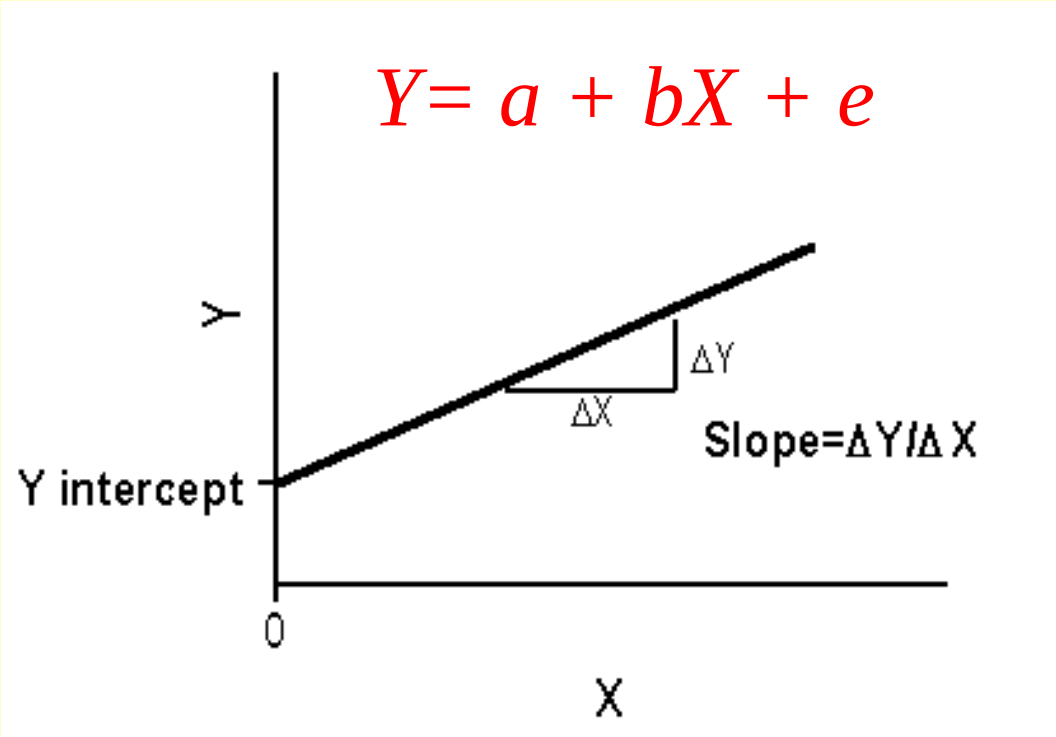
Simple Linear Regression

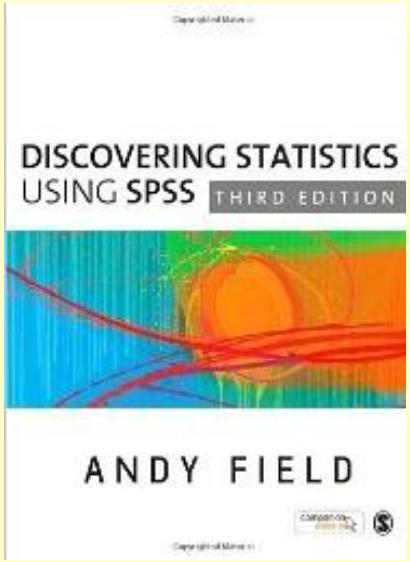


http://www.pelagicos.net/classes_biometry_fa16.htm

Reading - Field: Chapter 7

AIMS

- Understand linear regression with one predictor
- Learn how to assess fit of a linear regression
  - Total Sum of Squares
  - Model Sum of Squares
  - Residual Sum of Squares
  - F
  - R^2
- Learn how to do Regression on SPSS
- Interpret regression model results



What is a Regression

The generic term *Regression* refers to methods that allow the prediction of the value of one (dependent) variable from another (independent).

Regression methods are based on various conceptual models of relationship between these two variables.

Examples include:

- Linear / non-linear regression
- Simple / multiple regression

What is Simple Linear Regression

Most simple version of regression:

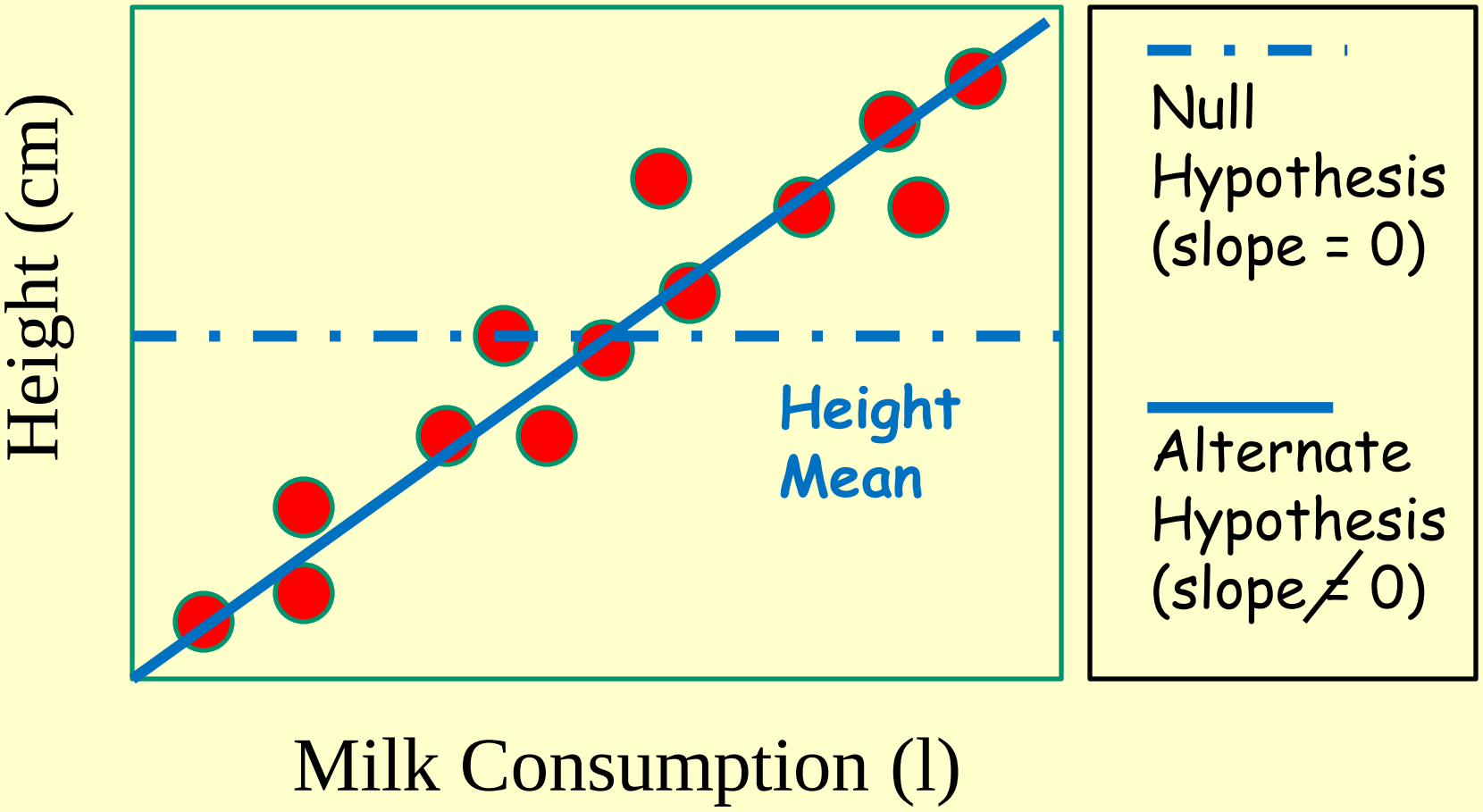
- One independent variable
- Linear relationship

