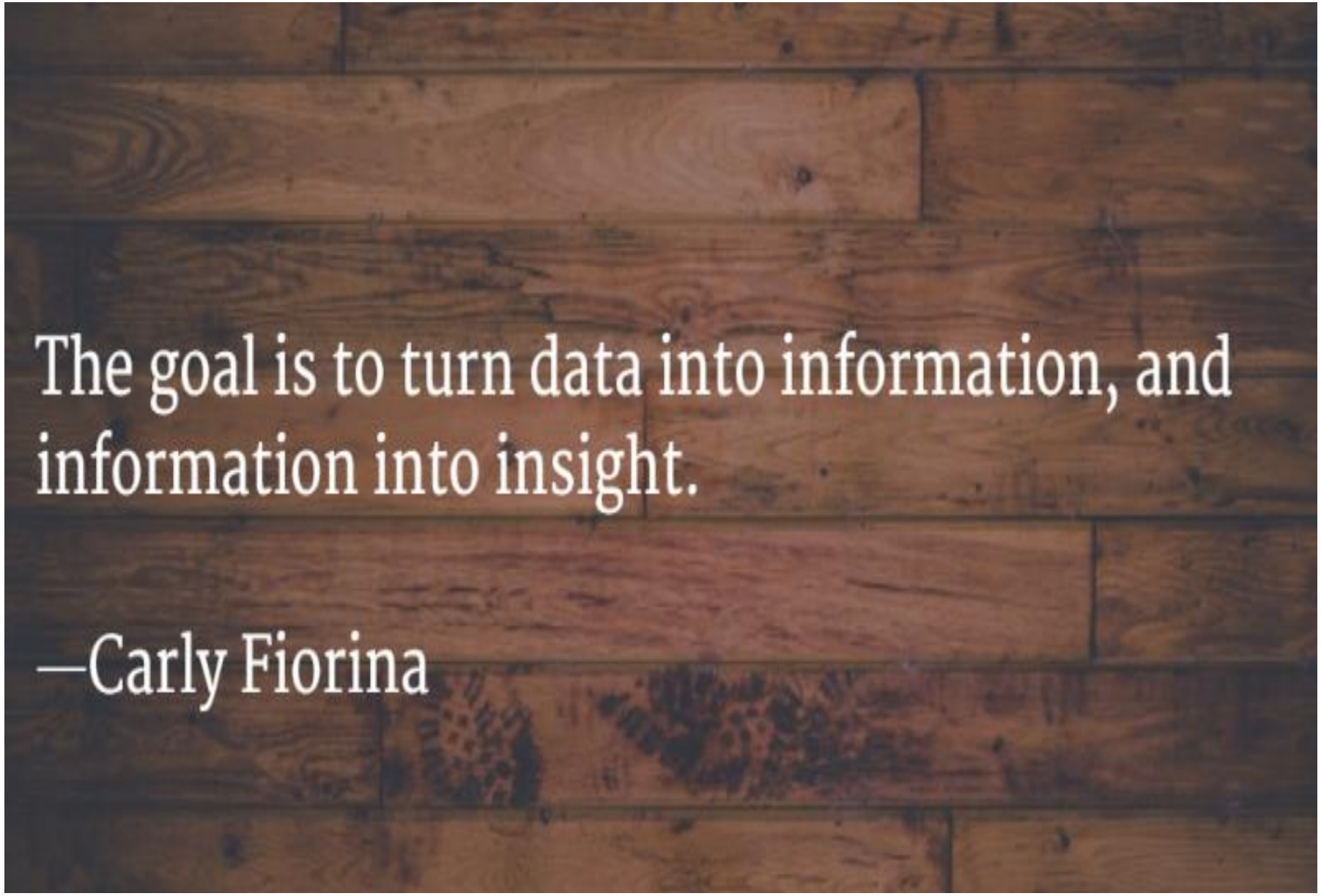
Quantitative Analytics tools for financial decisions

Linear Regression, Regression with Multiple Explanatory Variables

Module 2 Session 3 & 4

विद्या परं दैवतम्
IIM
VISAKHAPATNAM


The goal is to turn data into information, and information into insight.
—Carly Fiorina

Agenda for Last Session

Hypothesis Testing

विद्या परं दैवतम्
IIM
VISAKHAPATNAM

Today's Agenda

Simple Regression

Multivariate Regression

विद्या परं दैवतम्
IIM
VISAKHAPATNAM

Question

Suppose you start up a company that has developed a drug that is supposed to increase IQ. You know that the standard deviation of IQ in the general population is 15. You test your drug on 36 patients and obtain a mean IQ of 97.65. Using an alpha value of 0.05, is this IQ significantly different than the population mean of 100?

Question

$$z = \frac{97.65 - 100}{2.5}$$

$$= -0.94$$

Level of Significance =0.05, two tailed , 0.05/2 =0.025, Z value = -1., Since calculated value is less than tab null accepted.

