Statistical Data Analysis for Post-Graduate Students Using R Programming Language.

Strengthening Research skills in Eastern and Southern Africa

Makerere University Regional Centre for Crop Improvement
Research and Training for Africa
MaRCCI

RUFORUM
Capacity Building in Agriculture

Makerere University

Correlation and Regression with R

Assoc. Prof. Susan Balaba Tumwebaze

Dr. Thomas Odong

Dr.Hellen Namawejje

Correlation

- Correlation is a measure of degree of linear association between two variables, varies between -1 to 1

Types of linear correlation

- **Positive Linear Correlation**
- General trend in the plotted points is from **bottom left to top right.**
- **Negative Linear Correlation**
- General trend in the plotted points is from **top left to bottom right.**
- **No Linear Correlation**
- No general trend in plotted points or a non-linear trend.

Correlation Coefficient

- Is a method for determining the type and strength of a linear correlation. Correlation Coefficient is denoted by *r*

Correlation Coefficient

- The correlation coefficient will always be a number between -1 and 1.
- A positive value indicates a positive correlation and a negative value a negative correlation.

Correlation Coefficient

- A coefficient of r=1 for a data set indicates perfect positive linear correlation,
- and r=-1 indicates perfect negative linear correlation,
- while r=0 would indicate no linear correlation.

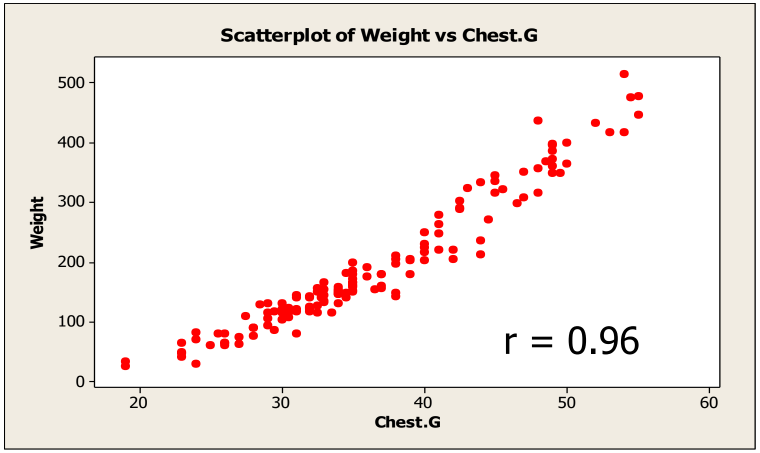
Scatterplots and Correlation

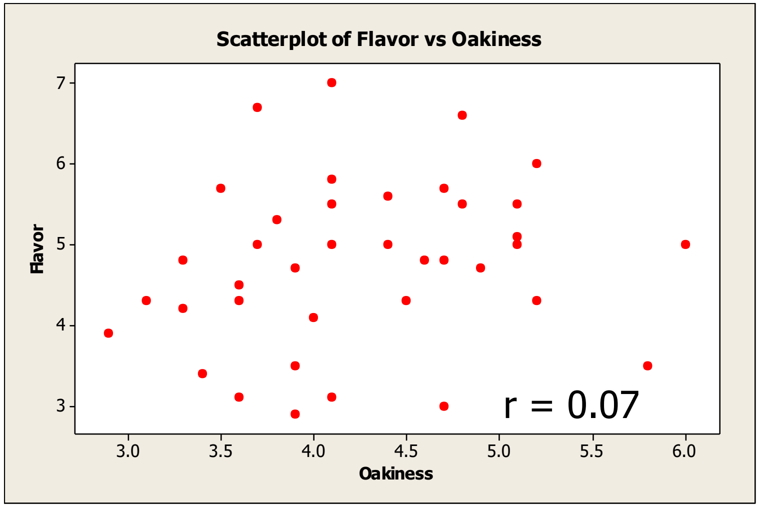
- To quantify the **strength and direction** (positive or negative) of the linear relationship between two variables we use Pearson’s Linear Correlation Coefficient (*r*)
- You will never have to compute this by hand!!!

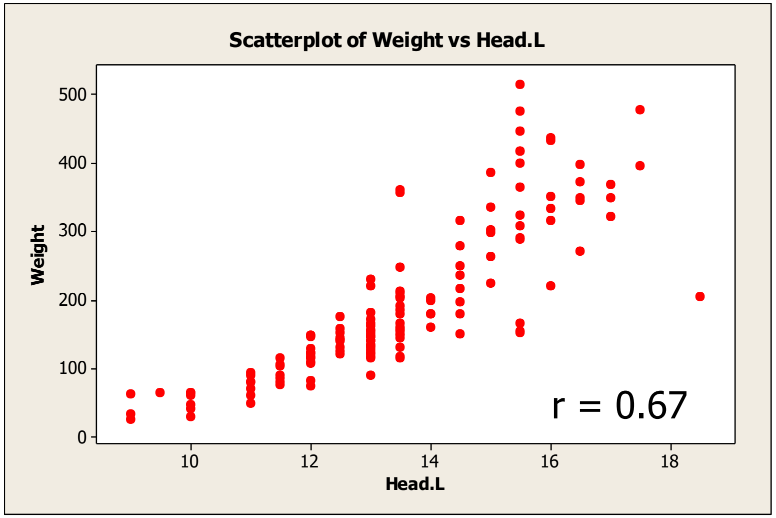
$$r = \frac{\sum \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}}{n - 1}$$

Scatterplots and Correlation

- The closer *r* is to +1, the stronger the positive relationship between the two variables

