Correlation and Regression

Tsitsi Bandason

BRTI

12th March 2019

Objective of the Session

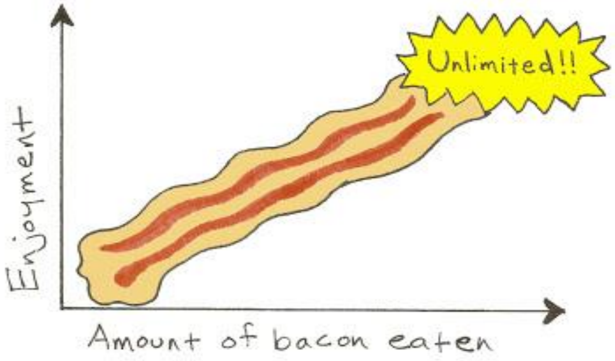
- To find relationships between quantitative variables and testing the validity of the relationship

Introduction

- Statistical analysis is a tool for processing and analysing data and drawing inferences and conclusions

- It is also a double edged tool easily lending itself to abuse and misuse
  - Abuse can occur when poor data is collected and sophisticated techniques used resulting in unreliable result
  - Misuse can occur when good data is collected and poor techniques are used resulting in poor results
  - Misuse can occur when good data is collected and good techniques are used but there is poor interpretation of results

Correlation



- **Correlation** is a bi-variate analysis that measures the strength and direction of relationship between two quantitative variables
  - High **Correlation** means **Strong** relationship
  - Direction of the relationship is indicated by the sign of the coefficient: + sign mean a positive relationship and a – sign means a negative relationship

Types of Correlation

- **Pearson's** coefficient of correlation (r) for symmetric, bell shaped data - for normally distributed variables

- **Spearman** rank correlation is correlation between ranks - for ordinal or skewed data (non-parametric)

- **Kendal's** tau is appropriate - for ordinal or skewed data with ties and/or with small sample (non-parametric)

Questions Answered by Pearson's Correlation

- Is there a statistically significant relationship between age, as measured in years, and bone density, measured in mg/m^2 ?

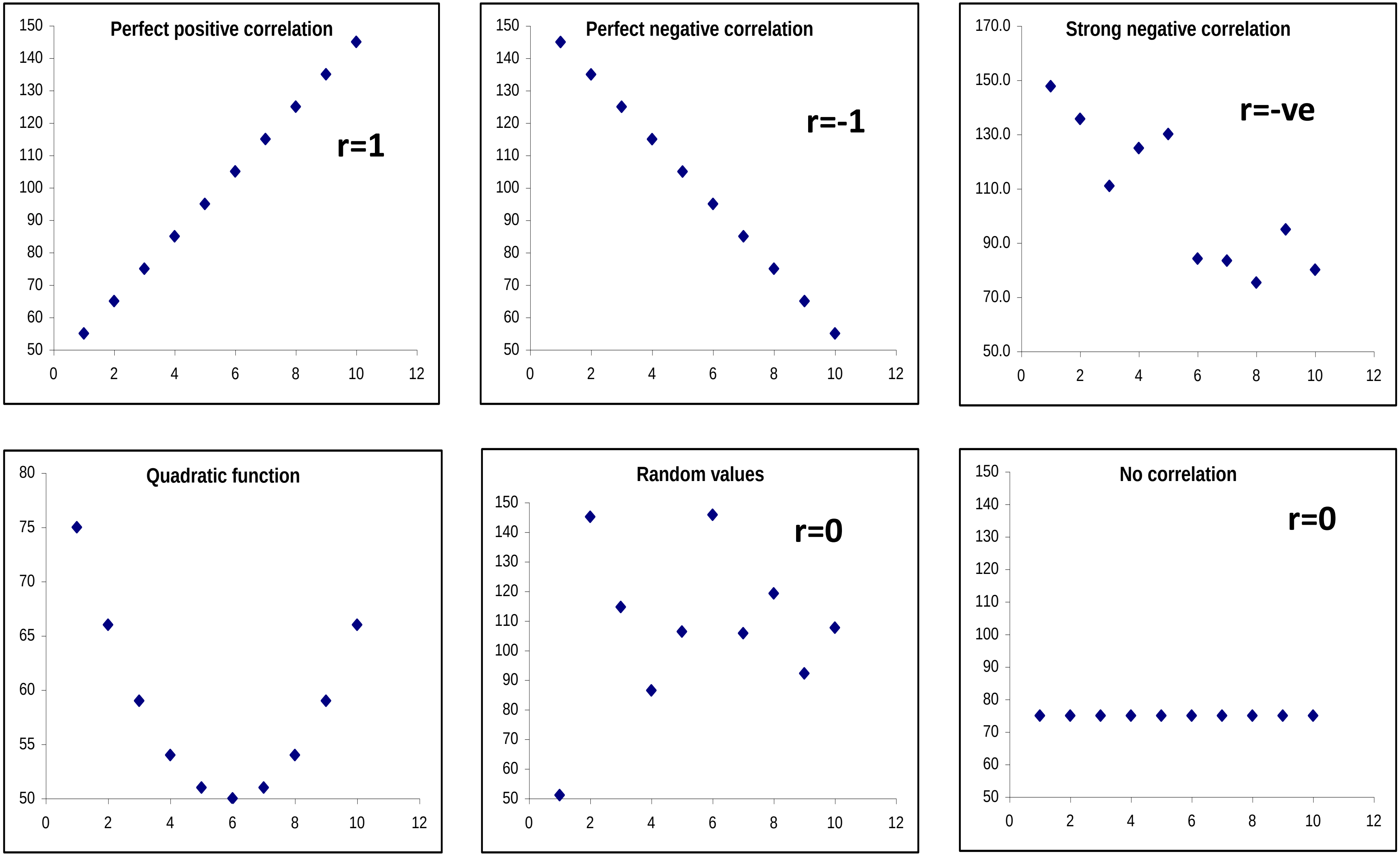
- Assumption
  - Variables are Normally distributed
  - There is a linear relationship between them.
  - The null hypothesis is that there is no relationship between them