विद्या परं देवतम्
IIM
VISAKHAPATNAM

Question Normal Distribution

A companies share price is normally distributed with a mean weight of Rs 800 and a standard deviation of Rs 300 . A random sample of 16 days share price is taken. (a) What is the probability that the share price of the sample exceeds Rs 900 ?

μ= 800, $\bar{x}$ = 900 , S.E = 300/$\sqrt{16}$ =75

P($\bar{x}$> 900) = $\frac{900-800}{75}$ = 1.33

0.5-0.4082 = 0.0918 , Hence there is 9.18% probability.

Question Normal Distribution

A companies share price is normally distributed with a mean weight of Rs 800 and a standard deviation of Rs 300 . A random sample of 16 days share price is taken.
(a) What is the probability that the share price of the sample exceeds Rs 900 ?

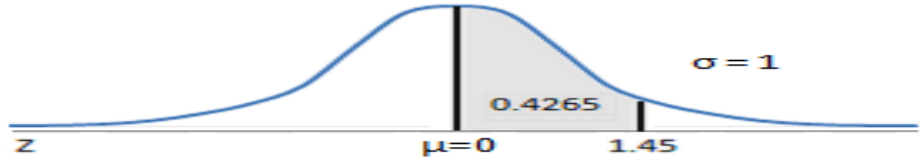
μ= 800, $\bar{x}$ = 900 , S.E = 300/$\sqrt{16}$ =75

P($\bar{x}$> 900) = $\frac{900-800}{75}$ = 1.33

0.5-0.4082 = 0.0918 , Hence there is 9.18% probability.

Areas Under the One-Tailed Standard Normal Curve

This table provides the area between the mean and some Z score. For example, when Z score = 1.45 the area = 0.4265.



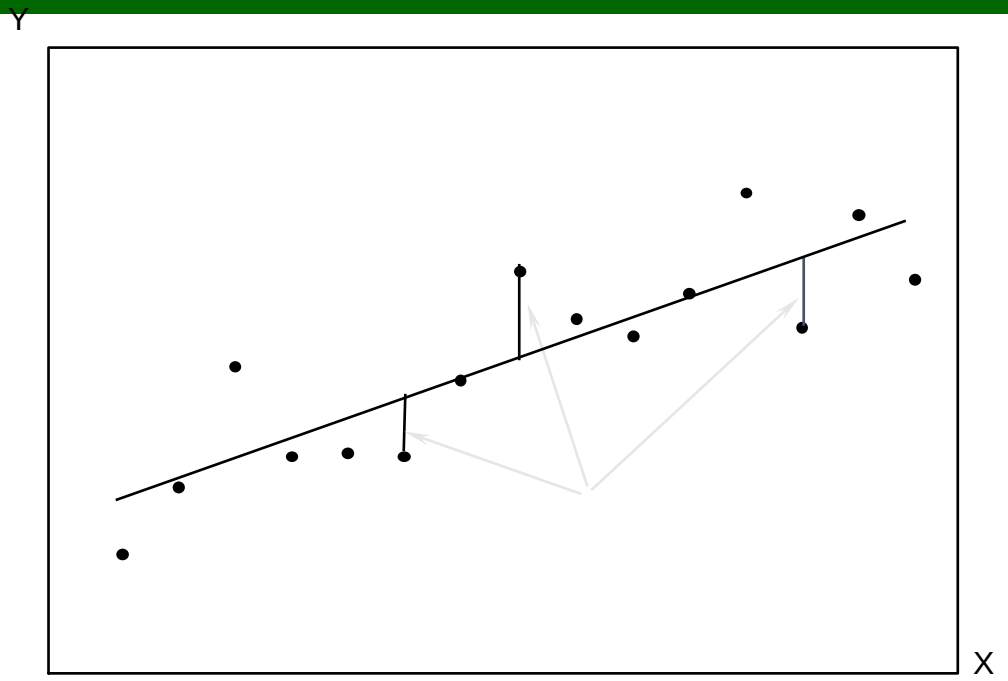
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Regression Analysis

The statistical technique that expresses the relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable, is called regression analysis.

Linear Regression model

Objective: The line that **BEST** fit the data.