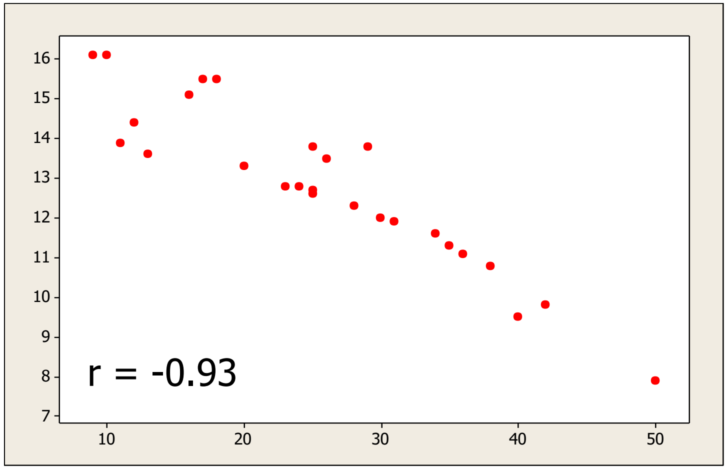


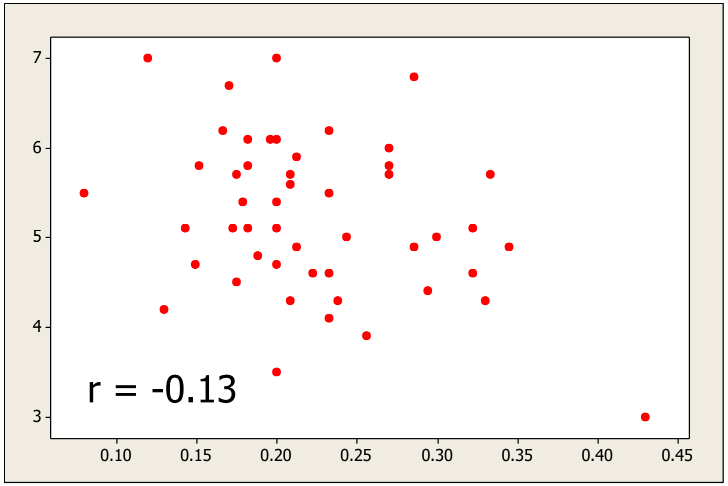


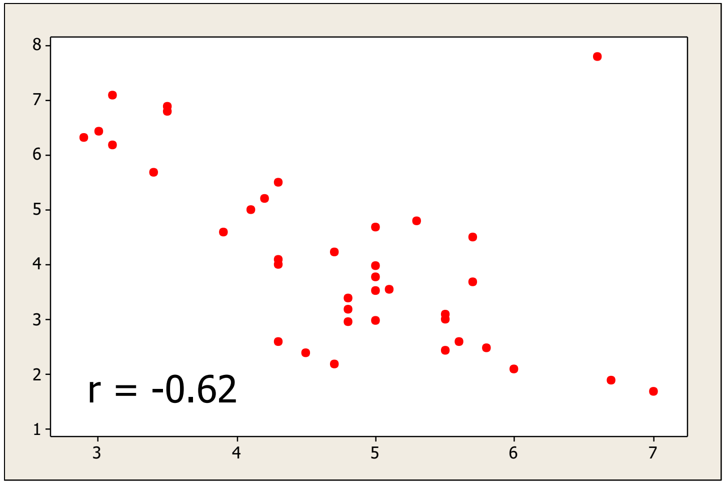


Scatterplots and Correlation

- The closer *r* is to -1, the stronger the negative relationship between the two variables







Scatterplots and Correlation

- Correlation is not causation
- Just because two variables are correlated does not mean that one causes the other to change
- Religion causes crime???
- Pearson’s correlation coefficient measures strength and direction of LINEAR relationships, not causal relationships

Correlation coefficient in R

- cor() computes the correlation coefficient
- cor.test() test for association between paired samples
- cor.test() returns both the correlation coefficient and p-value of the correlation

Correlation coefficient in R

Detailed commands

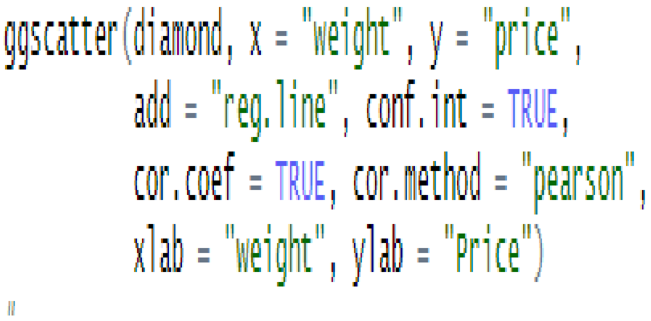
- cor (x, y, method = c("pearson", "kendall", "spearman"))

- cor.test(x, y, method=c("pearson", "kendall", "spearman"))

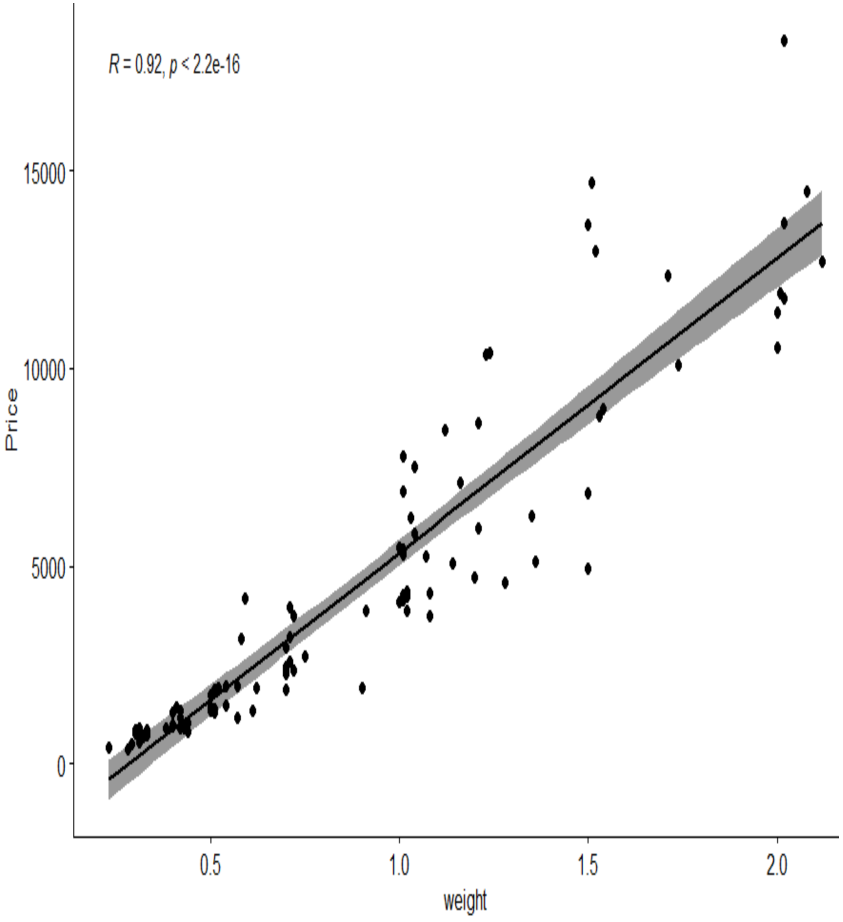
- Note: x, y: numeric vectors with the same length,
  method: correlation method