Pearson Correlation Interpretation

- Measures strength of linear relationship
- ***r*** lies between -1 and 1
  - If r = -1 there is perfect negative linear relationship
  - If r= 0 there is no linear relationship
  - If r=1 there is perfect positive linear relationship

- Can test whether a correlation coefficient ***r*** is statistically significant using a t-test

Scatter Plot of Relationships



How large should *r* be?

- Physical sciences – high correlations possible
- Biological sciences – investigate high
- Crude Scale

Degree of Relationship	Positive	Negative
Very strong	0.8 to 1.0	-1.0 to -0.8
Strong	0.6 to 0.79	-0.79 to -0.6
Moderate	0.4 to 0.59	-0.59 to -0.4
Weak	0.2 to 0.39	-0.39 to -0.2
Very Weak	0 to 0.19	-0.19 to 0

Steps for Correlation

- Check for normality of each variable (histogram and/or Q-Q Plot)

- Check whether there is a relationship between the variables and type by constructing a scatter diagram
  - Vertical Scale (Dependent): experimental results
  - Horizontal Scale (Independent): Time or classification

- Calculate the correlation coefficient
  - correlation between X and Y is the same as the correlation between Y and X

- Calculate the p-value to check whether the correlation coefficient is statistically significant

Pearson Correlation Formulas

- Correlation Coefficient

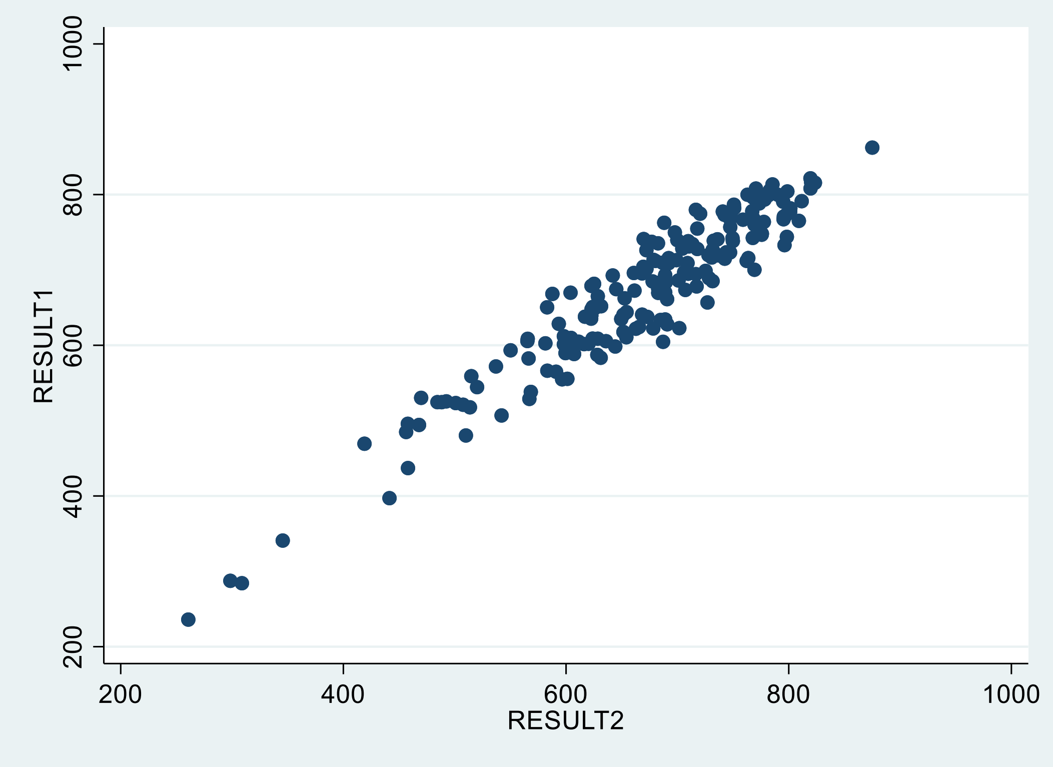
$$r = \frac{N\sum xy - \sum(x)(y)}{\sqrt{N\sum x^2 - \sum(x^2)][N\sum y^2 - \sum(y^2)]}}$$