Tests hypothetical model of a linear relationship between two variables:

- Dependent (outcome): Y axis
- Independent (driver): X axis

Simple Linear Regression - How to

Identify the line that best describes relationship between X and Y variables



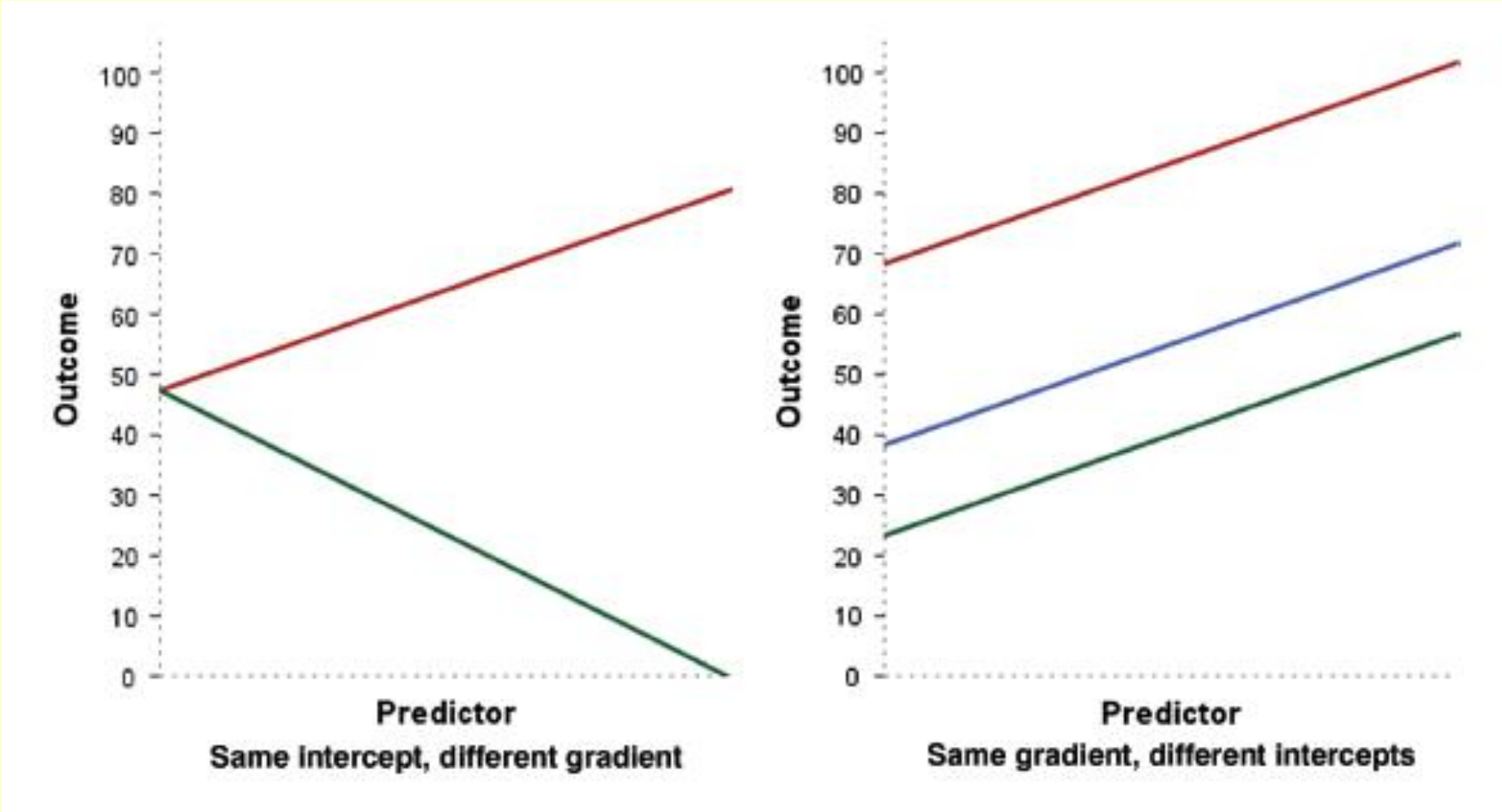
Describing a Straight Line

Model uses linear relationship between X and Y:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

- ϵ_i Error (unexplained portion of variation ~ N (μ, σ))
- b_i
  - Regression coefficient for predictor variable
    - Gradient (slope) of the regression line
    - Direction / Magnitude of Relationship
- b_0
  - Intercept (value of Y when X = 0)
  - Point where regression line crosses Y-axis

Intercepts and Gradients



Calculating Slope of Best-fit Line

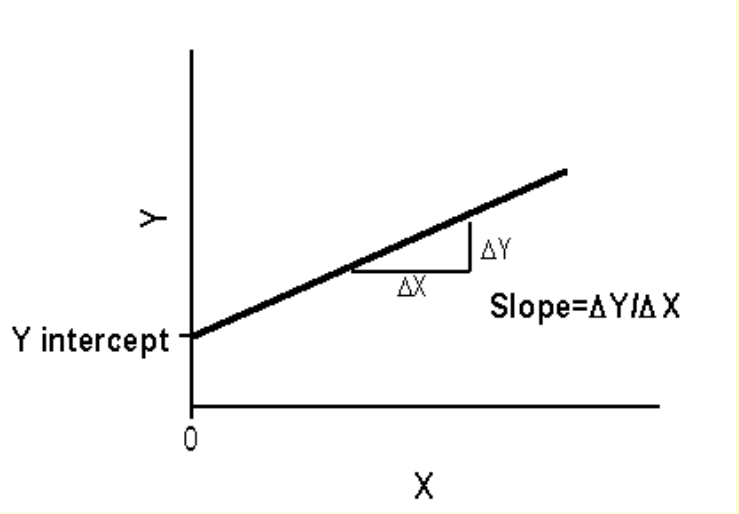
The Regression Coefficient = Slope of Best-fit Line

Covariance between X and Y divided by the variance in X

$$b = \frac{Covariance = \frac{\sum (Xi - \overline{X})(Yi - \overline{Y})}{(n-1)}}{Variance = \frac{\sum (Xi - \overline{X})(Xi - \overline{X})}{(n-1)}}$$

Quantifies best-fit slope of line relating X and Y variables

$$Y = a + bX + e$$



Linear Regression - Assumptions

Linear Regression makes four assumptions:

- (In addition to reliance on "random sampling").

- Variables either interval or ratio measurements.

- Variables normally distributed – No Outliers.

- Linear relationship between the two variables.