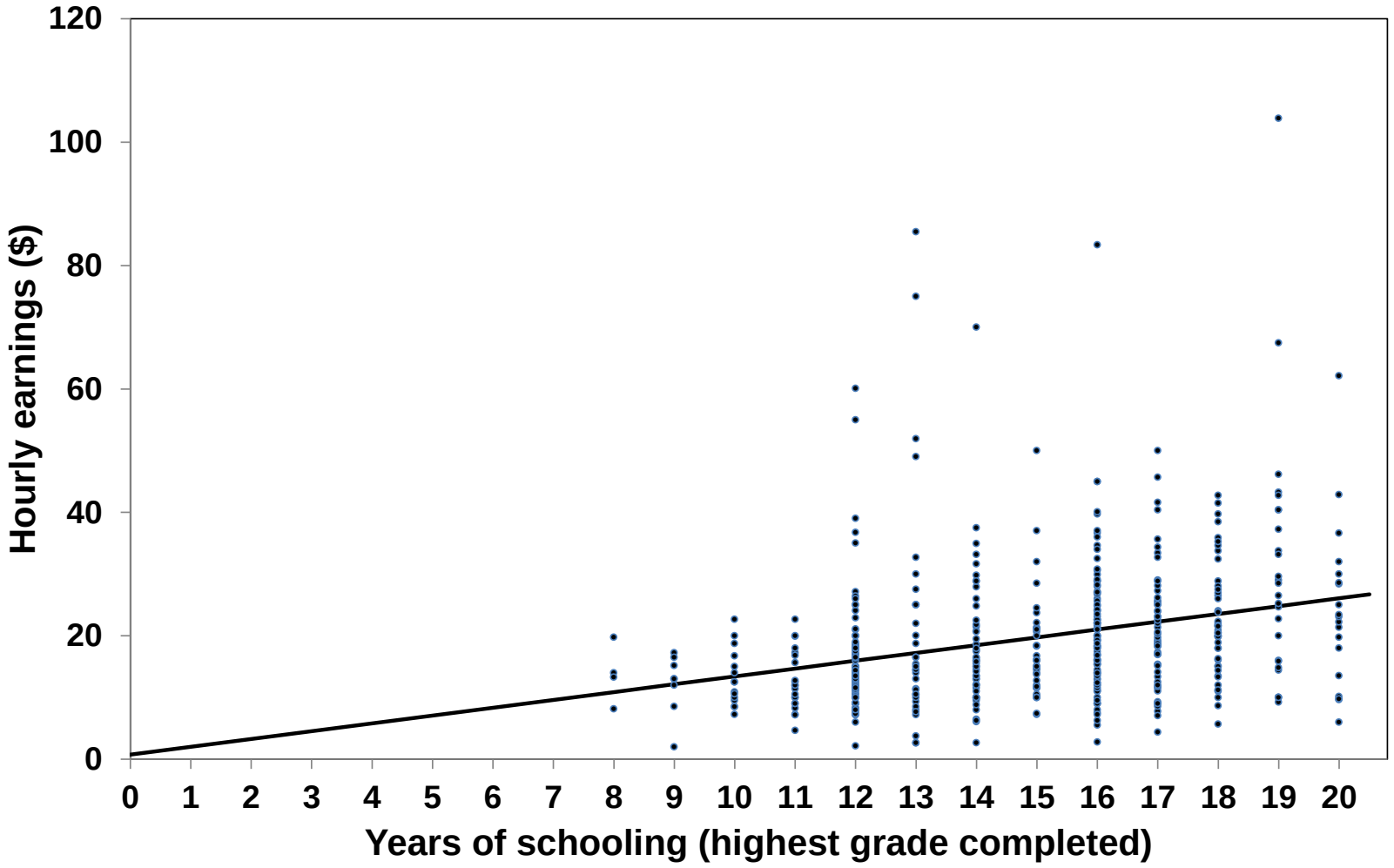


Minimize the sum of difference between actual and fitted values?

Or

Minimize the sum of squares of difference between the actual and fitted value?

Regression Analysis



विद्या परं देवतम्
IIM
VISAKHAPATNAM

Regression Analysis

- **A time series is a set of observations *x* , *each one being recorded at a specific time t* .**
- Time series data, as the name suggests, are data that have been collected over a period of time on one or more variables.

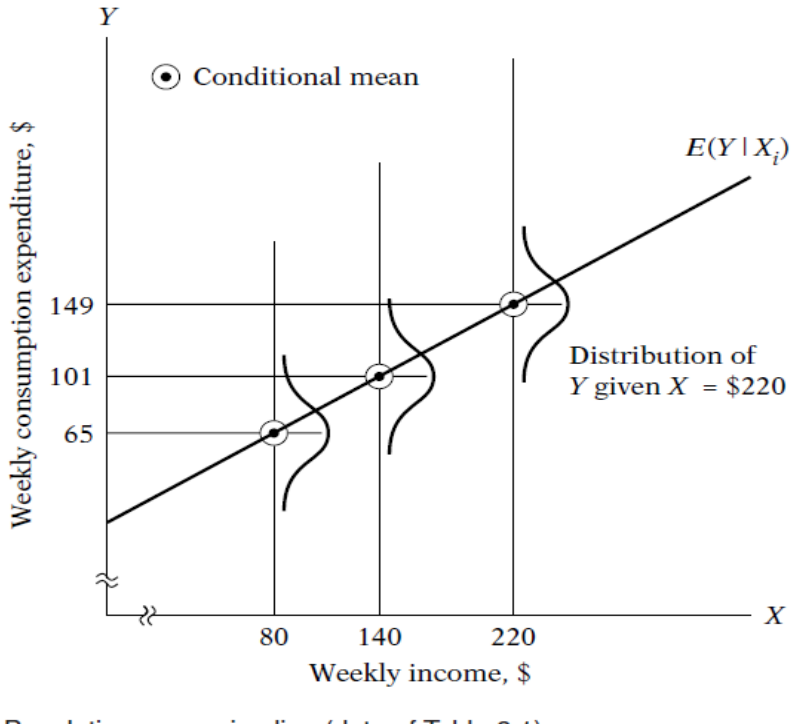
- **Problems that can solved with Time-Series:**
  - How the value of a country's stock index has varied with that country's macroeconomic fundamentals.
  - How the value of a company's stock price has varied when it announced the value of its dividend payment.
  - The effect on a country's exchange rate with increase in its trade deficit.
  - How interest rate are determined.
  - Finding out risk in an asset class.