- Significance test

$$\mathbf{t} = \mathbf{r}\sqrt{\frac{\mathbf{n}-2}{\mathbf{1}-\mathbf{r}^2}}$$

Scatter Plot in STATA

- twoway scatter reading1 reading2



Pearson Correlation Result in STATA

- pwcorr reading1 reading2, star(0.05) sig

```
             |   result1   result2
-------------+--------------------
    reading1 |    1.0000
             |
    reading2 |    0.9485*   1.0000
             |    0.0000
             |
```

- r= +0.95, which means as Reading1 increases, Reading2 also Increases. There is a very strong positive correlation

- $100r^2$ =90%, means 90% of the variability of the data is explained by this relationship result1 and result 2

Correlation Notes

- If r=0, that does not mean there is no relationship
  - there might be a strong non-linear relationship (examine the graphical data)

- Check for scatter plot outliers - can affect the coefficient

- Causation cannot be directly inferred from a strong correlation coefficient (background information is essential)

- Correlation is useful for generating hypotheses

Correlation vs Regression

- Correlation describes the strength and direction of an association between two variables (X and Y/Y and X)

- Regression describes the causal/trend of the relationship and predicts/forecasts future values of Y given X.
  - Helps us to understand how much the Y which is the dependent variable will change when there is a change in X which is the independent variable
  - Helps us predict trends and future values of Y

Correlation vs Regression

- Unlike correlation, it is important which variable goes on which axis for regression

- When we want to explain the variation of variable Y by variable X, variable Y is the dependent and goes on the vertical axis and X is the independent variable and goes on the horizontal axis
  - Dependent also called Response or Outcome variable
  - Independent also called Explanatory or Predictor variable

Linear Regression

- **Linear regression** is an analysis that assesses whether one or more independent variables explain the dependent variable

- If Y represents the dependent variable and X the independent variable, this relationship is described as the regression of Y on X.)
  - The relationship can be represented by a simple equation called the regression equation
  - The direction in which the line slopes depends on whether the correlation is positive or negative

Types of Regression Equation

- **Simple linear regression**: 1 dependent variable and 1 independent variable
- **Multiple linear regression**: 1 dependent variable and 2 or more independent variables
- **Logistic regression**: 1 dichotomous dependent variable and 1 or more nominal, ordinal, interval or ratio-level independent variables
- **Ordinal regression**: 1 ordinal dependent variable and 1 or more nominal or dichotomous independent variables
- **Multinomial regression**: 1 dependent nominal variable, 1 or more interval or ratio of dichotomous independent variables

Assumptions of Linear Regression

- Linear relationship
- Normality
- No or little multicollinearity - multi-variable
- No auto-correlation - residuals
- Homoscedasticity   - residuals
- A sample size of at least ≥ 30

Steps for Linear Regression

- Determine the correlation
- Estimate the model – fit the line
- Evaluate the validity of model

Simple Linear Regression Equation

- Equation $Y= \beta_0 + \beta_1 X + \varepsilon$ ($Y= a + bX$)
  - Y is the estimate of dependent/outcome variable
  - β_0 is the regression coefficient for the intercept
  - β_1 is the regression coefficient for the slope (the change in the mean value of Y for a unit change in X)
  - X is the score on the independent variable
  - ε is the random error term

- $Y= \beta_0$ if $X=0$

Simple Linear Regression Formula

- Calculation of Intercept and Slope using Least squares estimates

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

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

- When you have calculated you can estimate the regression equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Simple Linear Regression
Abuse of Statistical Packages

- The regression line is the best fit line that predicts Y to best possible accuracy.

