise to their maximum (i.e., until they can longer continue exercising due to physical exhaustion). This can put off those individuals who are not very active/fit and those individuals who might be at higher risk of ill health (e.g., older unfit subjects). For these reasons, it has been desirable to find a way of predicting an individual's VO2max based on attributes that can be measured more easily and cheaply. To this end, a researcher recruited 100 participants to perform a maximum VO2max test, but also recorded their "age", "weight", "heart rate" and "gender". Heart rate is the average of the last 5 minutes of a 20 minute, much easier, lower workload cycling test. The researcher's goal is to be able to predict VO2 max based on these four attributes: age, weight, heart rate and gender.

We have six variables:

1. $VO_2max$ (y) → dependent variable
2. Age ($x_1$)
3. Weight ($x_2$)
4. Heart Rate ($x_3$)
5. Gender ($x_4$)
6. CaseNo

→ ind. variables

# Case I – Understanding the Output

Interpreting and Reporting the Output:

The first table of interest is the **Model Summary** table. This table provides the *R*, $R^2$, adjusted $R^2$, and the standard error of the estimate, which can be used to determine how well a regression model fits the data:

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .760[a] | .577 | .559 | 5.69097 |

a. Predictors: (Constant), gender, age, heart_rate, weight

The "**R**" column represents the value of *R*, the ***multiple correlation coefficient***. *R* can be considered to be one measure of the quality of the prediction of the dependent variable; in this case, $VO_2$max. A value of 0.760, in this example, indicates a good level of prediction. The "**R Square**" column represents the $R^2$ value (also called the coefficient of determination), which is the proportion of variance in the dependent variable that can be explained by the independent variables (technically, it is the proportion of variation accounted for by the regression model above and beyond the mean model). You can see from our value of 0.577 that our independent variables explain 57.7% of the variability of our dependent variable, $VO_2$max. However, you also need to be able to interpret "**Adjusted R Square**" (*adj.* $R^2$) to accurately report your data.

# Case I – Understanding the Output

Statistical Significance

The *F*-ratio in the **ANOVA** table (see below) tests whether the overall regression model is a good fit for the data. The table shows that the independent variables statistically significantly predict the dependent variable, $F(4, 95) = 32.393$, $p < .0005$ (i.e., the regression model is a good fit of the data).

**ANOVA**[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 4196.483 | 4 | 1049.121 | 32.393 | .000[b] |
| | Residual | 3076.778 | 95 | 32.387 | | |
| | Total | 7273.261 | 99 | | | |

a. Dependent Variable: VO2max

b. Predictors: (Constant), gender, age, heart_rate, weight

SSM

SSE

SST

p-value

p-value < LOS(0.05)

Fcal > Ftab (4, 95)

32.393 > = 2.247

# Case I – Understanding the Output

Estimated Model Coefficients

The general form of the equation to predict VO2max from age, weight, heart_rate, gender, is:

$\hat{y}$ $\beta_0$ $\beta_1 x_{age}$ $\beta_2 x_{weigh}$ $\beta_3 x_{hr}$ $\beta_4 x_{gender}$

Predicted VO2max = 87.83 – (0.165 x age) – (0.385 x weight) – (0.118 x heart_rate) + (13.208 x gender)

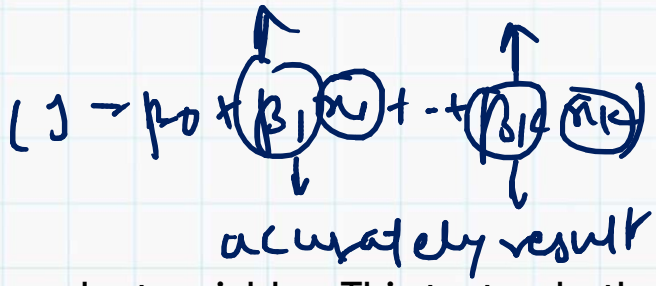
**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | **87.830** | 6.385 | | 13.756 | .000 | 75.155 | 100.506 |
| | age | **-.165** | .063 | -.176 | -2.633 | .010 | -.290 | -.041 |
| | weight | **-.385** | .043 | -.677 | -8.877 | .000 | -.471 | -.299 |
| | heart_rate | **-.118** | .032 | -.252 | -3.667 | .000 | -.182 | -.054 |
| | gender | **13.208** | 1.344 | .748 | 9.824 | .000 | 10.539 | 15.877 |

a. Dependent Variable: VO2max

Unstandardized coefficients indicate how much the dependent variable varies with an independent variable when all other independent variables are held constant. Consider the effect of age in this example. The unstandardized coefficient, $B_1$, for age is equal to -0.165 (see **Coefficients** table). This means that for each one year increase in age, there is a decrease in $VO_2$max of 0.165 ml/min/kg.