Regression Analysis

Dependent variable	Explanatory variable
⇕	⇕
Explained variable	Independent variable
⇕	⇕
Predictand	Predictor
⇕	⇕
Regressand	**Regressor**
⇕	⇕
Response	Stimulus
⇕	⇕
Endogenous	Exogenous
⇕	⇕
Outcome	Covariate
⇕	⇕
Controlled variable	Control variable

विद्या परं देवतम्
IIM
VISAKHAPATNAM

Regression Analysis



Regression Analysis

You are fitting a below mentioned straight -line equation.

$$y_i = \beta_1 + \beta_2 x_2$$

where $\beta 1$ and $\beta 2$ are unknown but fixed parameters known as the **regression coefficients.**

In regression analysis our interest is in estimating the PRFs**.**

विद्या परं देवतम्
IIM
VISAKHAPATNAM

Regression Analysis

Assumptions of Linear Regression Model

1. **Linear regression model.** The regression model is **linear in the parameters,**

   ***X* values are fixed in repeated sampling.** Values taken by the regressor *X* are considered fixed in repeated samples. More technically, *X* is assumed to be *non stochastic.*

2. **Homoscedasticity or equal variance of *ui.***

3. **No autocorrelation between the disturbances.**

4. ***4.* Zero covariance between *ui* and *Xi*,**

5. ***5.* The number of observations *n* must be greater than the number of parameters to be estimated.**

Regression Analysis

- **PRECISION OR STANDARD ERRORS**

- It is evident that least-squares estimates are a function of the sample data. But since the data are likely to change from sample to sample, the estimates will change ipso facto.

- In statistics the precision of an estimate is measured by its standard error.

Standard error of estimate or **the standard error of the regression (se).**

- *It is simply the standard deviation of the Y values about the estimated regression line and is often used as a summary measure of the "goodness of fit" of the estimated regression line.*

Regression Analysis

PROPERTIES OF LEAST-SQUARES ESTIMATORS

- It is **linear,** that is, a linear function of a random variable, such as the dependent variable *Y* in the regression model**.**
- It is **unbiased,** that is, its average or expected value, $E(\hat{\beta}2)$, is equal to the true value.
- It has minimum variance in the class of all such linear unbiased estimators; an unbiased estimator with the least variance is known as an **efficient estimator.**

विद्या परं देवतम्
IIM
VISAKHAPATNAM

Regression Analysis

THE COEFFICIENT OF DETERMINATION- R2

A MEASURE OF "GOODNESS OF FIT"

We now consider the **goodness of fit** of the fitted regression line to a set of data; that is, we shall find out how "well" the sample regression line fits the data.

The **coefficient of determination** *r* 2 (two-variable case) or *R*2 (multiple regression) is a summary measure that tells how well the sample regression line fits the data.

$$R^2 = \frac{ESS}{TSS}$$

Where ESS = Explained sum of square
TSS = Total sum of square

Regression Analysis

Output Interpretation

Regression Statistics										
Multiple R	0.983559									
R Square	0.967389									
Adjusted R Square	0.964672									
Standard Error	19.00868									
Observations	14									
ANOVA										
	df	SS	MS	F	Significance F					
Regression	1	128625.3	128625.3	355.9772	2.75E-10					
Residual	12	4335.96	361.33							
Total	13	132961.2								
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%		
Intercept	-22.5464	10.43676	-2.16029	0.051685	-45.2861	0.193368	-45.2861	0.193368		
X Variable 1	3.269721	0.1733	18.86736	2.75E-10	2.892132	3.64731	2.892132	3.64731		

विद्या परं देवतम्
IIM
VISAKHAPATNAM

Regression Analysis

Null Hypothesis of Regression Co-efficient H_0 = All coefficients =0

Significance testing…

H0: $\beta 1 = 0$ (no linear relationship)

H1: $\beta 1 \neq 0$ (linear relationship does exist)

$$T_{n-2} = \frac{\hat{\beta} - 0}{s.e.(\hat{\beta})}$$

Residual Analysis: check assumptions

$$e_i = Y_i - \hat{Y}_i$$

The residual for observation i, e_i, is the difference between its observed and predicted value

