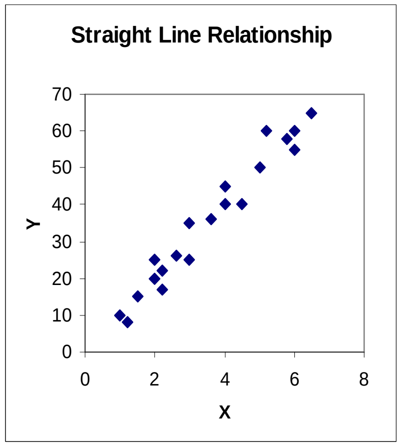
Module 3: Models for Data

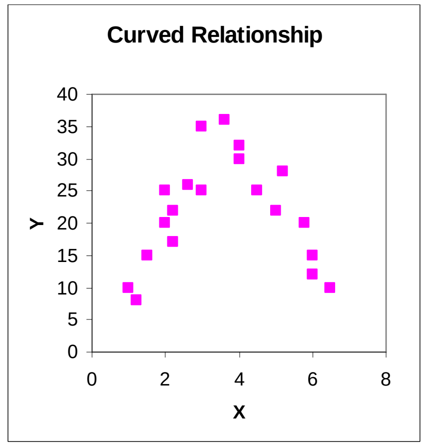
3.2 Regression Analysis

Regression Analysis

- Regression analysis is useful when you are looking at the relationship between two or more variables.
- It's useful for trend analysis.
- If two variables are involved, the relationship could be a straight line or could contain various types of curvature.

Regression Analysis





Regression Analysis

- If three variables are involved, the lines become planes or curved surfaces.
- If the relationship involves multiple variables, the surfaces are multi-dimensional.

Regression Analysis

Data: Predictor Variables Response

x_{11}	x_{21}	...	x_{p1}	y_1
x_{12}	x_{22}	…	x_{p2}	y_2
.	.	.	.	
.	.	.	.	
.	.	.	.	
x_{1n}	x_{2n}	…	x_{pn}	y_n

Regression Analysis

Model:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

where ε is Normal (μ,σ)

Once the model coefficients are estimated, the model can be used to calculate a predicted y_i for any set of x_i's.

Regression Analysis

- The regression coefficients (Betas) are estimated using least squares.
- Least squares minimizes the sum of the squared differences between the data values and their predicted values
- These differences are prediction errors
- So, we calculate estimates of the Betas that minimize the sum of the squared errors

Module 3.2

Regression Analysis

SSE = $$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

There are other important sums of squares

The Total Sum of Squares (SST) = $$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

The Regression Sum of Squares (SSR) can be easily calculated by subtraction SSR = SST - SSE

Module 3.2

Regression Analysis

- R^2 is the coefficient of determination. It is the proportion of the variation in the Y variable that is accounted for by the regression model.
- R^2 =SSR/SST = 1 - SSE/SST
- It is a measure of the usefulness of the regression model

Regression Analysis

- Mean Squares are calculated by dividing Sums of Squares by the appropriate degrees of freedom.
- A Mean Square is a measure of a variance
- If you divide a Mean Square by another, the resulting statistic has an F distribution with the degrees of freedom of the df of the numerator and the df of the denominator
- You can use these F statistics to test for significance. In this case, for the significance of the regression model.

Regression Analysis

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Regression	SSR	p	MSR	MSR/MSE
Error	SSE	n-p-1	MSE	
Total	SST	n-1		

To carry out a a test at 95% confidence, compare the calculated F ratio against a value from a table of the F distribution with p and n-p-1 degrees of freedom (Table A2.4 in Manly).

Regression Analysis

- If regressing on more than one variable, each variable must be tested for significance in addition to testing the significance of the overall model
- If the estimate of a Beta coefficient is not significantly different from zero, the variable should not be included in the model.

Regression Analysis

- There are two ways to build a model with multiple variables: forward stepwise and backward stepwise.
- Many statistical packages incorporate these.
- Forward adds one variable at-a-time, testing each one for significance as it is added.
- Backward stepwise starts with all of the variables in the equation and drops the nonsignificant ones out one-at-a-time.