How Good is the Fit of the Model

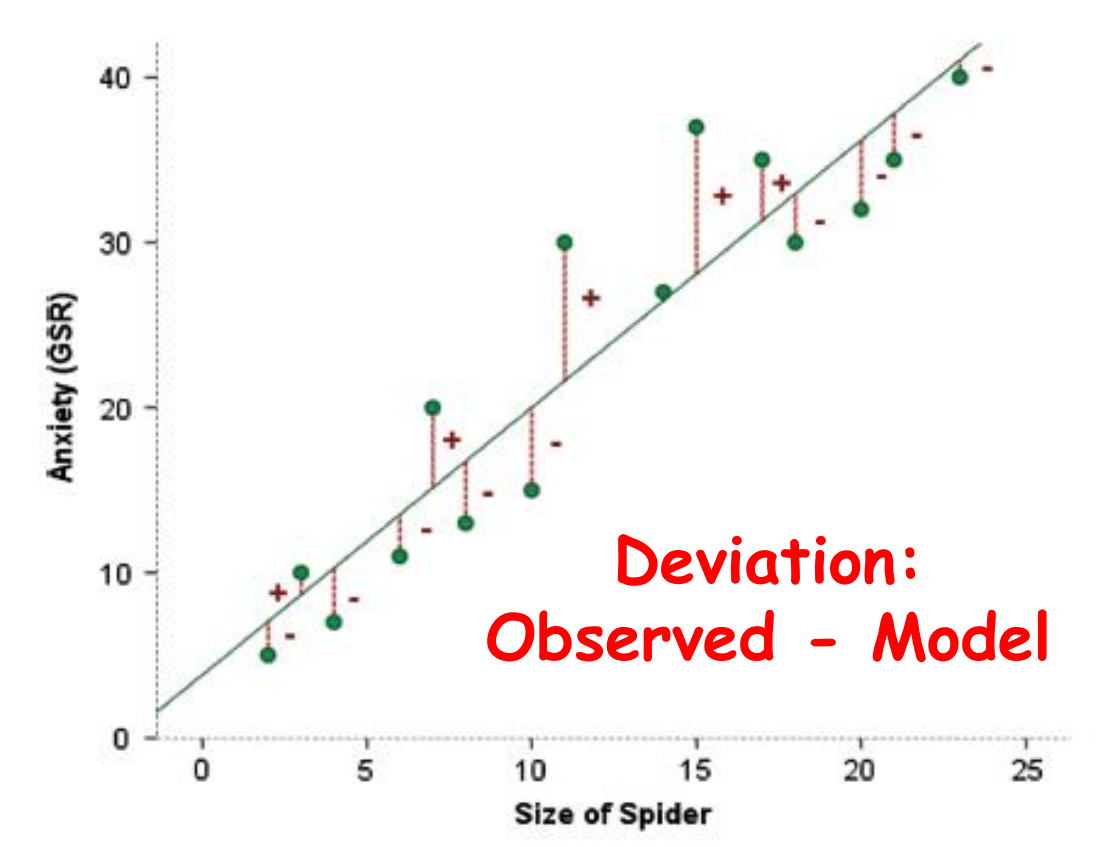
Regression line is based on observations. But, this model might not reflect reality.

- We need a way of testing how well the model fits the observed data: similar to a variance
- Sum of Squares: Sum of the squared deviations (both positive and negative)
- Mean Square: Sum of Squares divided by the degrees of freedom

Measuring Fit

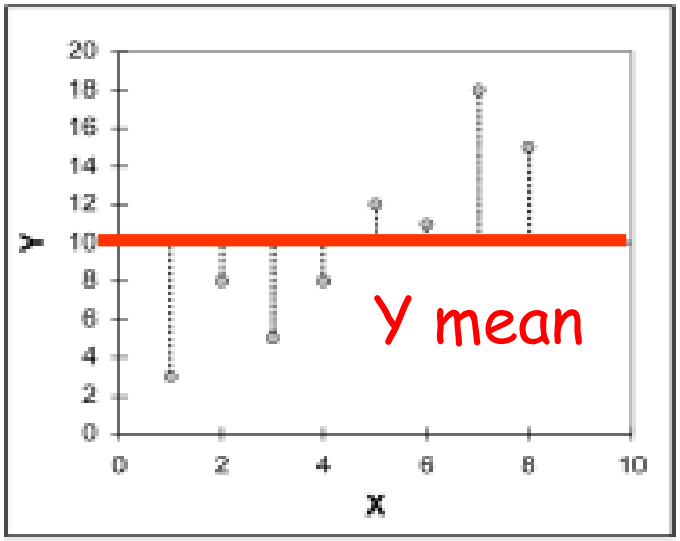
Calculate squared deviations for all data points

Sum of Squares: sum of all squared differences between observed and modeled Y values



Sum of Squares: Σ (Deviations) ^2

Three Different Sum of Squares



df = sample size - 1

SS_T uses the differences between the observed data and the mean value of Y

SS_T - Total SumSquares

Squared difference between observed Y values and their mean calculated from the data

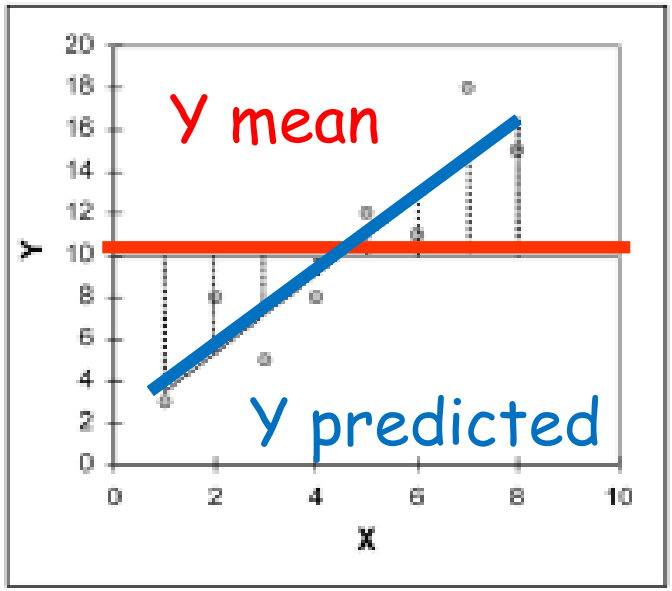
Sum Squares:

Σ (Yi - Y mean) ^2

Mean Squared:

SumSquares / df

Three Different Sum of Squares



df = 1 (linear model)

SS_M uses the differences between the mean value of Y and the regression line

SS_M

Model SumSquares

Squared difference between predicted Y values from regression model and mean of Y data

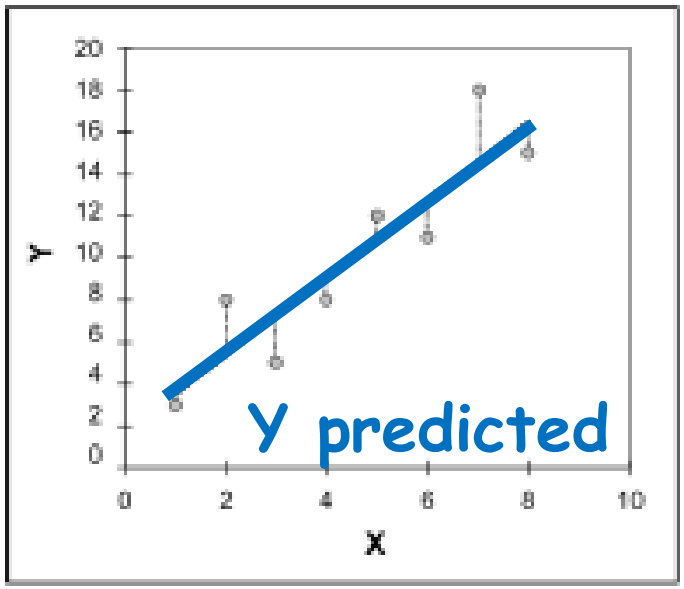
Sum Squares:

Σ (Y predicted – Y mean) ^2

Mean Squared:

SumSquares / 1

Three Different Sum of Squares



df = sample size - 2

SS_R uses the differences between the observed data and the regression line

SS_R

Residual (Error) SumSquares

Squared difference between predicted Y values from regression model and observed Y data