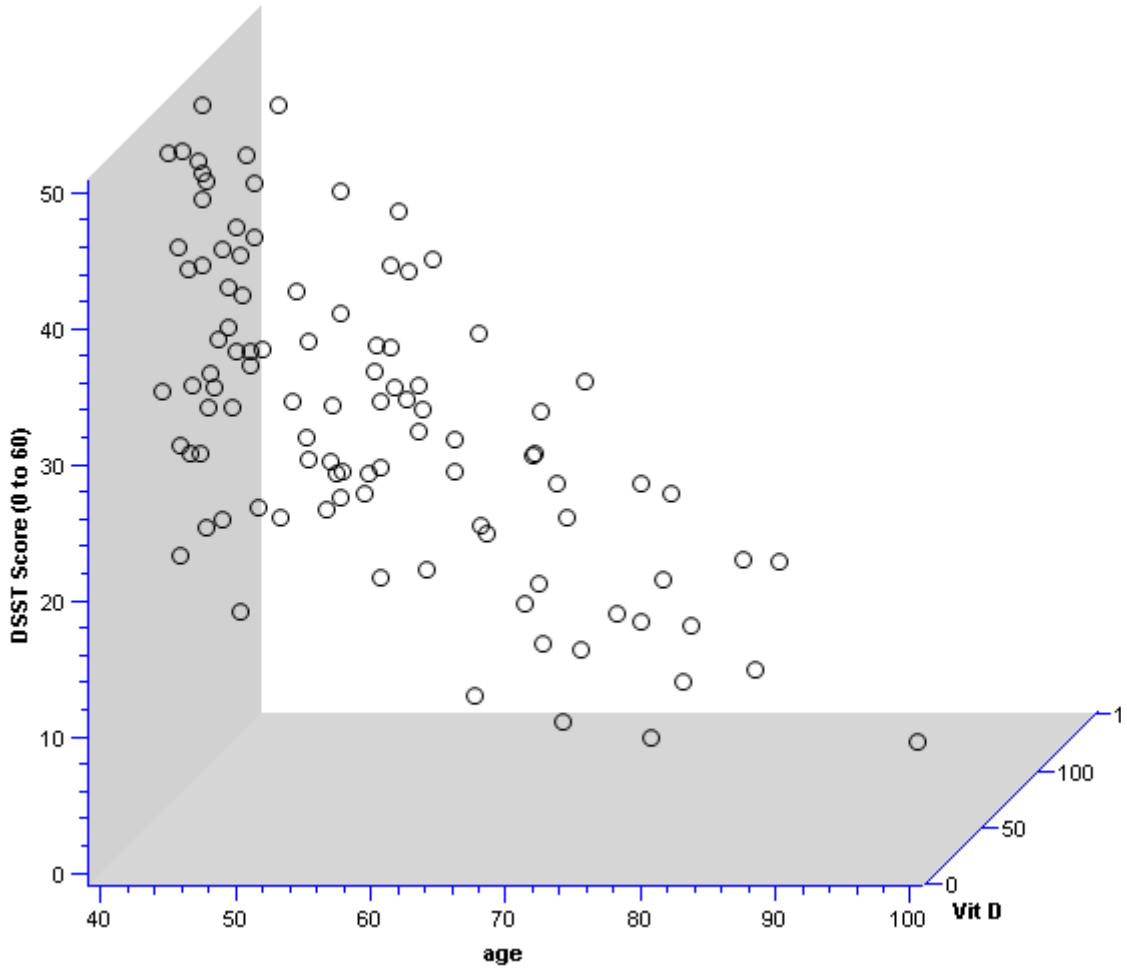
Check the assumptions of regression by examining the residuals

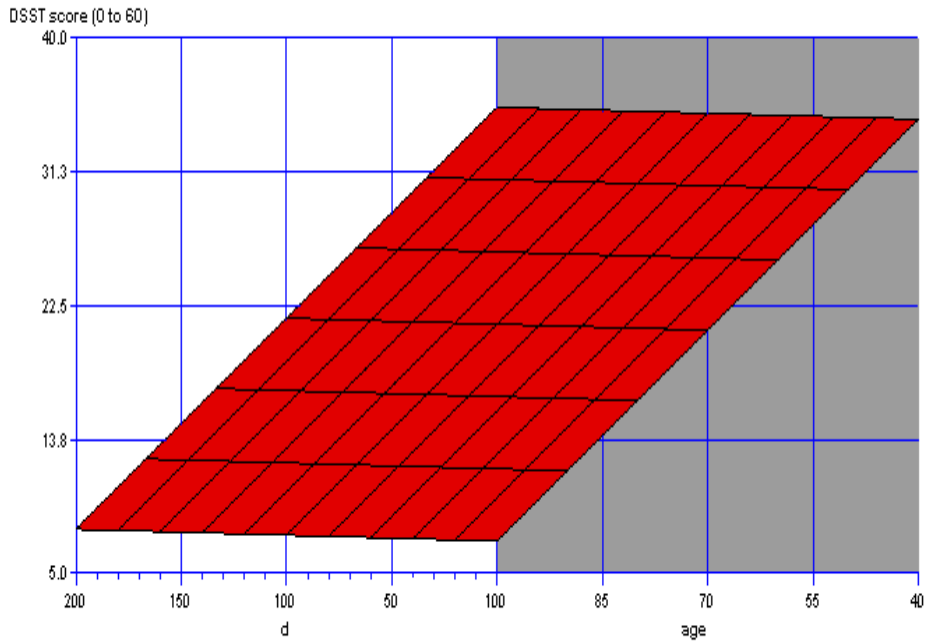
- Examine for linearity assumption
- Examine for constant variance for all levels of X (homoscedasticity)
- Evaluate normal distribution assumption
- Evaluate independence assumption