Module 3.2

Regression Analysis

- To test if a coefficient is significant, the estimate is compared to its standard error (recall that all statistics have standard errors).
- The ratio $b_j/SE(b_j)$ has a t distribution with n-p-1 degrees of freedom
- So, to determine significance compare the calculated statistic against the value from a table of the t distribution

Module 3.2

Regression Analysis

- Residual analysis should always be done after fitting a regression equation to check to see if the model form is adequate.
- An example of model inadequacy would be fitting a straight line to a curved relationship.
- The residuals would show the curvature

Module 3.2

Example Regression Analysis

X	Y	X	Y
1	10	3.6	75
1.2	28	4	95
1.5	15	4	60
2	45	4.5	72
2	35	5	87
2.2	20	5.2	69
2.2	60	5.8	48
2.6	81	6	57
3	45	6	70
3	69	6.5	55

Module 3.2

Example Regression Analysis

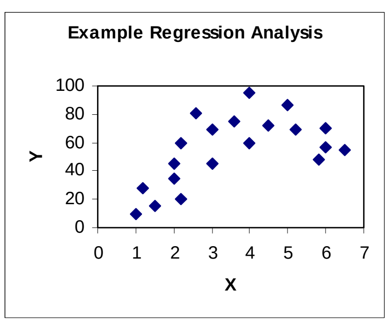
$Y = B_0 + B_1X$

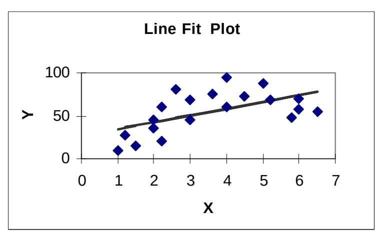
ANOVA	***SS***	***df***	***MS***	***F***	***Significance of F***
Regression	**3535.4**	**1**	**3535.4**	**8.7**	**0.0087**
Residual	**7351.8**	**18**	**408.4**		
Total	**10887.2**	**19**			

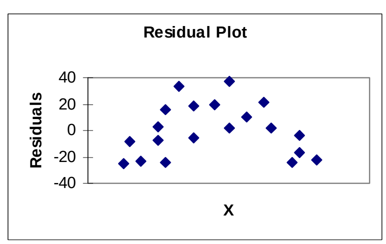
	Coefficients	***Standard Error***	***t Stat***	***P-value***
Intercept	**26.99**	**10.48**	**2.58**	**0.02**
X Variable	**7.80**	**2.65**	**2.94**	**0.01**

$R^2 = 0.32$

Example Regression Analysis







Conclusion: Not the right model for this data!

Example Regression Analysis

- So, what to do?
- Looks like it needs a squared term to make it into a quadratic
- $Y = B_0 + B_1X + B_2X^2$
- To do this, create a column in Excel that squares X and regress against both variables

Example Regression Analysis

- $Y = B_0 + B_1X + B_2X^2$

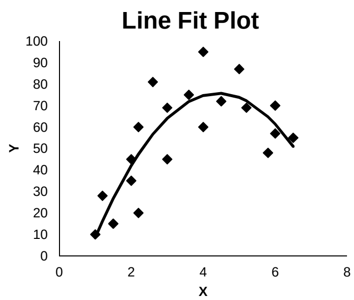
ANOVA	*SS*	*df*	*MS*	*F*	*Significance of F*
Regression	**7429.2**	**2**	**3714.6**	**18.3**	**0.00006**
Residual	**3458.0**	**17**	**203.4**		
Total	**10887.2**	**19**			

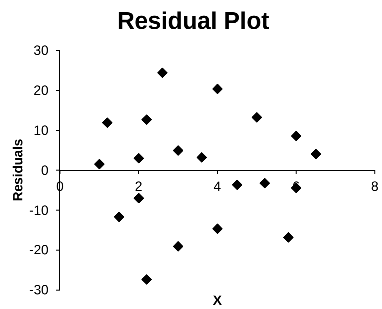
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**-36.54**	**16.29**	**-2.24**	**0.04**
X Variable	**50.74**	**9.99**	**5.08**	**0.00009**
X^2 Variable	**-5.74**	**1.31**	**-4.38**	**0.0004**

$R^2 = 0.68$

Example Regression Analysis

Much Better!





Module 3.2

21