- You can find a linear regression equation for a set of data using Excel or STATA but that does not necessarily mean the equation is a good fit for your data.
  - Do a scatter plot first to see if relationship is linear
  - Conduct tests on the regression coefficients obtained and residuals
  - Check the model value

Simple Linear Regression Model Adequacy

- Coefficient of Determination R^2
  - is a measure of the amount of variability in the data accounted for by the regression model
  - R^2 =0.95 means 95% of the variability in the data is explained by the regression model, indicating a very good fit of the model
  - R^2 =0.5 means only 50% of the variability in the data is explained by the regression model

Simple Linear Regression Model Adequacy

- Residuals Check
  - Follow the normal distribution
  - Have a constant variance
  - Pattern is random when residuals are plotted in a time or run-order sequence

Question – Simple Linear Regression

- Does the duration on ART have and influence on the bone density?
  - Determine the duration on ART of adolescents and measure the bone density

- The linear regression analysis can then show whether the duration on ART (independent variable) has an effect on the bone density level.
  - Predict when an adolescent has been on ART for X-years the bone density is Y mg/m^2 .
  - Show trend of bone density decrease for every additional year on ART

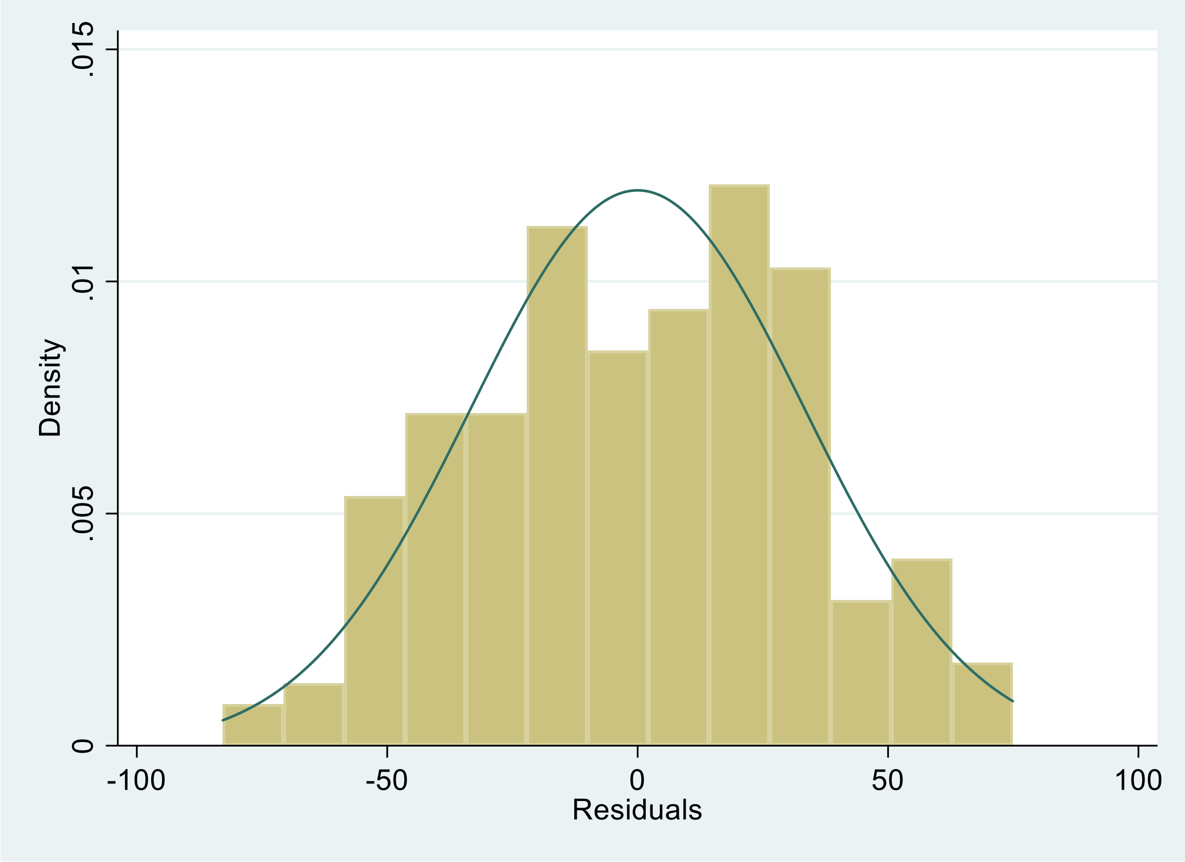
Regression in STATA

- regress reading1 reading2

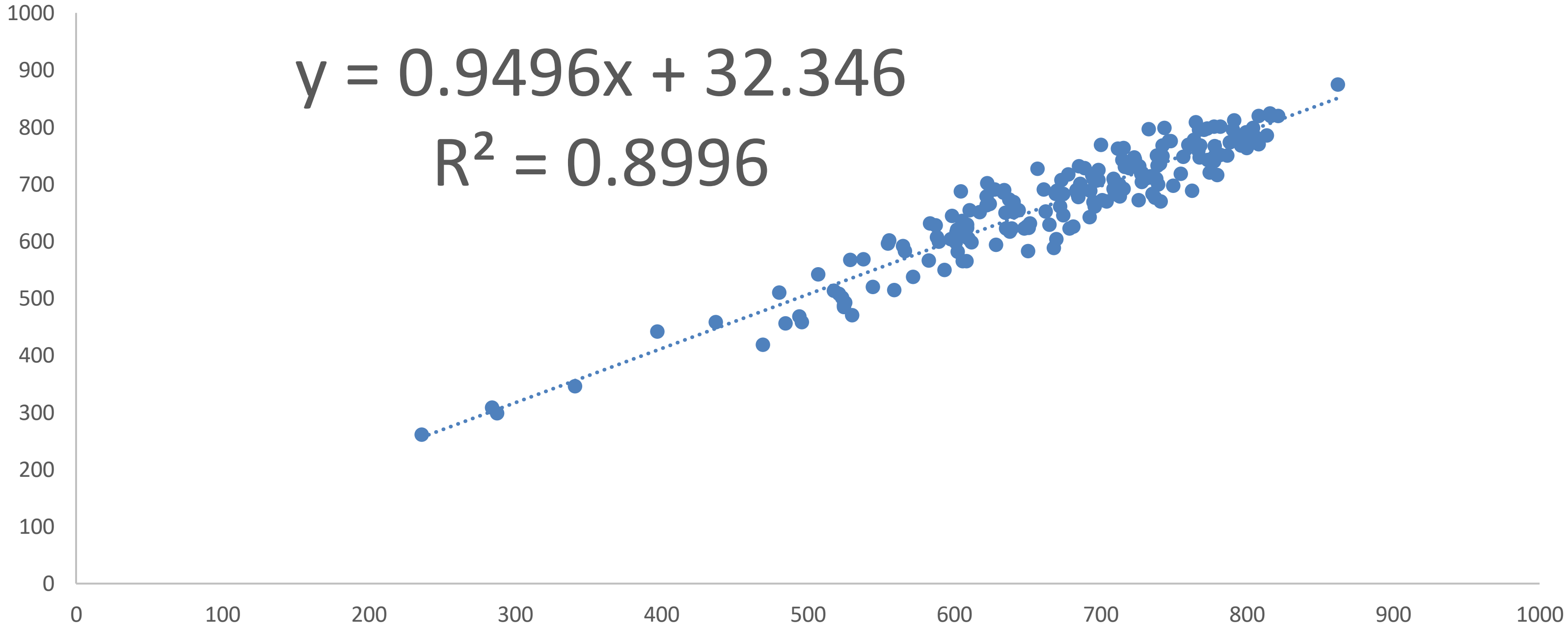
```
 Source |       SS           df       MS      Number of obs   =       184
-------------+----------------------------------   F(1, 182)       =   1631.18
       Model |  1824037.79         1  1824037.79   Prob > F        =    0.0000
    Residual |  203518.169       182   1118.2317   R-squared       =    0.8996
-------------+----------------------------------   Adj R-squared   =    0.8991
       Total |  2027555.96       183  11079.5408   Root MSE        =     33.44

------------------------------------------------------------------------------
     reading1|      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     reading2|   .9473902   .0234573    40.39   0.000      .901107    .9936734
       _cons |   36.49349   15.85037     2.30   0.022     5.219379    67.76761
```