Graphical Analysis of Residuals

- Can plot residuals vs. X

Multiple linear regression (We fit a Plane)





विद्या परं देवतम्
IIM
VISAKHAPATNAM

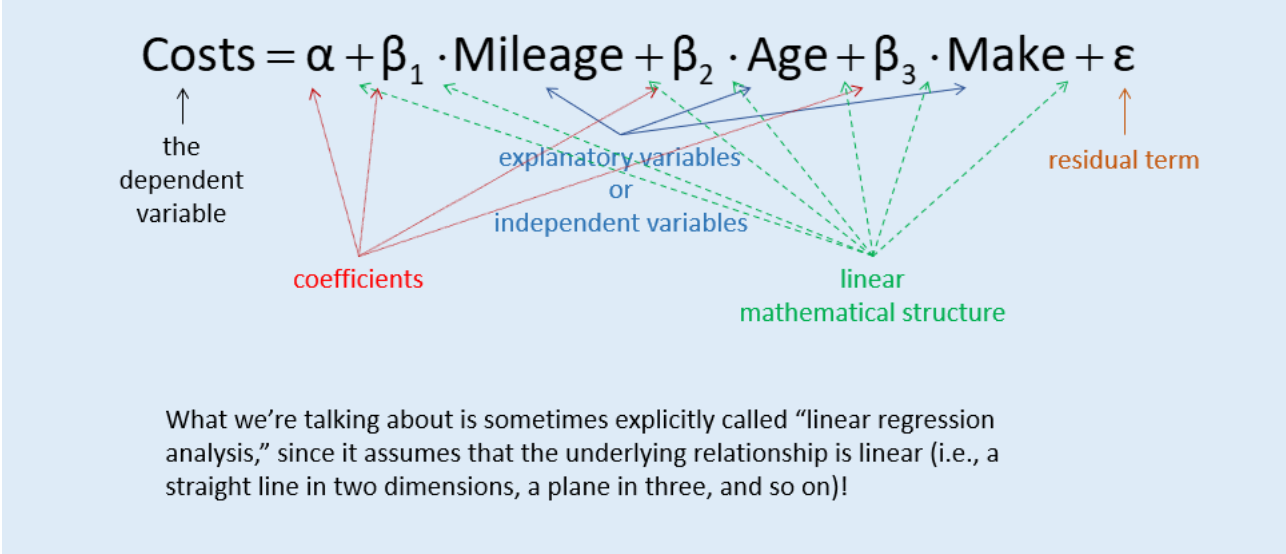
Functions of multivariate analysis:

- Control for confounders
- Test for interactions between predictors (effect modification)
- Improve predictions

विद्या परं देवतम्
IIM
VISAKHAPATNAM

The Regression Model

$$\text{Costs} = \alpha + \beta_1 \cdot \text{Mileage} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Make} + \varepsilon$$