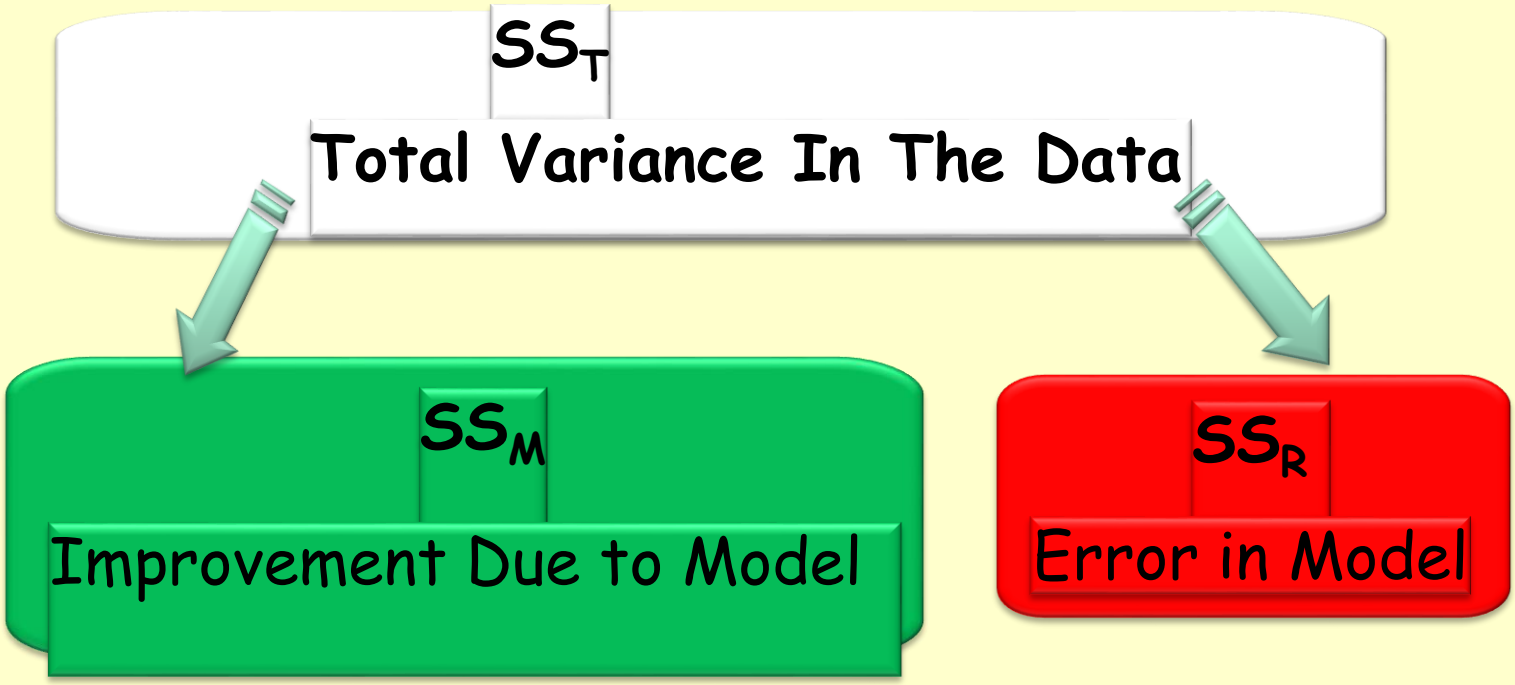
Sum Squares:

Σ (Yi - Y predicted) ^2

Mean Squared:

SumSquares / df

Testing the Model - ANOVA



If model results in better prediction than using the mean, then we expect SS_M to be much greater than SS_R

Linear Regression -Output

Variable List

a. Predictors: (Constant), Advertsing Budget (thousands of pounds)

b. Dependent Variable: Record Sales (thousands)

Sum Squares

Model		Sum of Squares
1	Regression	433687.833
	Residual	862264.167
	Total	1295952.000

SS_M

SS_R

SS_T

Mean Squares

Model	Sum of Squares	df	Mean Square
Regression	433687.833	1	433687.833
Residual	862264.167	198	4354.870
Total	1295952.000	199	

MS_M

MS_R

Testing the Model: R squared

R^2 - Coefficient of Determination

The proportion of total variance accounted for by the regression model

Ranges from
0 (none) to 1 (all)

$$R^2 = \frac{SS_M}{SS_T}$$

Testing the Model: F Test

F statistic: Mean Squares Ratio

- Sums of Squares: sums of squared deviations
- Calculate averages called Mean Squares, MS

F statistic =
ratio of model MS
(regression variance)
divided by residual MS
(error variance)