Visualize using a scatter plot

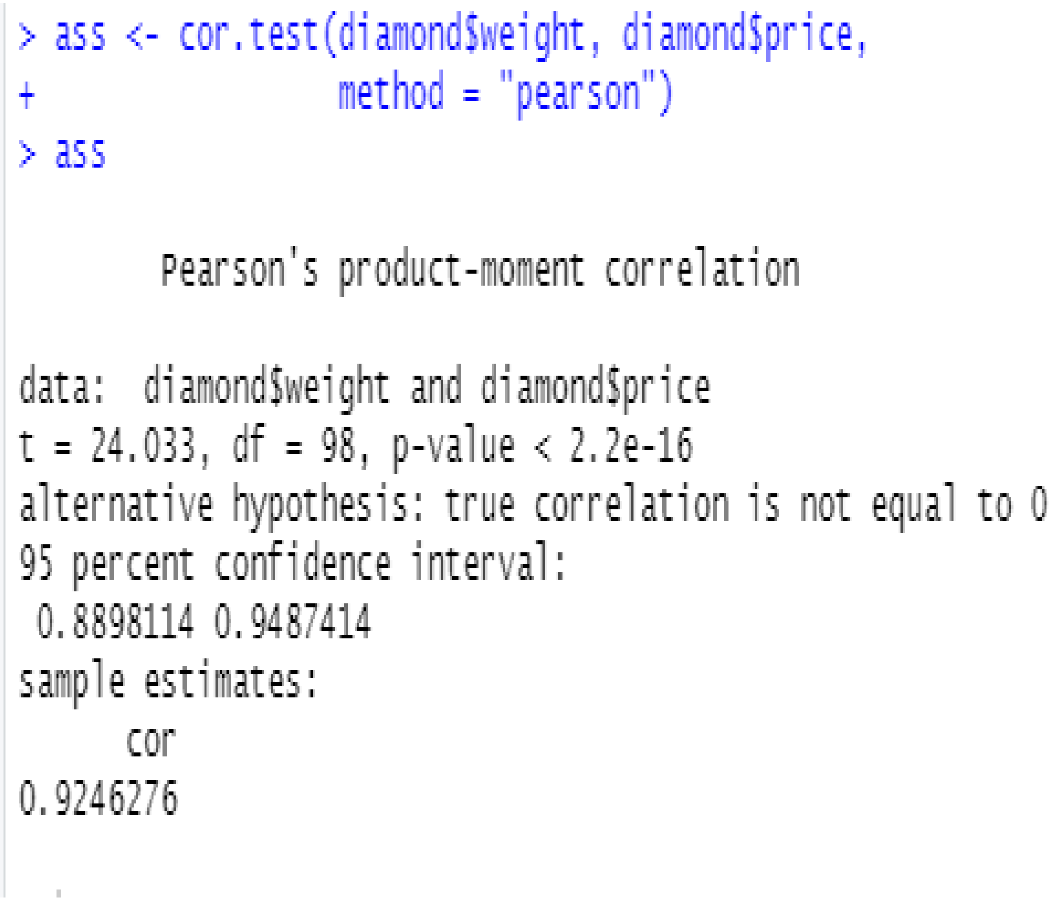
- ggscatter() comand



```
ggscatter(diamond, x = "weight", y = "price",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "weight", ylab = "Price")
```



Pearson correlation in R



```
> ass <- cor.test(diamond$weight, diamond$price,
+                 method = "pearson")
> ass

        Pearson's product-moment correlation

data:  diamond$weight and diamond$price
t = 24.033, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8898114 0.9487414
sample estimates:
      cor
0.9246276
```

Interpretation of results

- t is the t-test statistic value (t = 24.033),
- df is the degrees of freedom (df= 98),
- p-value is the significance level of the t-test (p-value = 2.2e^{-16}).
- conf.int is the confidence interval of the correlation coefficient at 95%
- (conf.int = [0.8898, 0.949]);
- sample estimates is the correlation coefficient (Cor.coeff = 0.92).

- Conclusion

The p-value of the test is 2.2e^{-16}, which is less than the significance level alpha = 0.05. We can conclude that weight and price are significantly correlated with a correlation coefficient of 0.92 and p-value of 2.2e^{-16} .

Correlation matrix

- Is used to investigate the dependence between multiple variables at the same time

- All variables have to be numerical, so you have to remove categorical variables

- You will need a Hmisc package for the rcorr() command to be used
- rcorr() computes both the p-values and the correlation coefficients

Linear regression

- If a pair of variables has a significant linear correlation, then the relationship between the data values can be roughly approximated by a linear equation.
- The process of finding the linear equation which best fits the data values is known as **linear regression** and the line of best fit is called the **regression line**.

Linear regression

Regression analysis concerns the study of relationship between variables with the objective of identifying, estimating and validating the relationship.

Linear regression

- The linear relationship is a straight-line
- $y=bo +b1x$;
- Where a=slope of line (for each unit increase in x, how much does y change)
- b=y-intercept (when $x=0$, what does y equal).

Ordinary Least Squares Regression

- Slope is represented by b_1 and describes the change in *y* for each 1 unit change in *x*

- The *y*-intercept is represented by b_0 and is the predicted value for the response variable (*y*) when $x = 0$

Linear regression

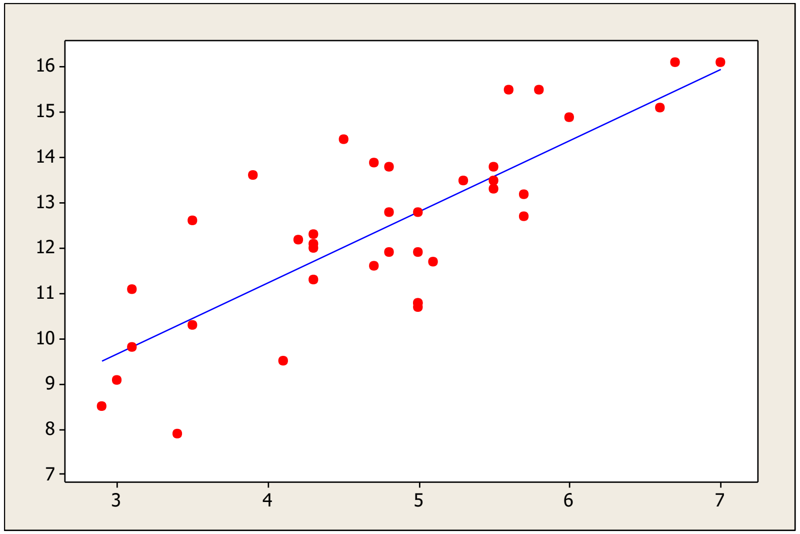
- The line of best fit gives the predicted value y-hat for (i.e., y) at a given level of x.
- Each observation will not fall exactly on the line of best fit and the difference between observed and predicted y value is defined as the residual (or error) in the predicted line.
- $e_i = y_i - \text{y-hat}_i$

Linear regression

- The line of best fit is that which minimizes the sum of the squared error terms, is obtained by the method of least squares (LOS).
- The slope and intercept obtained by LOS are called least square estimates.