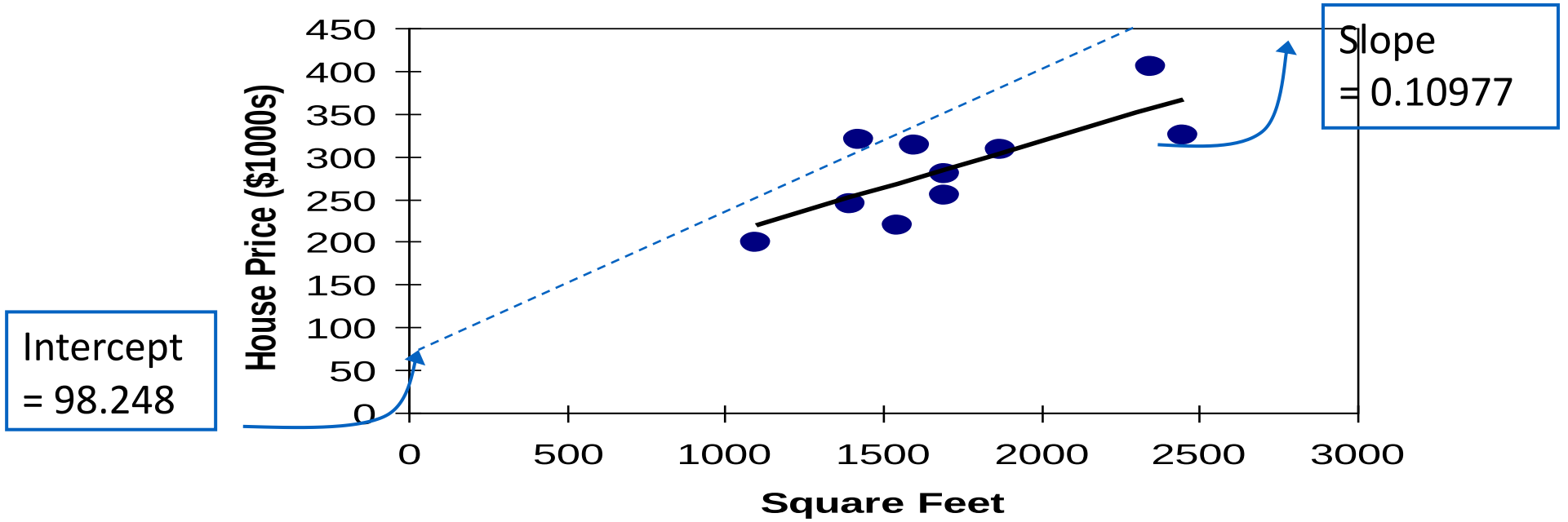


What we're talking about is sometimes explicitly called "linear regression analysis," since it assumes that the underlying relationship is linear (i.e., a straight line in two dimensions, a plane in three, and so on)!

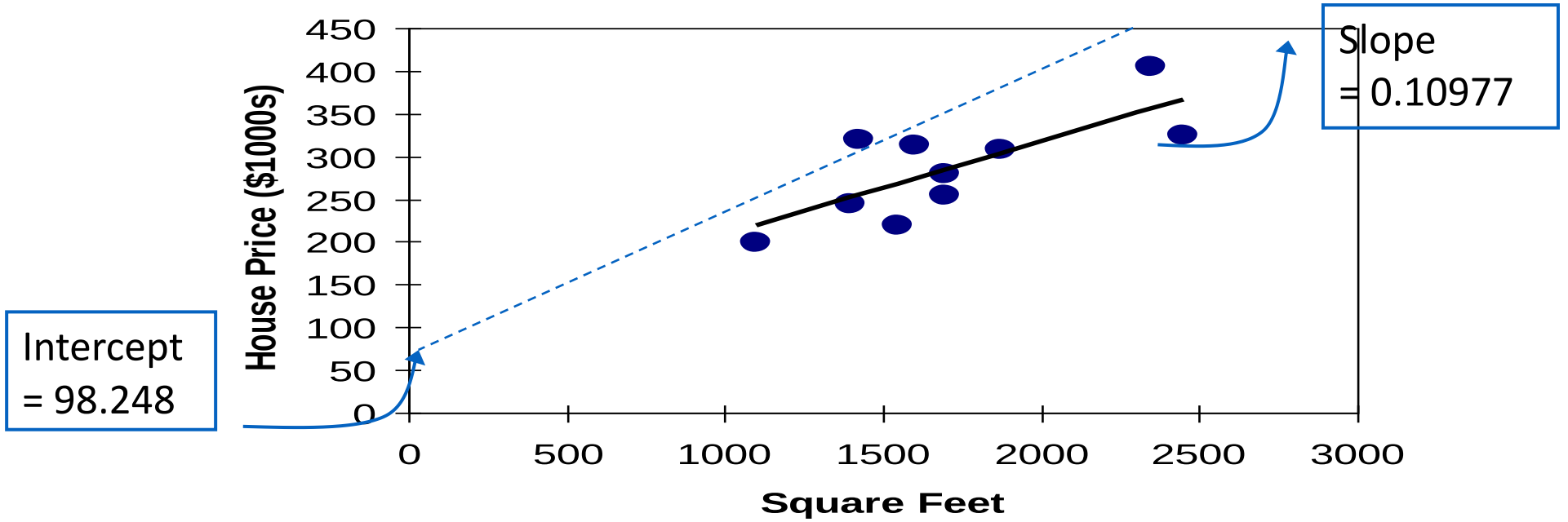
विद्या परं देवतम्
IIM
VISAKHAPATNAM

The Regression Model

- House price model: scatter plot and regression line



- House price model: scatter plot and regression line



Interpretation of the Intercept, b_0

$$\text{Price} = 98.24833 + 0.10977\ (\text{sales})$$

b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)

Excel Output

Regression Statistics	
Multiple R	**0.76211**
R Square	**0.58082**
Adjusted R Square	**0.52842**
Standard Error	**41.33032**
Observations	**10**

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**1**	**18934.9348**	**18934.9348**	**11.0848**	**0.01039**
Residual	**8**	**13665.5652**	**1708.1957**		
Total	**9**	**32600.5000**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

विद्या परं देवतम्
IIM
VISAKHAPATNAM

Standard Error of Estimate

The standard deviation of the variation of observations around the regression line is estimated by