# Case I – Understanding the Output



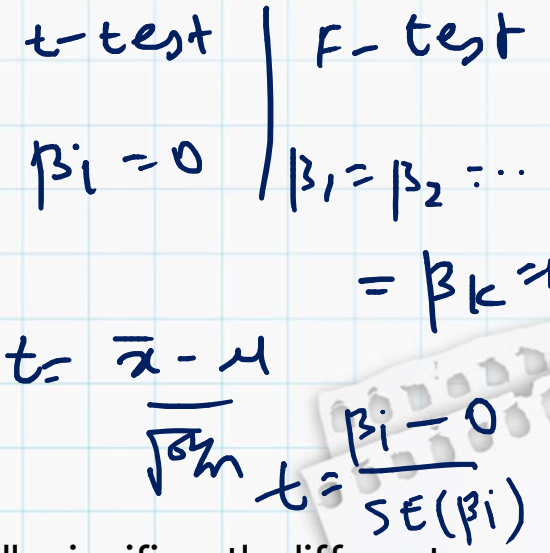
Statistical significance of the independent variables.

You can test for the statistical significance of each of the independent variables. This tests whether the unstandardized (or standardized) coefficients are equal to 0 (zero) in the population. If p < .05, you can conclude that the coefficients are statistically significantly different to 0 (zero). The t-value and corresponding p-value are located in the "t" and "Sig." columns, respectively, as highlighted below:

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Lower Bound | Upper Bound |
| 1 | (Constant) | 87.830 | 6.385 | | **13.756** | **.000** | 75.155 | 100.506 |
| | age | -.165 | .063 | -.176 | **-2.633** | **.010** | -.290 | -.041 |
| | weight | -.385 | .043 | -.677 | **-8.877** | **.000** | -.471 | -.299 |
| | heart_rate | -.118 | .032 | -.252 | **-3.667** | **.000** | -.182 | -.054 |
| | gender | 13.208 | 1.344 | .748 | **9.824** | **.000** | 10.539 | 15.877 |

a. Dependent Variable: VO2max



You can see from the "**Sig.**" column that all independent variable coefficients are statistically significantly different from 0 (zero). Although the intercept, $B_0$, is tested for statistical significance, this is rarely an important or interesting finding.

# Case I – Understanding the Output

Overall Conclusion:

A multiple regression was run to predict $VO_2$max from gender, age, weight and heart rate. These variables statistically significantly predicted $VO_2$max, $F(4, 95) = 32.393$, $p < .0005$, $R^2 = .577$. All four variables added statistically significantly to the prediction, $p < .05$.

Question: Predict the VO2 max for 40 years age, 55kg weight, heart rate 70 beats per minute of a female.

**Predicted VO2max = 87.83 – (0.165 x age) – (0.385 x weight) – (0.118 x heart_rate) + (13.208 x gender)**

$VO_2 max = 51.79$

| A | B | C |
|---|---|---|
| 1 | 0 | |
| 0 | 1 | |

# Case I – Understanding the Output

# Case II – Literacy Rate Example

Literacy rate is a reflection of the educational facilities and quality of education available in a country, and mass communication plays a large part in the educational process. In an effort to relate the literacy rate of a country to various mass communication outlets, a demographer has proposed to relate literacy rate to the following variables: number of daily newspaper copies (per 1000 population), number of radios (per 1000 population), and number of TV sets (per 1000 population). Here are the data for a sample of 10 countries:

# Case II – Literacy Rate Example

| Country | newspapers | radios | tv sets | literacy rate |
|---|---|---|---|---|
| Czech Republic / Slovakia | 280 | 266 | 228 | 0.98 |
| Italy | 142 | 230 | 201 | 0.93 |
| Kenya | 10 | 114 | 2 | 0.25 |
| Norway | 391 | 313 | 227 | 0.99 |
| Panama | 86 | 329 | 82 | 0.79 |
| Philippines | 17 | 42 | 11 | 0.72 |
| Tunisia | 21 | 49 | 16 | 0.32 |
| USA | 314 | 1695 | 472 | 0.99 |
| Russia | 333 | 430 | 185 | 0.99 |
| Venezuela | 91 | 182 | 89 | 0.82 |

# Case II – Literacy Rate Example

Below is the Minitab output from a Multiple Linear Regression analysis.