Ordinary Least Squares Regression

- The regression line is $\hat{y} = b_1 x + b_0$
- This is a straight line that passes through the points from your data set on the scatterplot



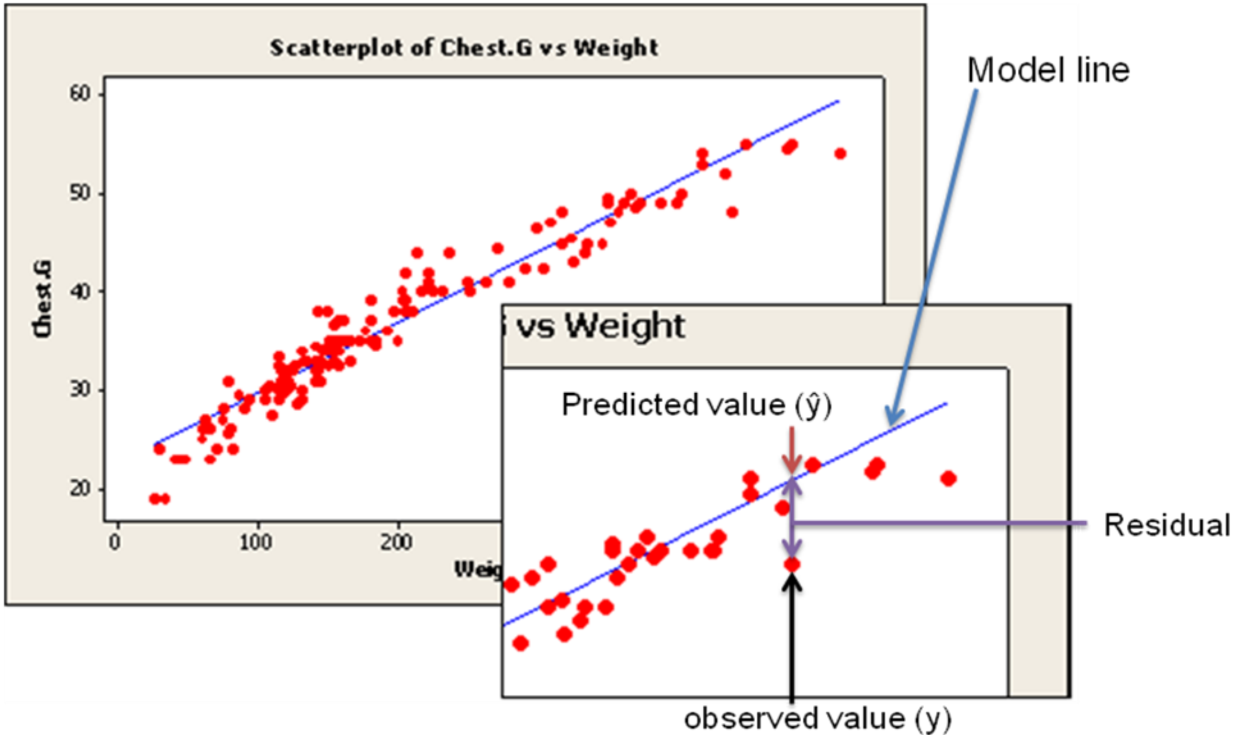
Ordinary Least Squares Regression

- Notice that the regression line doesn’t go through every point
- An OLS regression model tries to balance the differences between the straight-line model and the observed data points
- A line of best fit!

Ordinary Least Squares Regression

- The model does not go through each point
- Some points are closer to the model
- Some points are farther from the model
- The residual is the difference between $\hat{y}$ the observed (*y*) and predicted (   ) values
- Observed – Predicted = Residual

Ordinary Least Squares Regression



SOURCE: SUNY-ESF OPEN RESOURCE

Ordinary Least Squares Regression

- Ordinary least squares regression minimizes the sum of the squared differences between the observed and predicted values
- The best fitting line has the smallest sum of the squared residuals
- Σ residuals^2