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

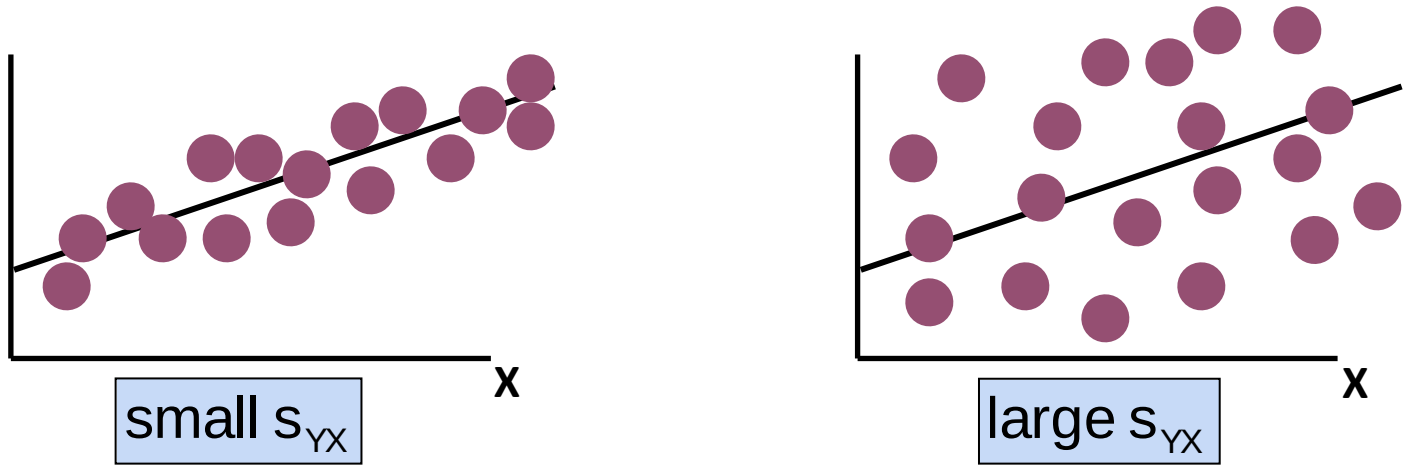
Where

SSE = error sum of squares

n = sample size

Comparing Standard Errors

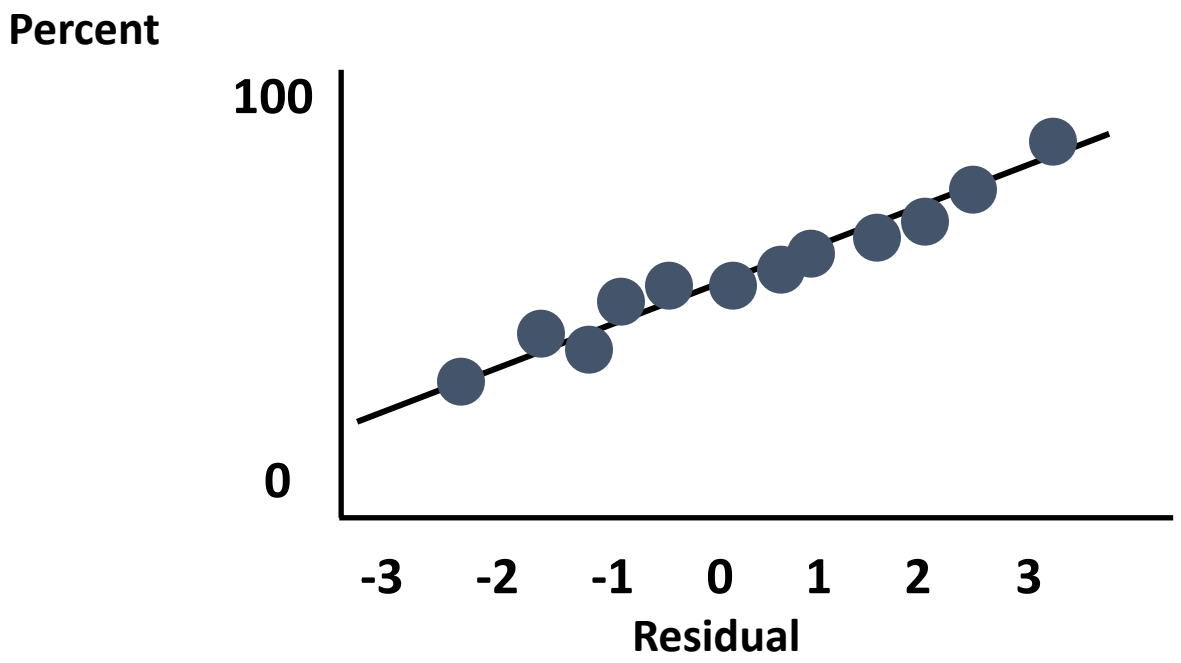
S_{YX} is a measure of the variation of observed Y values from the regression line



The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data

Residual Analysis for Normality

A normal probability plot of the residuals can be used to check for normality:



Regression Estimator

Regression estimator should be BLUE

B= Best
L= Linear
U= Unbiased
E= Estimator