$$F = \frac{MS_M}{MS_R}$$

MODEL

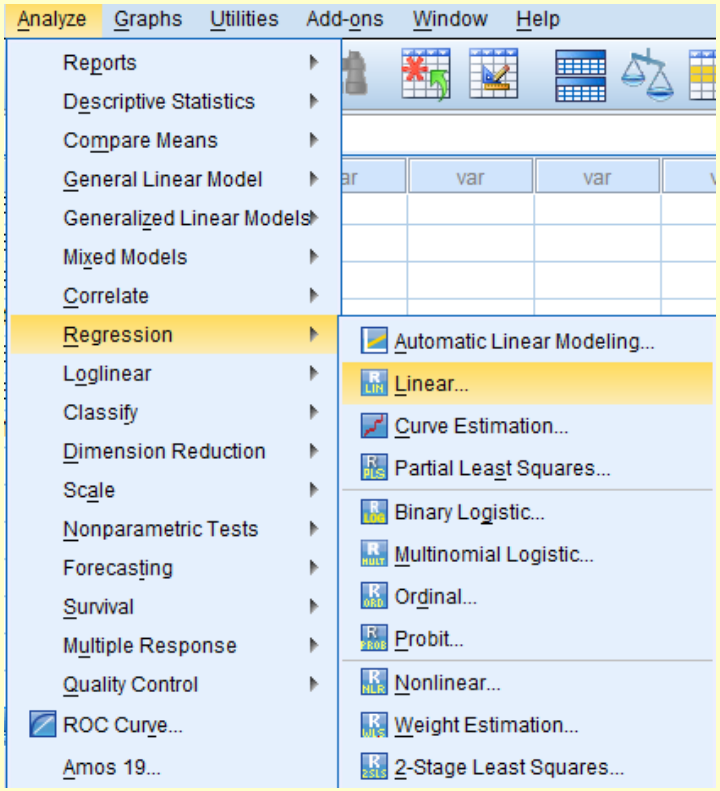
ERROR

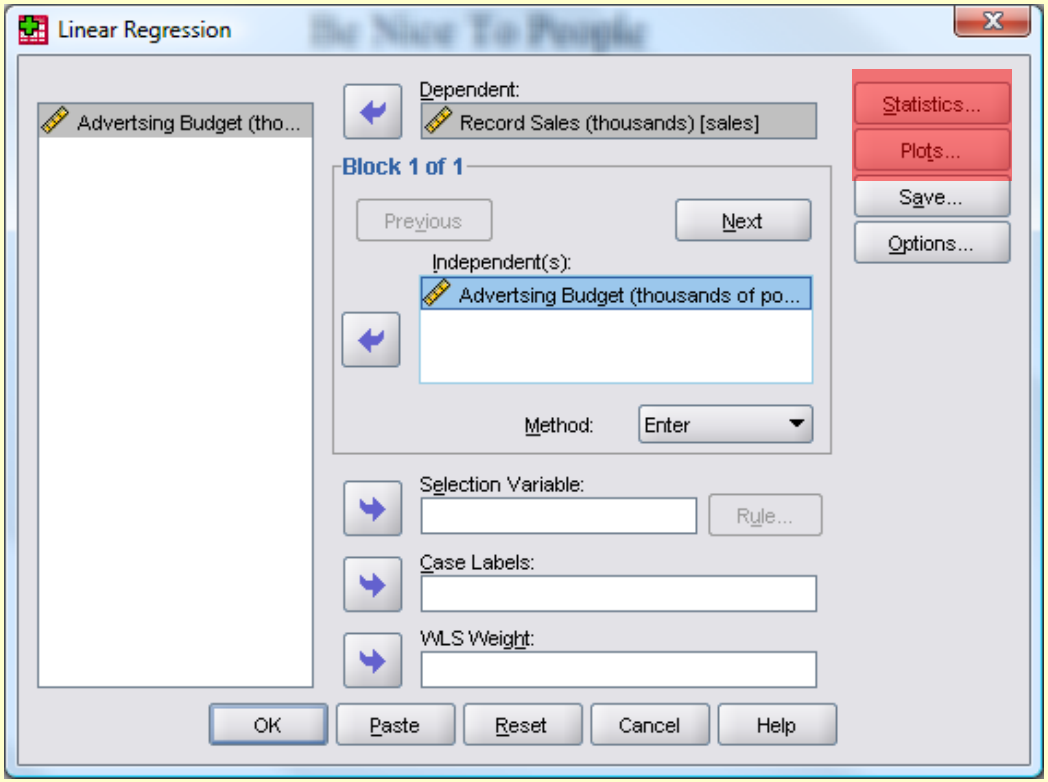
NOTE: F ranges from 0 to a very large number
The larger the F value, the stronger model

Linear Regression – An Example

- A record company boss was interested in predicting record sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales in week after release
- Predictor variable:
  - Amount (£s) spent promoting record before commercial release

Linear Regression - SPSS





Select Dependent / Independent Variables

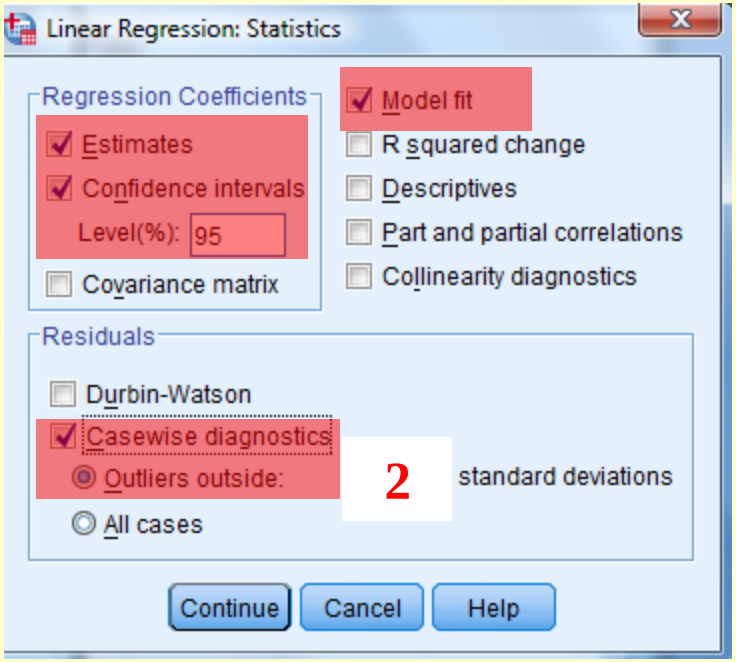
Request Additional Plots and Statistics

Linear Regression - SPSS

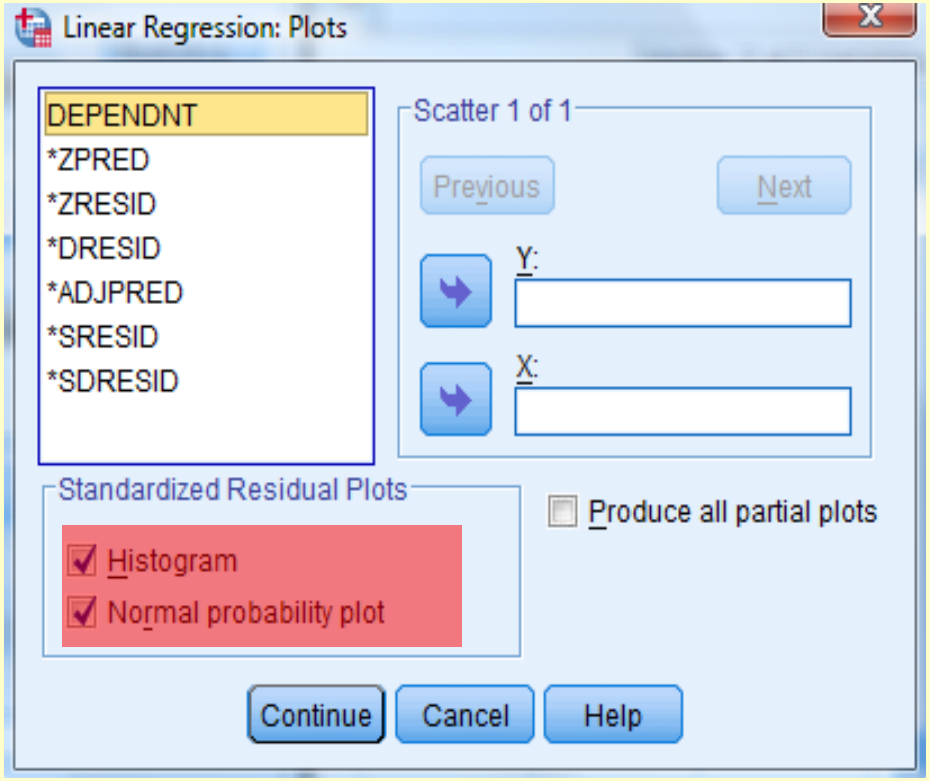
Coefficient Estimates
(point estimate +/- 95% C.I.)

Model Fit (R squared)

Outliers (> 2 or 3 SD)



Linear Regression – SPSS



- Normality of Errors:
  p-p plot.

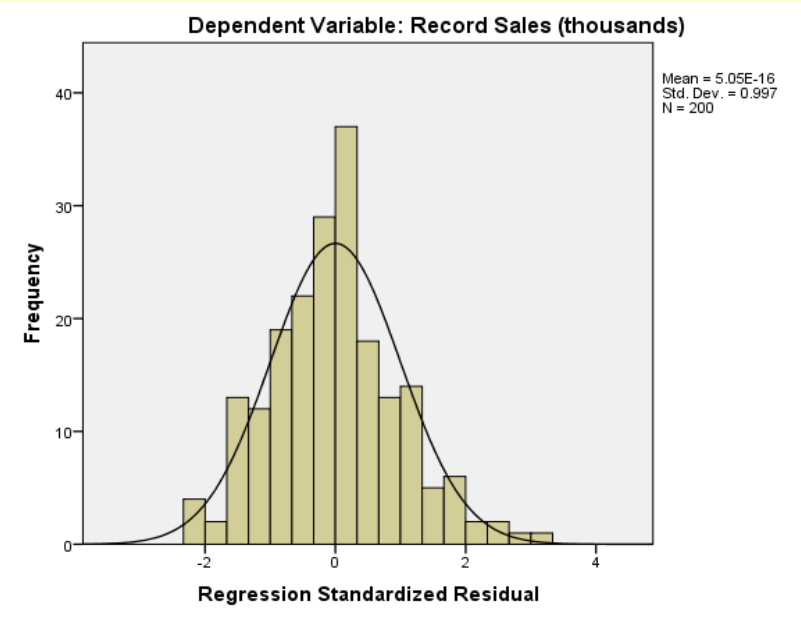
- Homoscedacity / Independence:

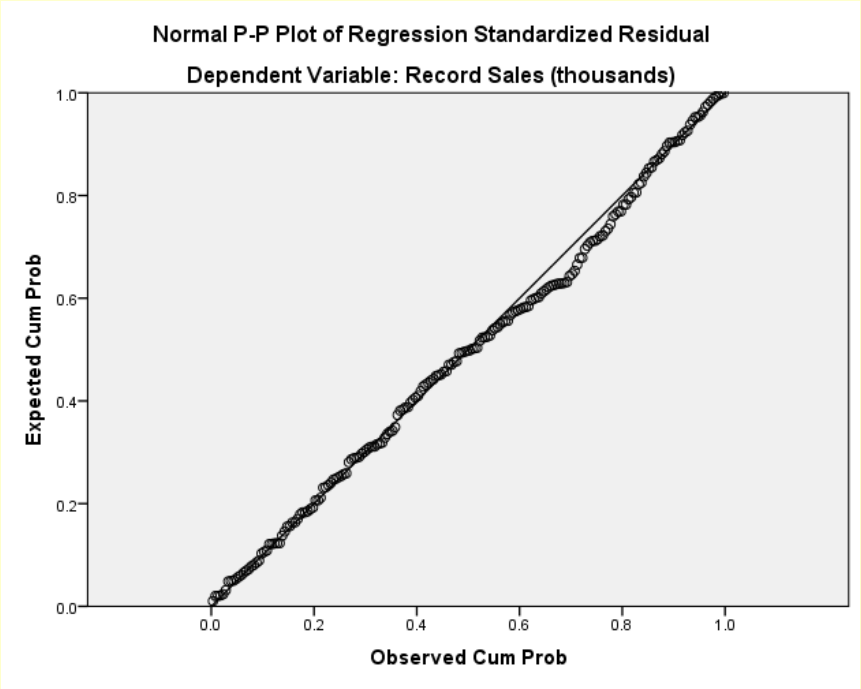
  Plot ZRESID against ZPRED.

Outlier Plots
- Histograms
- Probability Plots

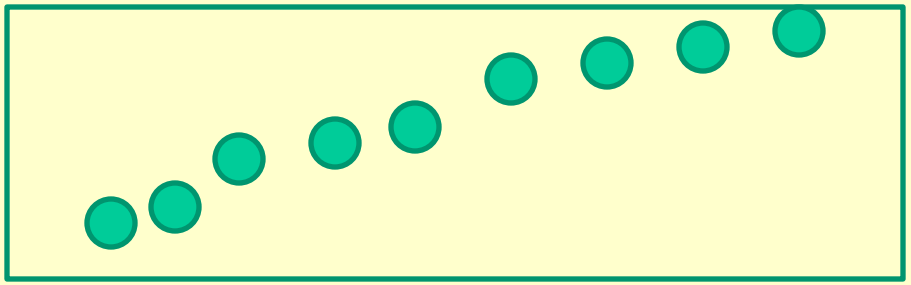
Linear Regression – SPSS

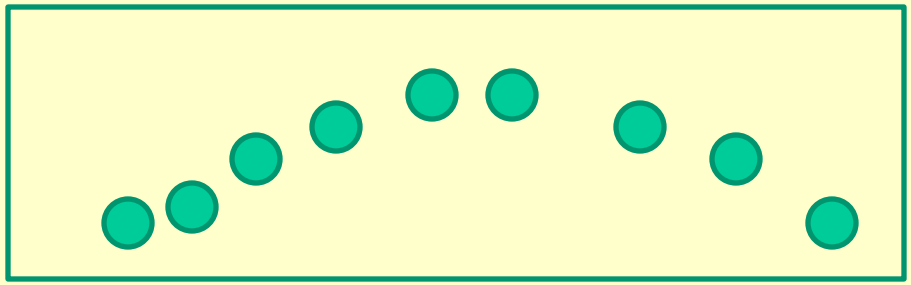
Output: Residuals (look for normality) – perform test