$x_1 = 200$, $x_2 = 800$, $x_3 = 250$

$\beta$

| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | 0.51486 | 0.09368 | 5.50 | 0.002 |
| newspaper copies | 0.0005421 | 0.0008653 | 0.63 | 0.554 |
| radios | -0.0003535 | 0.0003285 | -1.08 | 0.323 |
| television sets | 0.001988 | 0.001550 | 1.28 | 0.247 |

S = 0.186455  R-Sq = 69.9%  R-Sq(adj) = 54.8%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 3 | 0.48397 | 0.16132 | 4.64 | 0.053 |
| Residual Error | 6 | 0.20859 | 0.03477 | | |
| Total | 9 | 0.69256 | | | |

# Case II – Literacy Rate Example

Question 1: What is the response variable? What are the explanatory variables?

Response variable: Y = literacy rate.

Explanatory variables:

1. X1 = number of daily newspaper copies,
2. X2 = number of radios, and
3. X3 = number of TV sets
   (all per 1000 people in the population of the country).

# Case II – Literacy Rate Example

Question 2: Write the least-squares regression equation for this problem. Explain what each term in the regression equation represents in terms of the problem

$$\hat{y} = 0.51486 + 0.00054x1 - 0.00035x2 + 0.00199x3$$

where

$\hat{y}$= predicted literacy rate

x1 = the number of daily newspaper copies in the country (per 1000 people)

x2 = the number of radios in the country (per 1000 people)

x3 = the number of TV sets in the country (per 1000 people)

# Case II – Literacy Rate Example

*Interpretation of the coefficients in the multiple linear regression equation*

- Let's start with the interpretation of the coefficient for newspaper copies (x1). Like the slope in simple linear regression, it tells us that we predict the literacy rate to increase by 0.00054 for every additional daily newspaper copy in that country (per 1000 people in the population).

- But, there is more. To properly interpret the coefficient of daily newspaper copies, the other two variables can't be changing – only the number of daily newspaper copies increases by 1. So, a way to interpret the coefficient of number of daily newspaper copies is as follows:

- For every additional daily newspaper copy per 1000 people in a population, literacy rate is predicted to increase by 0.00054, keeping the number of radios and TV sets the same.

# Case II – Literacy Rate Example

*Interpretation of the coefficients in the multiple linear regression equation*

- Since the coefficient is negative, we'd expect the literacy rate to be lower for every additional radio per 1000 people in the population (for countries with the same number of daily newspaper copies and TV sets per 1000 people in the population).

# Case II – Literacy Rate Example

Question 3: What are the degrees of freedom for the t* value in this problem?

- Recall, the degrees of freedom for any hypothesis test or confidence interval that involves a t-statistic is DFE = n – v – 1,
- where v = the number of explanatory variables in the model.
- In our problem, n = 10 and v = 3.
- Therefore, the degrees of freedom for the t* critical value is 10 – 3- 1 = 6.

# Case II – Literacy Rate Example

Question 4: Determine the lower and upper bounds for the 95% confidence interval for .β3

**Formula for confidence interval for a coefficient ($\beta_i$):**

$$b_i \pm (t^*_{n-v-1})(SE(b_i))$$

β-coeff  p-value  S.E

b3 = 0.00199, SE(b3) = 0.00155, and t = 2.447.

Therefore, the lower bound = (0.00199) – (2.447)(0.00155) = -0.00180.

The upper bound = (0.00199) + (2.447)(0.00155) = 0.00578.

0.00199

We write the 95% confidence interval for B3 as (-0.00180, 0.00578).