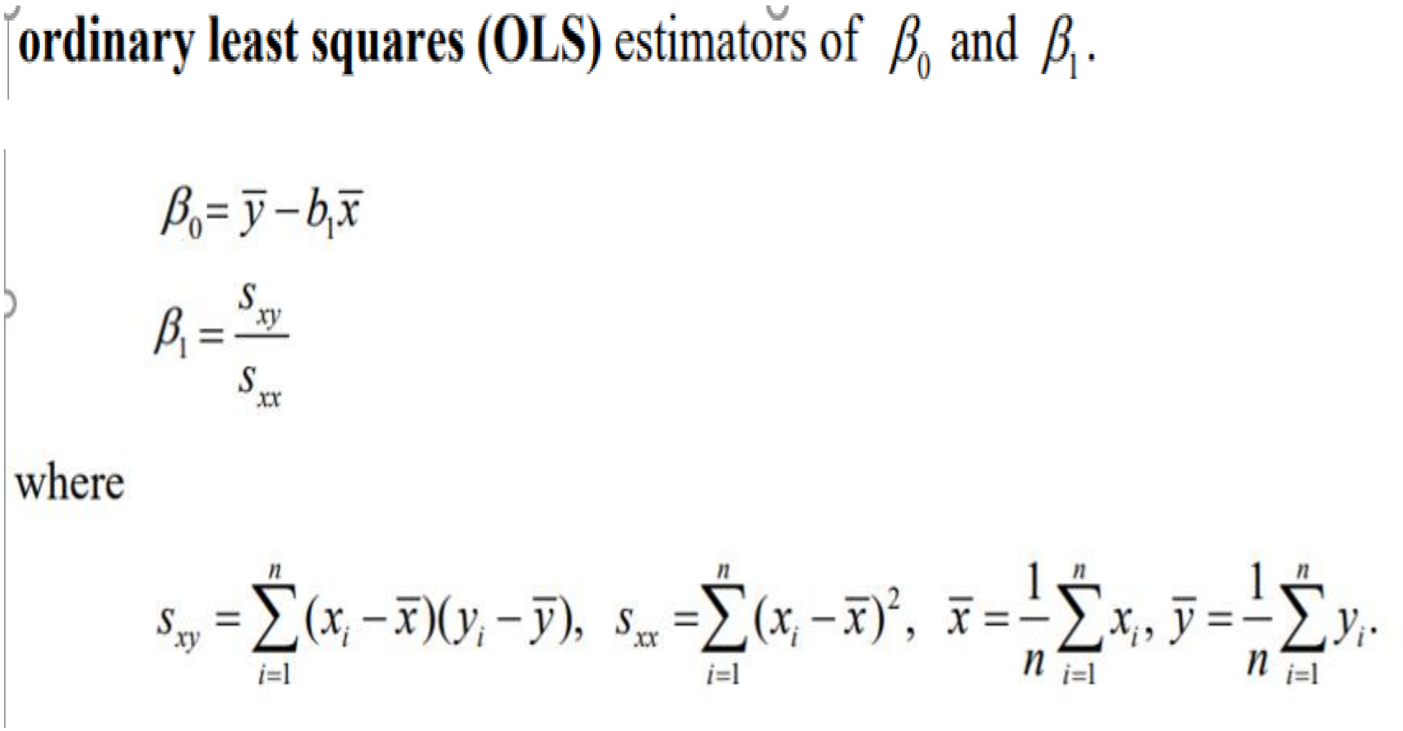
Linear regression

- The line of best fit gives the predicted value y-hat for (i.e., y) at a given level of x.
- Each observation will not fall exactly on the line of best fit and the difference between observed and predicted y value is defined as the residual (or error) in the predicted line.
- $e_i = y_i - \text{y-hat}_i$

Linear regression

- The line of best fit is that which minimizes the sum of the squared error terms.
- The slope and intercept parameters are obtained by LOS called least square estimates.
- From the straight line model, the least squares principle involves the determination of bo and b1 to minimize the vertical distance of the observed value from the fitted line.

Ordinary Least Squares Mathematical expressions (Option One)



ordinary least squares (OLS) estimators of β_0 and β_1.

$$\beta_0 = \overline{y} - b_1\overline{x}$$

$$\beta_1 = \frac{s_{xy}}{s_{xx}}$$

where

$$s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}),\ \ s_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2,\ \ \overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i,\ \overline{y} = \frac{1}{n}\sum_{i=1}^{n}y_i.$$

Ordinary least squares (Option B)

- Regression parameters that give minimum error variance are:

$$b_1 = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2} \qquad \text{and} \qquad b_0 = \bar{y} - b_1\bar{x}$$

where,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

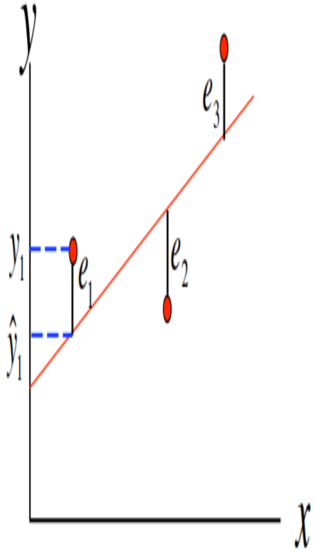
$$\Sigma xy = \sum_{i=1}^{n} x_i y_i \qquad \Sigma x^2 = \sum_{i=1}^{n} x_i^2$$

Regression Residuals

Residuals are the deviations of observed and predicted values

$$e_1 = y_1 - \hat{y}_1$$

$$e_2 = y_2 - \hat{y}_2$$



Residuals Are Useful!

- They allow us to calculate the error sum of squares (SSE):

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

- Which in turn allows us to estimate σ^2:

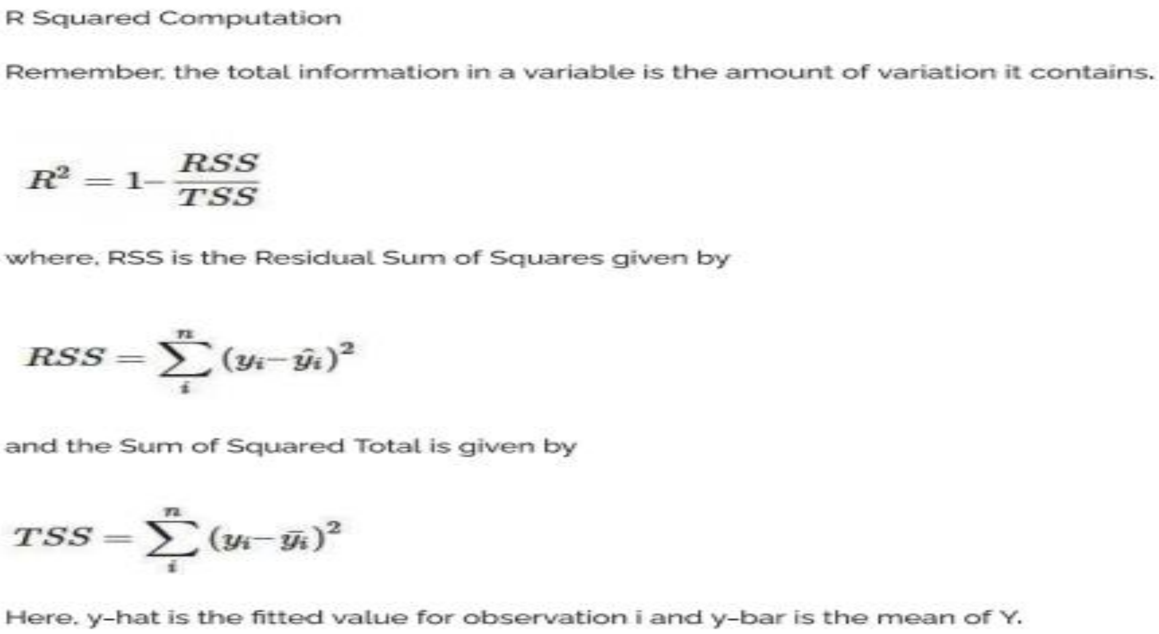
$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