Histogram of Residuals in STATA

histogram resid, normal



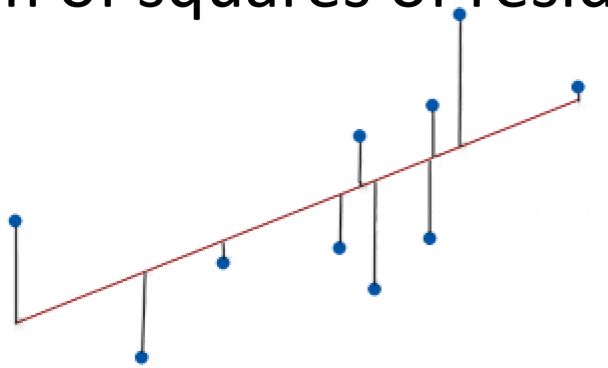
Excel Output



- For every unit increase in x, y increases on average by 0.95 of X
- 90% of the variability in the data is explained by the regression model, indicating a good fit of the model

Notes for Simple Linear Regression

- The aim is to fit a straight line to the data that best describes the relationship and gives an estimate of the relationship of variable X and Y in the population
  - The most useful line is the one that minimises the distance between the data points and the line using the least squares regression
  - The line of 'best fit' is the one that gives the smallest sum of squares of residuals



Notes for Simple Linear Regression

- For clinical/biological data the regression line should not be extended outside range of the data it comes from (zero values of X sometimes do not have meaningful value of Y

QUIZ

- Which do you use to determine the following
  - Do the values of Y tend to be higher (or lower) for higher values of X
  - What is value of Y likely to be when we know the value of X

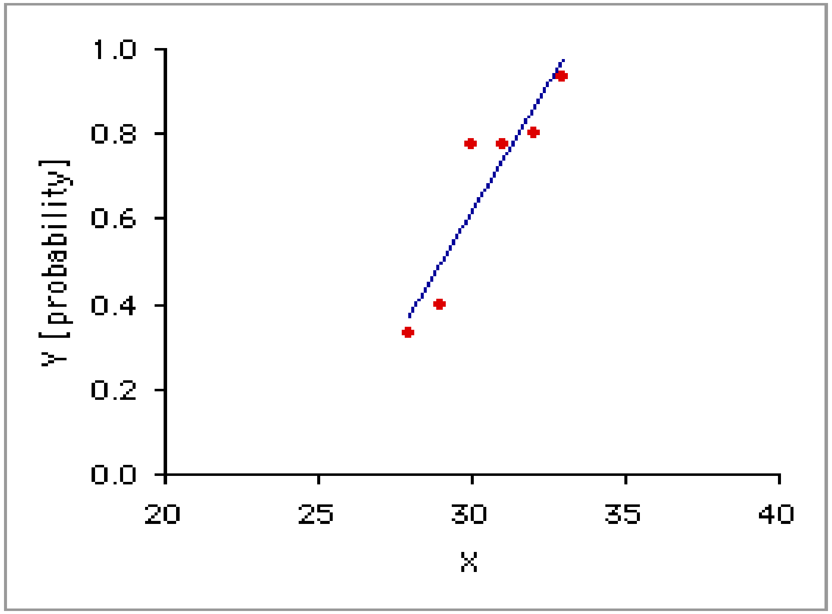
- How to you explain this result
  - Relationship between height and skeletal maturity is given by

    Height = 97.9 + 0.215 Skeletal Maturity

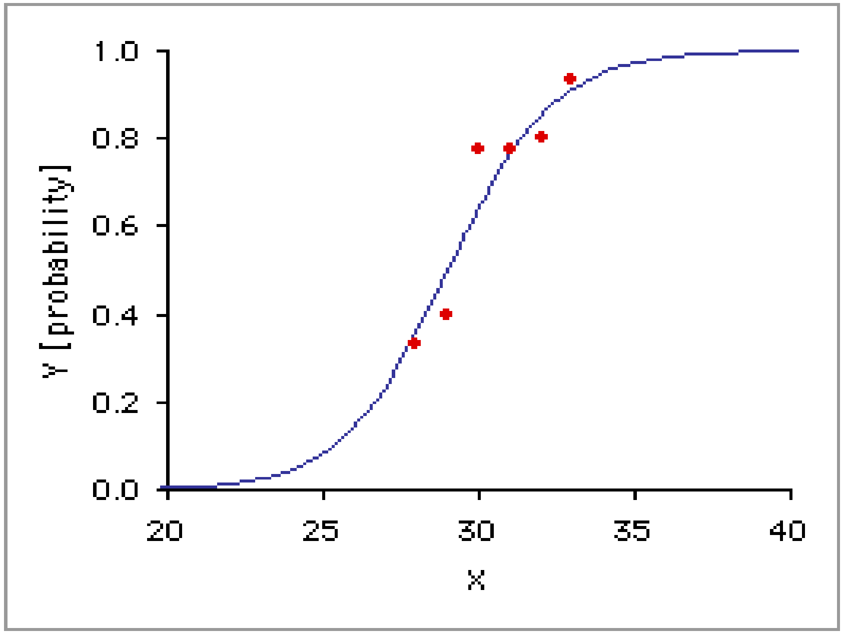
    This means, when skeletal maturity is 0, height is 97.9cm