# Case II – Literacy Rate Example

Question 5: Determine the lower and upper bounds for the 95% confidence interval for $\beta 1$ and $\beta 2$.

# Case II – Literacy Rate Example

Question 6: Predict literacy rate for a country that has 200 daily newspaper copies (per 1000 in the population), 800 radios (per 1000 in the population), and 250 TV sets (per 1000 in the population).

$\hat{y} = 0.8436$

# Case II – Literacy Rate Example

Question 7: Verify that the F-statistic in the output above equals MSM / MSE.

MSM = 0.16132

and

MSE = 0.03477.

0.16132/0.03477 = 4.6396 or 4.64 rounded to two decimal places.

# Case II – Literacy Rate Example

Question 8: Conclusion on R square and adjusted R square value

→ ∴ $R^2$ = 69.9 %. We can say that our independent variables $x_1$, $x_2$, $x_3$ explains total 69.9 % of variation of our dependent variable i.e. Literacy rate.

# Case II – Literacy Rate Example

Question 9: What are the degrees of freedom for this F-statistic?

numerator df = DFM = # explanatory variables = 3.

denominator df = n – v – 1 = 10 – 3 – 1 = 6.

# Case II – Literacy Rate Example

Question 10: State a conclusion in the context of the problem.

There is suggestive, but weak, evidence to indicate that at least one of number of daily newspaper copies, number of radios, and/or number of TV sets help to explain a country's literacy rate (p-value = 0.053).

Some notes:

1) Even though the evidence is weak, we should continue the analysis to find out for sure if there is at least one explanatory variable that is a significant predictor of literacy rate and, if so, which one or ones. Anytime the p-value is less than 0.1 for the F-test, we should continue the analysis.

2) Remember, the conclusion states that there is suggestive evidence that at least one explanatory variable is a significant predictor of literacy rate. It does NOT tell us how many or which one or ones are significant predictors of literacy rate – only that there is at least one that is.

3) If the F-test indicates no evidence to reject the null hypothesis, then there is no need to continue the analysis as there is no evidence to indicate that any of the explanatory variables are helpful in explaining the response variable. However, if there is even the slightest bit of evidence to reject the null hypothesis from the F-test (i.e., p-value < 0.10), we should continue the analysis. This will involve doing t-tests on each explanatory variable, as we will see below.

# Case II – Literacy Rate Example

Question 11: Calculate the t-statistic (with degrees of freedom) for newspaper copies.

t =0.0005421 / 0.0008653 = 0.6265.

the degrees of freedom for the t-test is DFE = n – v – 1 = 6.

Calculate the t-statistic (with degrees of freedom) for radio.

Calculate the t-statistic (with degrees of freedom) for Tv sets.

# Case II – Literacy Rate Example

Question 12: What are the degrees of freedom for the t-tests in the final model?

Recall, the degrees of freedom for the t-test is DFE = n – v – 1. There are only 2 explanatory variables left in the model, so the degrees of freedom for the t-tests = 10 – 2 – 1 = 7.

# Case II – Literacy Rate Example

Question 13: Which of the following is true?

A] TV sets would remain in the model because we always need to have at least one explanatory variable in the model.

B] TV sets would remain in the model since its p-value is less than 0.05.

A is not correct because it is possible that a backwards selection process will eliminate all variables. But, remember that we'll stop eliminating variables once all remaining variables have p-values less than 0.05, which is the case here. Therefore, C is also incorrect.

# Case III – The Box Office Collection

The result of regression analysis done by one of the MBA students groups in the sports and entertainment economics class is reported in the table below. The dependent variable is the gross revenue from 190 movie show in theatres in the US.

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.291252677 |
| R Square | 0.084828122 |
| Adjusted R Square | 0.07014622 |
| Standard Error | 50181924.73 |
| Observations | 191 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 3 | 4.36489E+16 | 1.45496E+16 | 5.777734 | 0.000845912 |
| Residual | 187 | 4.70908E+17 | 2.51823E+15 | | |
| Total | 190 | 5.14557E+17 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* |
|---|---|---|---|---|---|
| Intercept | 12846788.93 | 24363241.93 | 0.527302112 | 0.598609 | -35215334.51 |
| Ticket_Prices | 7483427.727 | 3206667.474 | 2.333708683 | 0.020674 | 1157535.286 |
| Star_Power | 26121958.92 | 7605710.943 | 3.434519024 | 0.000731 | 11117936.92 |
| Movie_Rating | -9330073.57 | 7453336.119 | -1.251798311 | 0.212207 | -24033501.03 |

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

# Case III – The Box Office Collection

Question: What is the response variable? What are the explanatory variables?