- As well as an important statistic referred to as the coefficient of determination:

$$r^2 = 1 - \frac{SSE}{SST} \qquad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Coefficient of determination/R-Squared

- R-squared tells us the proportion of variation in the dependent (response) variable that has been explained by the model



R Squared Computation

Remember, the total information in a variable is the amount of variation it contains.

$$R^2 = 1 - \frac{RSS}{TSS}$$

where, RSS is the Residual Sum of Squares given by

$$RSS = \sum_{i}^{n} (y_i - \hat{y_i})^2$$

and the Sum of Squared Total is given by

$$TSS = \sum_{i}^{n} (y_i - \bar{y_i})^2$$

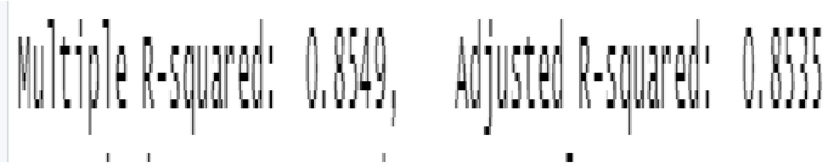
Here, y-hat is the fitted value for observation i and y-bar is the mean of Y.

Coefficient of determination R^2

- Commonly used measure is the coefficient of determination or R^2
- Proportion of total (data) variability explained by the regression model

$$R^2 = \frac{\text{Explained Variation of the model}}{\text{Total variation of the model}}$$

- R^2 is near 1, the model fits the data well, when near 0, then the model poorly fits the data



```
Multiple R-squared:  0.8549,    Adjusted R-squared:  0.8535
```

The model can explain 85% of data variability

Inferences Concerning the Slope

- Parameter of interest: $\beta 1$
- Hypothesis being tested
- $H0 : \beta 1 = 0$ (no linear relationship)
- $H1 : \beta 1 \neq 0$ (there is linear relationship)
- Test statistic is the t-test
- $t =\hat{}\ (\beta 1 – 0)/s/\sqrt{SSxx}$
- *Degrees of freedom (d.f.) = n-2*
- *Page 447-448 Johnson &Bhattacharyya, 2006*

Linear regression

- We assume that the response Y is a random variable and normally distribution with mean μ and variance σ-squared
- The unknown random errors are assumed to be independent, normally distributed with mean zero and variance σ-squqred.

Analysis of Variance table

- Notation
- Total sum of squares
- Regression sum of squares
- Residual sum of squares

Analysis of Variance table

- Notation:
- *TSS := SSyy = (y – y-bar)-sq (Total SS of deviations).*
- *SSR = (ˆy – y-bar)-sq (SS of deviations due to regression or explained deviations)*
- *SSE = (y – ˆy)-sq (SS of deviations for the error or unexplained deviations)*
- *TSS = SSR + SSE*

Analysis of Variance table

Source	DF	SS	MS	F	P-value
Regression	1	7581.8	7581.8	419.36	0.000
Residual Error	29	524.3	18.1		
Total	30	8106.1			

S = 4.252 R-Sq = 93.5% R-Sq(adj) = 93.3%
The regression equation is Volume = - 36.9 + 5.07 Diameter

Interpretation of results

- **Test**
- *H0 : β1 = 0 (no linear relationship)*
- *Ha : β1 = 0 (there is linear relationship)*
- T.S.: *F = MSR/ MSE =* 419.36
- Rejection Region: (critical value: *F.05,1,29 = 4.18)*
- Reject *H0 if F > 4.18* (OR: Reject *Ho if α > p-value)*
- *Conclusion: At 5% significance level there is sufficient statistical evidence to indicate a linear relationship between diameter (x) and Volume (y)*

Exercise

- Nine samples of soil were prepared with varying amounts of inorganic phosphorus x. Corn plants, grown in each soil, were harvested at the end of 38 days and analyzed for phosphorus content. From this value, the plant available phosphorus in the soil was estimated. The data is shown below.
- a) compute the intercept and slope of the relationship between inorganic phosphorus in soil (ppm)- X and estimated plant available phosphorus (ppm)-Y.
- Construct analysis of variance table for analysis
- Compute the coefficient of determination and correlation coefficient, interpret your results.

Example

X	1	4	5	9	11	13	23	23	28
Y	64	71	54	81	76	93	77	95	109

Residuals

- The individual deviations of the observations from the fitted values are called residuals.
- $e= yi- \beta o-\beta 1x$

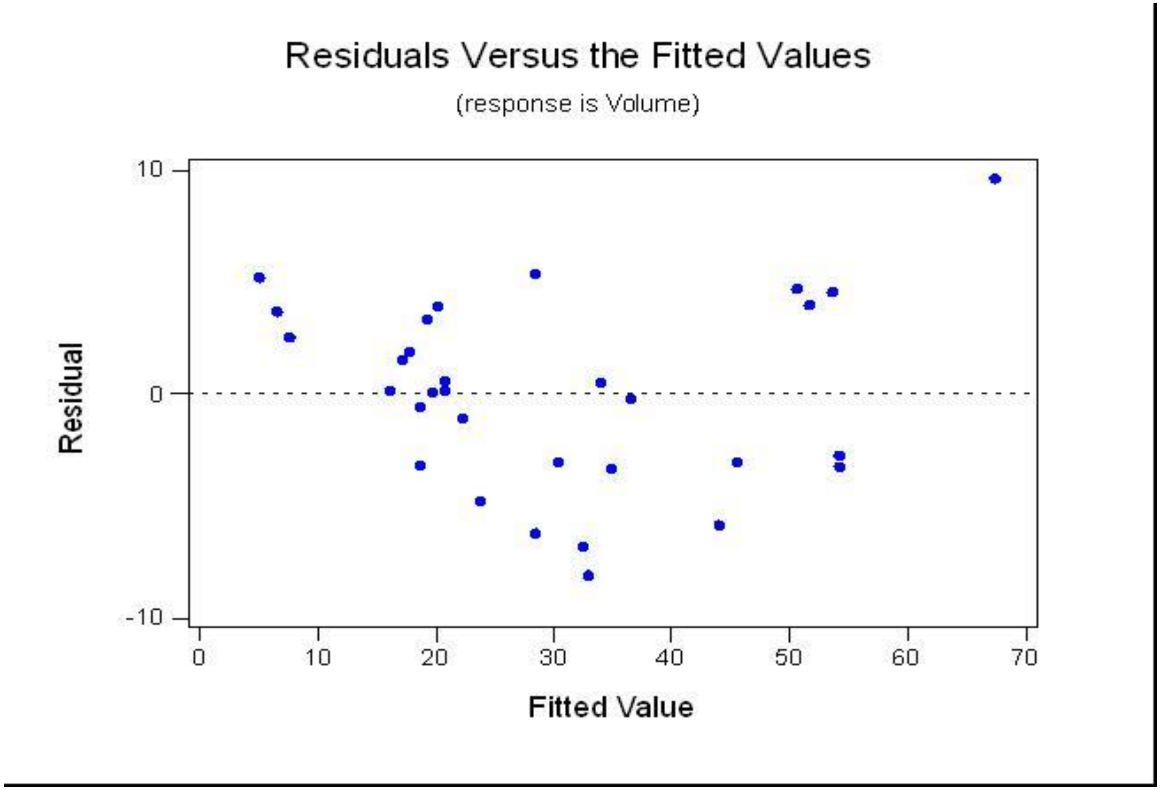
Residual plots to check the adequacy of a statistical model