Logistic vs Linear Regression

- Logistic regression, as shown in Graph B, fits the relationship between X and Y with a special S-shaped curve that is mathematically constrained to remain within the range of 0.0 to 1.0 on the Y axis



A. Simple Linear Regression



B. Logistic Regression

Logistic Regression

- Used to evaluate whether or not an event occurred
  - suitable when outcome/event is measured on a dichotomous scale (binary)
  - e.g presence(Y=1) or absence (Y=0) of disease e.g. HIV

Logistic Regression Equation

- Equation: Logit (p) = log ($\frac{p}{1-p)}$)=β_0 + βX_1
  - p= probability of disease = P(Y=1)
  - β is the rate of change in the “log odds” of the event Y
- Because of these complicated algebraic translations, logit regression coefficients are not easy to interpret
  - We usually translate using exponent function e^{β}
  - The coefficient is called the odds ratio

Odds Ratio

- The ratio $(\frac{p}{1-p)})$ = e^{β}
- It is an odds ratio and is a function of the probability
- Odds indicates how much more likely a certain event occurs in one group relative to the other eg. HIV positive vs HIV negative

Odds Ratio Notes

- Odds ratio = 1: implies no association, that is, the predictor does not affect presence of disease

- Odds ratio >1: implies association, whereby the predictor increases the presence of the disease

- Odds ratio <1: implies association, whereby the predictor reduces the presence of the disease

- If the confidence interval of the Odds ratio crosses 1 e.g. 95%CI 0.9-1.1 this implies there is no statistically difference between the two groups
  - P < 0.05 indicates a statistically significant difference between groups

Odds Ratio Notes

- Comparing drug effect to its placebo has OR: 0.5 95%CI 0.3-0.6”
  - The odds of death when drug is used is 0.5 times less than when it is not used
  - The odds of death when drug is used is 50 % less than when it is not used with the true population effect between 70% and 40%.

With Single Binary Predictor

- Y= 1 if develops disease or Y= 0 if does not develop disease
- X=1 if exposed to a factor or X=0 if not exposed

- If X=0 then Logit (p) = $\beta_0 + \beta(x=0) = \beta_0$
- If X=1 then Logit (p) = $\beta_0 + \beta(x=1) = \beta_0 + \beta$
  - Odd ratio = $e^{\beta}{}_{0}$ or = $e^{\beta}{}_{0}{}^{+\beta}1$

Assumptions of Logistic Regression

- The dependent variable should be dichotomous in nature (e.g., Yes vs. NO).
- Avoid outliers in the data, which can be assessed by converting the continuous predictors to standardized scores
- There should be no high correlations (multicollinearity) among the predictors (if multi-variable)

Model Adequacy

- Hosmer-Lemeshow goodness of fit test
  - Checks how closely the observed and the predicted probabilities match using the Chi-square statistic
  - If sample size is small , the model can fit well but can fail with a larger dataset

- R^2 developed for binary logistic regression
  - Should be interpreted with extreme caution as they have many computational issues which cause them to be artificially high or low

Model Adequacy

- Over-fitting
  - Avoid adding too many independent variables as this increases the amount of variance explained in the log odds
  - reduces the generalizability of the model

Abuse of Statistical Packages

- You can find a logistic regression equation for a set of data
  - Does not necessarily mean the equation is a good fit for your data.
- Conduct tests on the regression coefficients
- Check the model value

Question – Logistic Regression

- How does the probability of having Osteoporosis (yes vs. no) change for every additional year person lives after the age of 30 years?
  - Determine if the person has Osteoporosis and age

- Logistic regression answers
  - causal relationship
  - forecast an outcome
  - show trend

Example – 1 Binary Predictor

- p03resc = 1 if one has HIV or p03resc = 0 if does not have

- p07orph =1 if one is an orphan or p07orph =0 if one is not

- If p07orph =0 then Logit (p) = β_0

- If p07orph =1 then Logit (p) = $\beta_0 + \beta$

Tabulation in STATA

- tab p03resc p07orph, col chi

```
           |        P07ORPH
   P03RESC |         0          1 |     Total
-----------+----------------------+----------
         0 |     8,031      1,148 |     9,179 
           |     97.37      82.41 |     95.21 
-----------+----------------------+----------
         1 |       217        245 |       462 
           |      2.63      17.59 |      4.79 
-----------+----------------------+----------
     Total |     8,248      1,393 |     9,641 
           |    100.00     100.00 |    100.00 

          Pearson chi2(1) = 584.3505   Pr = 0.000
```