Question: Write the least-squares regression equation for this problem.

Question: Interpretation of the coefficients in the multiple linear regression equation

Question: What are the degrees of freedom for the t* value in this problem? $n-v-1 = 190-3-1 = 186$

Question: Determine the lower and upper bounds for the 95% confidence interval for β1 β2 β3 → $\beta_i \pm t^* SE(\beta_i)$

Question: Predict gross review of movie with 250$ ticket price, 4 star movie rating and star power of 9.

Question: Conclusion on R square and adjusted R square value

Question: State a conclusion in the context of the problem.

# Case III – The Interest Rate

100 variables → $x_1, x_2, x_3 \dots x_{100}$ variables

Selection Method :-

→ start Method ✓

| $y_1$ | $x_{100}$ |
|---|---|

- Forward selection
- Backward Elimination → $y = \beta_0 + \beta_1 x_1 + \dots + \beta_{100} x_{100}$ → $R^2$ $\bar{R}^2$
- Mixed selection ⇌

$R^2$ $\bar{R}^2$

# Case IV – Housing Price

Below is the summary of the multiple regression analysis for estimation of housing price based on various parameters:

SUMMARY OUTPUT

| Regression Statistics | |
|---|---|
| Multiple R | 0.831897 |
| R Square | 0.692052 |
| Adjusted R Square | 0.66498 |
| Standard Error | 180450.3 |
| Observations | 100 |

ANOVA

| | df | SS | MS | F | gnificance F |
|---|---|---|---|---|---|
| Regression | 8 | 6.66E+12 | 8.32E+11 | 25.5631 | 3.24E-20 |
| Residual | 91 | 2.96E+12 | 3.26E+10 | | |
| Total | 99 | 9.62E+12 | | | |

| | Coefficients | tandard Err | t Stat | P-value | Lower 95% | Upper 95% | ower 95.0% | pper 95.0% |
|---|---|---|---|---|---|---|---|---|
| Intercept | 4201625 | 2611155 | 1.609106 | 0.111057 | -985113 | 9388363 | -985113 | 9388363 |
| sqft_living | 179.7723 | 28.43326 | 6.322606 | 9.4E-09 | 123.2931 | 236.2515 | 123.2931 | 236.2515 |
| bedrooms | 22359.91 | 27512.84 | 0.812708 | 0.418506 | -32291 | 77010.78 | -32291 | 77010.78 |
| bathrooms | -94113.1 | 37866.29 | -2.48541 | 0.014767 | -169330 | -18896.3 | -169330 | -18896.3 |
| waterfront | 340047.7 | 95826.77 | 3.548567 | 0.000615 | 149699.6 | 530395.8 | 149699.6 | 530395.8 |
| view | 257387.9 | 57917.54 | 4.44404 | 2.48E-05 | 142341.8 | 372433.9 | 142341.8 | 372433.9 |
| condition | -22249.7 | 25621.51 | -0.8684 | 0.38746 | -73143.7 | 28644.28 | -73143.7 | 28644.28 |
| grade | 79672.97 | 26266.51 | 3.033253 | 0.003153 | 27497.78 | 131848.2 | 27497.78 | 131848.2 |
| yr_built | -2251.09 | 1290.687 | -1.7441 | 0.084518 | -4814.88 | 312.7037 | -4814.88 | 312.7037 |

i. Interpret the results of regression analysis
ii. Develop the regression equation

$\hat{y} = 4201625 + 179.77 x_1 + \ldots - 2251.09 x_8$

$R^2 \rightarrow 0.69$ i.e. 69% of the variation in the housing price is explained by all the ind. variable

At 5% LOS, if the model is useful for predictions

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_8 = 0$

$H_1$ : Not $H_0$

ANOVA Table

$F = 25.5231$, p-value = 0.0005

At 5% LOS, I enough evidence to conclude that the predictor is useful for predicting housing price

at least one of