- Statistical inferences can be made when the postulated model is adequate
- When a model is correct, the residuals are assumed to be normally distributed with mean zero and standard deviation of σ
- Residuals are assumed to have constant variance and no correlations

Residuals

- Histogram of residuals is used to check for normality of residuals
- The normal probability plot of residuals vs normal score also tests for normality of residuals
- To check for validity of the assumption of constant error variance, a plot of residuals versus the predicted value is constructed

Constant error variance



Residual plots

- If the points form a horizontal band around zero, then the assumption of constant variance is valid.
- If the residuals increase with increasing fitted values, this indicates the error variance increases with increasing response.
- If the residuals form a systematic pattern as shown in ---, the model may be inadequate and a square term or nonlinear x term should be considered.

Independence of residuals

- A plot of residuals versus time order detects the violation of the assumption of independence.

Regression

- When a scatter plot shows a relationship on a curve, then one or both variables need to be transformed to exhibit a linear relation.
- Natural log or square either the response or the independent variable.
- Polynomial regression is a special case of regression where the powers of the independent variable play a role of individual variable.

Multiple Linear Regression

- If the unexplained variation is so large
- The response variable y may depend not only on one x variable but other factors as well.
- To obtain a useful prediction model, all variables that significantly affect the response variable should be recorded.
- Multiple regression refers to a model of relationship where the response depends on two or more predictor variables.

Multiple Linear Regression

- A multiple regression model
- $Y_i = \beta_0+\beta_1X_1+\beta_2X_2 + \varepsilon_i$ i=1,….., n
- $X_{1,}$ X_2 are independent variables for the ith experimental unit and Yi is the corresponding response, β*i : unknown parameters.*
- *ε : random error due to other factors not included in the model.*
- Assumptions mentioned under linear relationship still hold.

Example using data on volume, Dbh and height

- Predicted equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
- The regression equation is
- Volume = - 58.0 + 4.71 Diameter + 0.339 Height

Predictor	Coef	StDev	T	P
Constant	-57.988	8.638	-6.71	0.000
Diameter	4.7082	0.2643	17.82	0.000
Height	0.3393	0.1302	2.61	0.014

- S = 3.882 R-Sq = 94.8% R-Sq(adj) = 94.4%

Example

- Analysis of Variance

```
Source          DF       SS        MS           F       P
Regression       2     7684.2     3842.1    254.97   0.000
Residual        28     421.9      15.1
Total           30     8106.1

Source            DF        Seq SS
Diameter     1              7581.8
Height            1          102.4
```

Interpretation

- The regression equation is
- Volume = - 58.0 + 4.71 Diameter + 0.339 Height
- This means that mean volume increases by 4.71 if diameter increases by 1 cm and height is fixed.
- Similarly 1m increase in height with diameter held constant will increase volume by 0.339

Example

- In an investigation into the relationship of water uptake, food intake and egg production, the following records from tweleve birds were noted.
- Compute the regression parameters (Intercept, slopes).
- Construct the analysis of variance table for the regression analysis.

Exercise

Water uptake X2 ml per bird/per day	Food uptake X1 g/bird/day	Egg production Y eggs/10days
175	90	2
342	150	5
252	114	6
362	152	9
284	122	6
219	117	7
229	114	4
260	117	6
88	55	0
132	72	0
254	106	2
199	93	0

Indicator variables

- Explanatory variable used in regression analysis are usually quantitative e.g. temperature, distance, e.t.c.
- Occasionally it is necessary to use qualitative variables as explanatory variables.
- For example; whether type of variety of tree species could be important for fast growth rate.

Indicator variables

- We can model the effect of variety using an indicator or dummy variable
- Qualitative variable has m levels, we need m-1 indicator variables to represent this variable.

Linear regression with grouped data

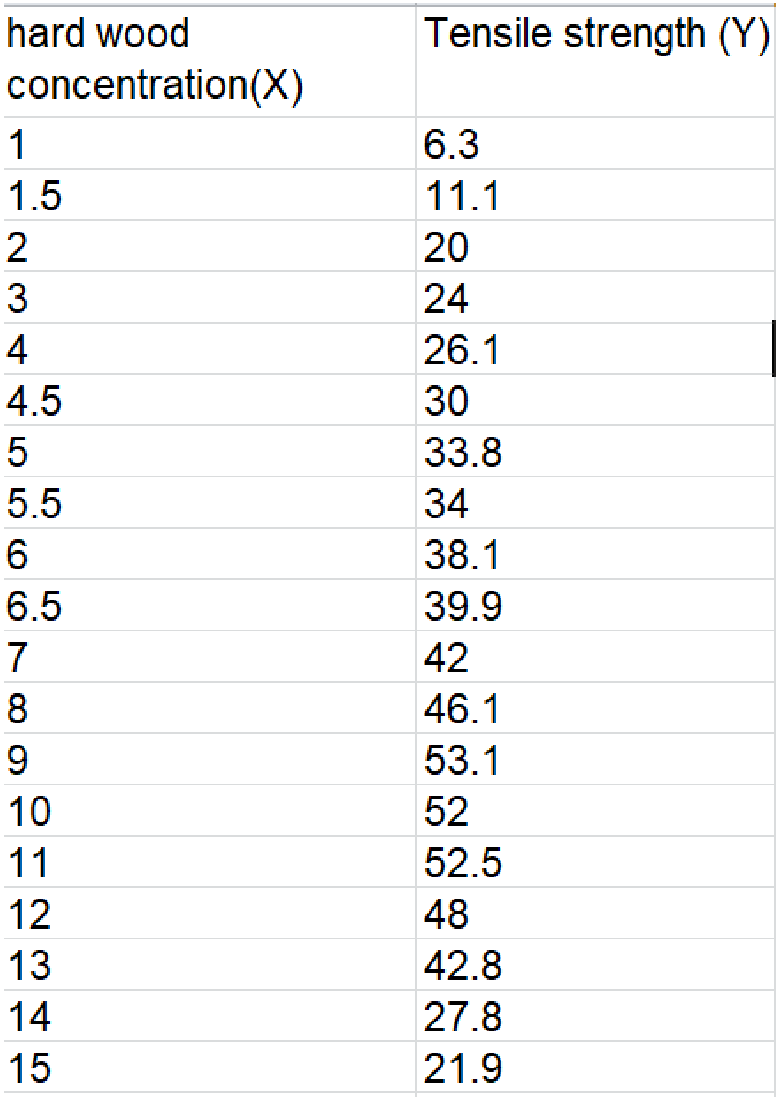
- Use indicator variable to represent the different groups.
- Single line
- Parallel lines
- Different lines