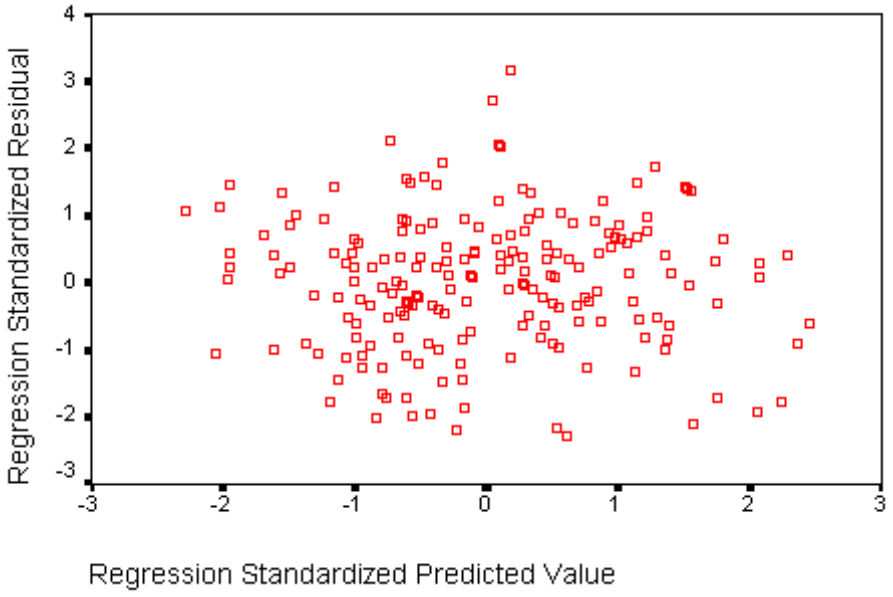


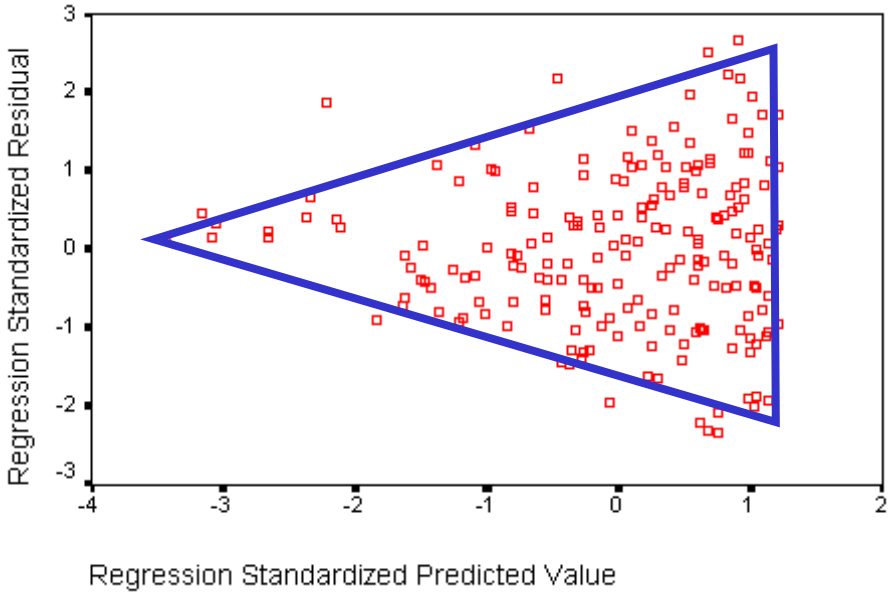
Independence & Homoscedasticity





Errors are not independent: obvious linear or non-linear patterns

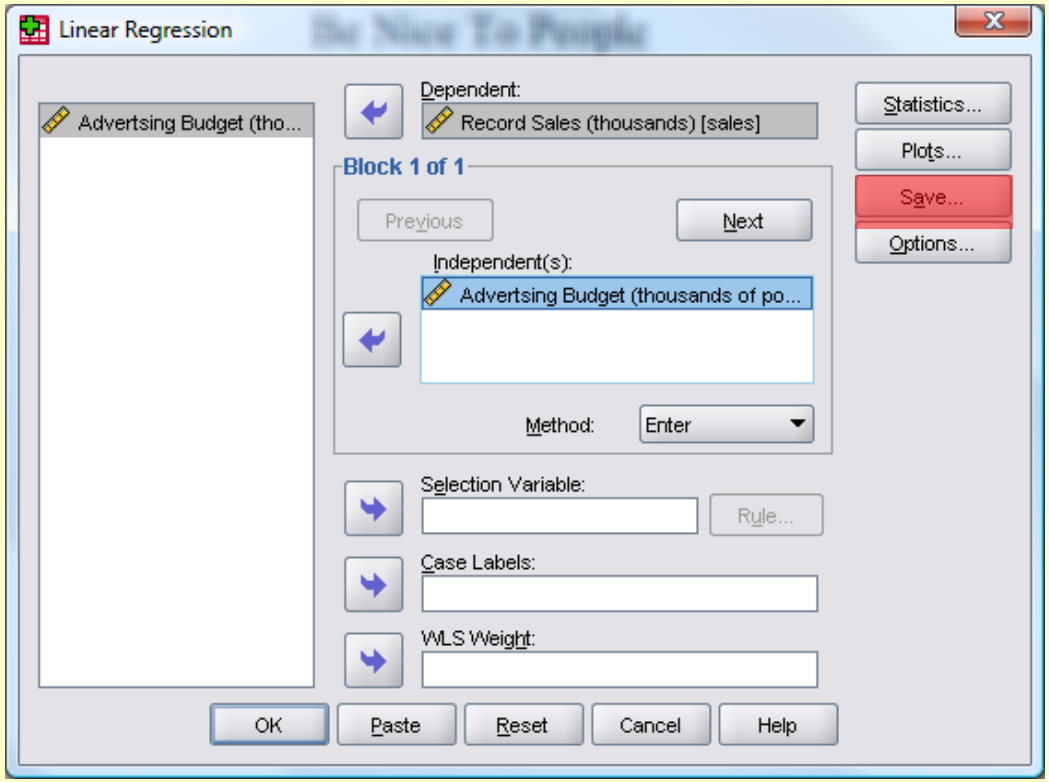




Error cloud: Equal variance with changing predicted value.

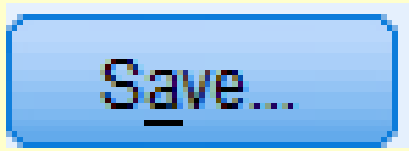
Triangular Pattern: Variance increases with predicted value.

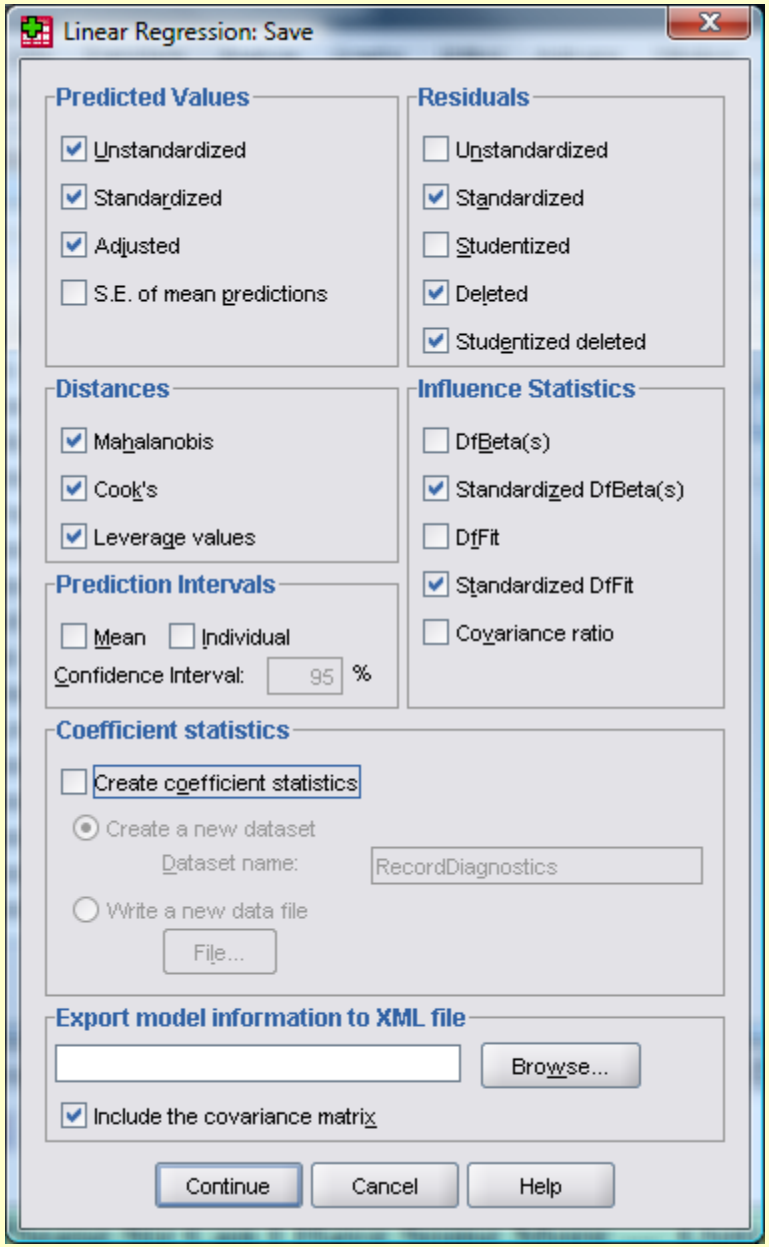
Linear Regression - SPSS



Saves output into data file:

- Predicted Values
- Model Residuals
- Distances & Influences





Linear Regression – SPSS

Output: Intercept and slope estimates (+/- SE)
Significance of intercept and slope (p values)
Ability to make linear predictions

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134.140	7.537		17.799	.000
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000

$Y = a + bX + e$ $\quad\quad$ $Y = 134.140 + 0.096X$

$$\text{Record Sales}_i = b_0 + b_1 \text{Advertising Budget}_i$$
$$= 134.14 + (0.09612 \times \text{Advertising Budget}_i)$$

Linear Regression – SPSS

Output: Best-fit linear relationship (+/- 95% CI)
Coefficient of Determination (% variance)

Model Summary

Model	R	R Square	Adjusted R Square
1	.578[a]	.335	.331

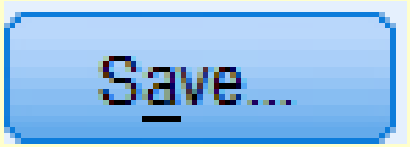
Output: ANOVA results (F-test)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[a]
	Residual	862264.167	198	4354.870		
	Total	1295952.000	199			