Logistic Regression in STATA

- logit p03resc p07orph

```
Iteration 0:   log likelihood = -1854.4047
Iteration 1:   log likelihood = -1669.3329
Iteration 2:   log likelihood = -1651.4466
Iteration 3:   log likelihood = -1651.3956
Iteration 4:   log likelihood = -1651.3956

Logistic regression                             Number of obs     =      9,641
                                                LR chi2(1)        =     406.02
                                                Prob > chi2       =     0.0000
Log likelihood = -1651.3956                     Pseudo R2         =     0.1095

------------------------------------------------------------------------------
     p03resc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     p07orph |   2.066649   .0984151    21.00   0.000     1.873759    2.259539
       _cons |  -3.611167   .0687954   -52.49   0.000    -3.746004    -3.47633
------------------------------------------------------------------------------
```

Logit Model

- Logit (p) = -3.611 + 2.066p07orph

- The coefficient 2.07 implies that a change in opharnhood status results in a 2.07times change in the log odds of being HIV positive

Logistic Regression in STATA

- logit p03resc p07orph, or
- logistic p03resc p07orph

```
 Logistic regression                              Number of obs     =      9,641
                                                 LR chi2(1)        =     406.02
                                                 Prob > chi2       =     0.0000
Log likelihood = -1651.3956                      Pseudo R2         =     0.1095

------------------------------------------------------------------------------
     p03resc | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     p07orph |   7.898308   .7773125    21.00   0.000     6.512729    9.578668
       _cons |   .0270203   .0018589   -52.49   0.000     .0236119    .0309207
------------------------------------------------------------------------------
```

The odds of being HIV positive are 7.9 times more likely when a child is an orphan as compared to when a child who is not an orphan

Notes for Regression

- Logistic regression assumes that the dependent variable is a stochastic event
  - Yes or No, Dead or Alive
  - If the likelihood of having a disease is greater than 0.5 it is assumed diseased, if it is less than 0.5 it is not diseased
  - Note the reference group/unit of the predictor variable (eg. Lowest or 0/kg)
  - **STATA : Outcome coding Yes=1 and No=0**

- Logistic regression is a predictive analysis and also explains the   relationship between one dependent binary variable and one or more independent variable
  - **Interpretation slightly different when more than one variable is used**

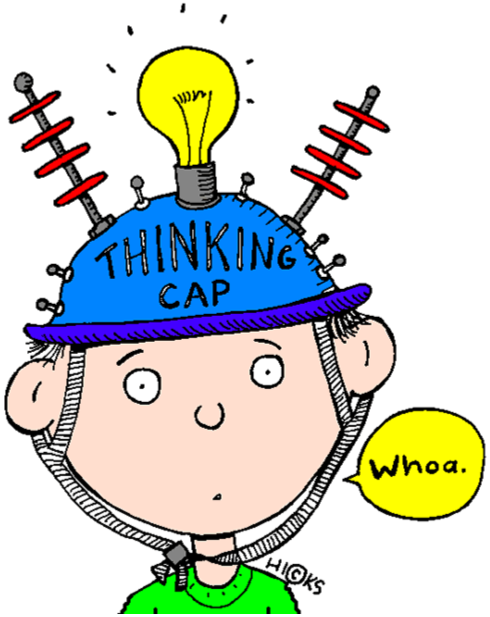
Hint

- When you are **interpreting** an **odds ratio** for Logistic regression
  - Check how it deviates from **1**
  - Easier to understand negative result when expressed as percentage (so multiply result by 100)
  - For an **odds ratio** of 0.75 (0.75-1=-0.25), means that in **one** group the outcome is 25% less likely
  - For. An **odds ratio** of 1.33 (1.33-1)means that in **one** group the outcome is 33% more likely

Quiz

- *Results : In logistic regression analysis, among workers of 8 major job groups, those who experienced prior acute injuries were more likely to have musculoskeletal symptoms in the same region as that of the injury (for the upper extremities), odds ratio [OR] 2.19, 95% confidence interval [CI] 1.51-3.16*
- What is the outcome variable
- What do these results mean in simple english

- You will never need to calculate this manually because of all the statistical packages available
- You need to be able to interpret the results



Thank You