Polynomial regression

- Polynomial can be fitted by using powers and products regresses variables.

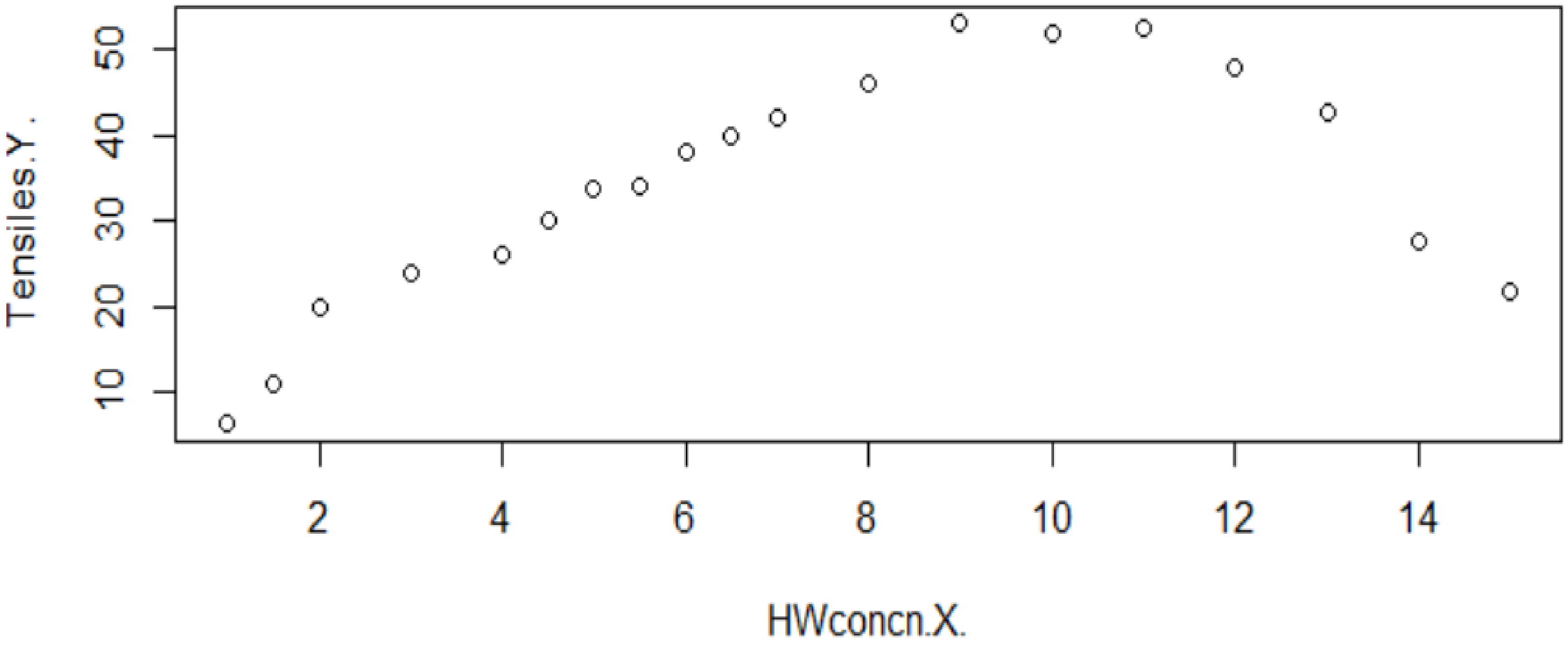
- Data concerning the strength of Kraft paper and percentage of hard wood in the batch of pulp from which the paper was produced.

Data on strength of Kraft paper



hard wood concentration(X)	Tensile strength (Y)
1	6.3
1.5	11.1
2	20
3	24
4	26.1
4.5	30
5	33.8
5.5	34
6	38.1
6.5	39.9
7	42
8	46.1
9	53.1
10	52
11	52.5
12	48
13	42.8
14	27.8
15	21.9

Scatter diagram of strength of Kraft paper



Transformation

- There are common transformations for non-normal data to achieve normality or stability of variance.

	Variable (y)	Transformation
1	Continuous	Square root(y), Natural log(y), 1/y
2	Count (>0)	Square root(y), Square root(y +0.375), Natural log(y)
3	Count (<0)	Square root(y), Square root(y +0.375), Natural log(y +c)
4	Proportion	Arcsin (Square root(y)), 1/sin (Square root(y))

Transformation

- Arcsin transformation is useful for binary proportions.
- If y is a percentage use 1/sin (Square root(y/100))
- Adding a small value c to a count variable, that take on value of zero enables the logarithmic transformation.

Non-linear regression

- Some time the relationship between the response and the independent variable is nonlinear. The curve of the relationship may be S-shaped.
- A function may be fitted by probit analysis(i.e., fitting with a normal cumulative distribution function) or a logistic function may be fitted.
- A logistic function has an S-shaped curve, has asymptotes at 0 and 1, indicating the estimated response is between 0 and 1.
- Usual assumptions for linear regression do not apply in logistic regression (see Montgomery and Peck, 1992)

Correlation & Regression in R

Go to

R