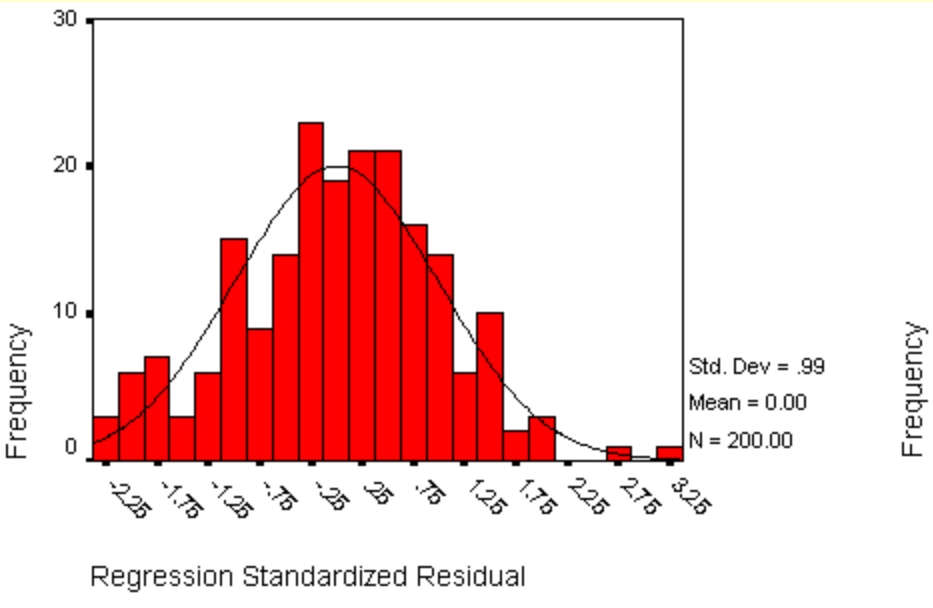
a. Predictors: (Constant), Advertsing Budget (thousands of pounds)
b. Dependent Variable: Record Sales (thousands)

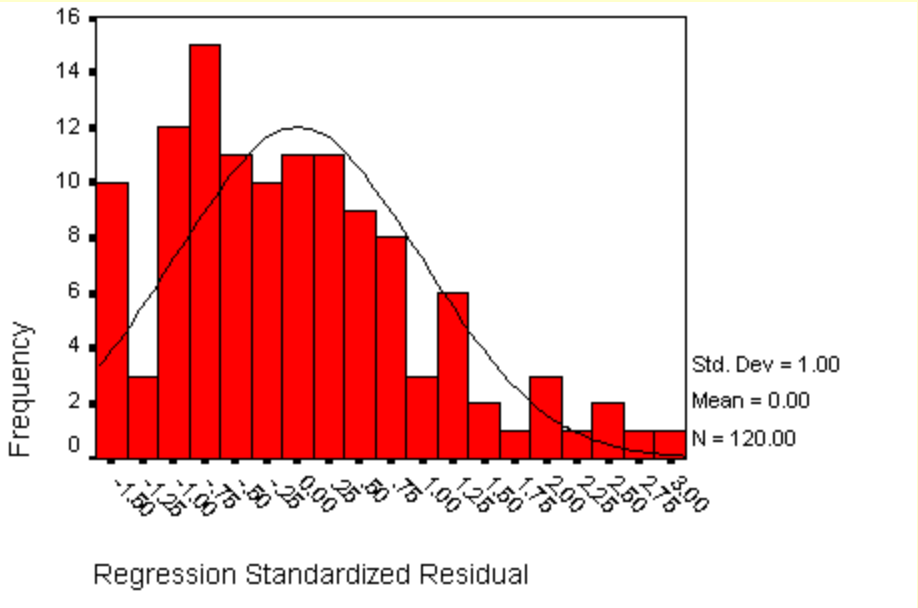
Normality of Errors: Histograms



NOTE: Save residuals using SPSS menu (when performing the linear regression)

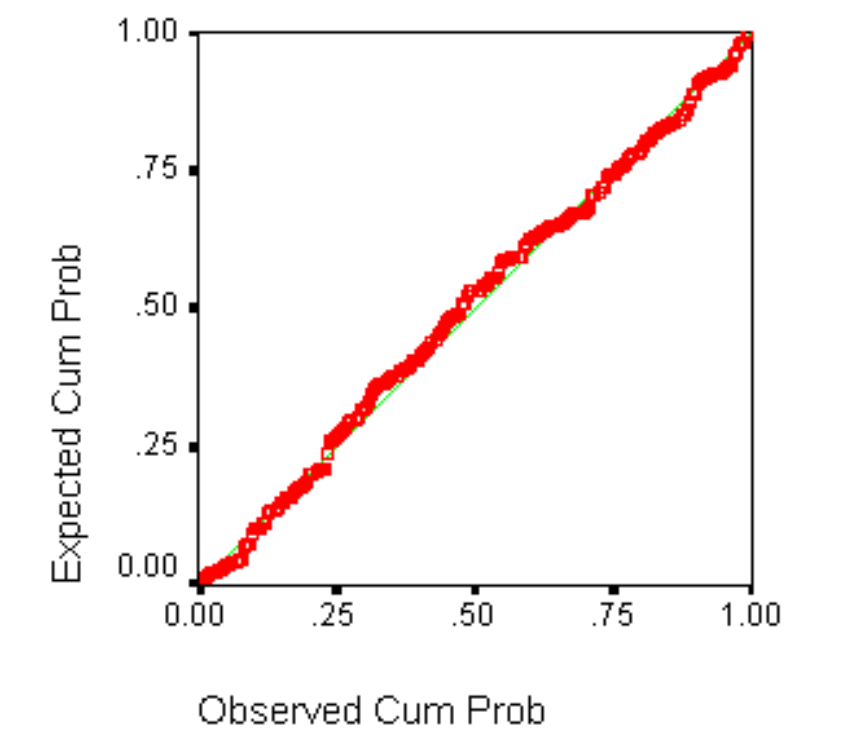
Visual Inspection of Residual Normality: Examine the distribution of residuals and check skew / kurtosis

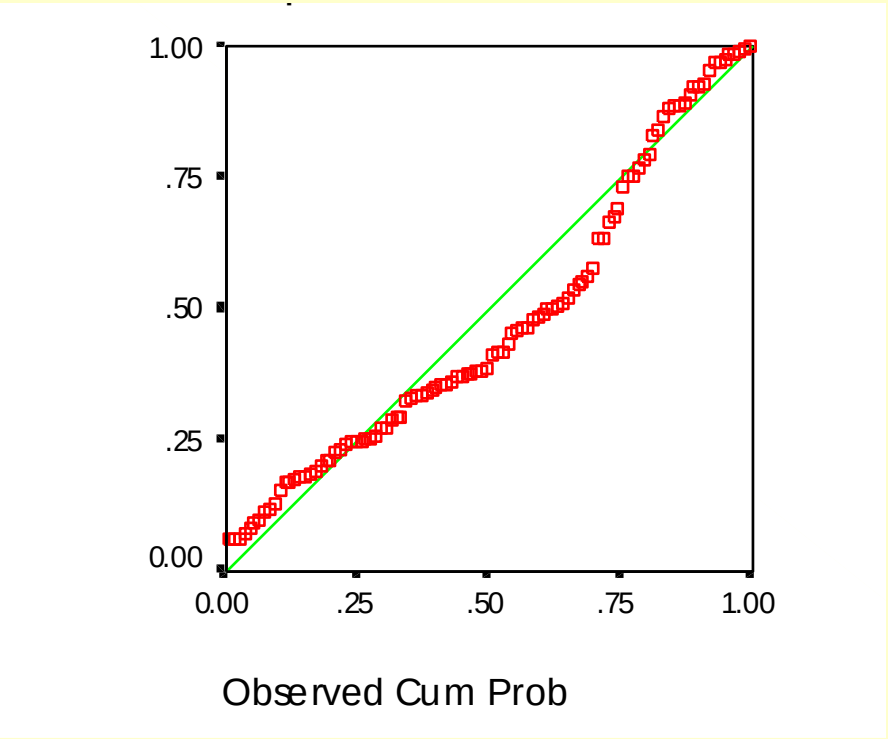




Normality of Errors: P-P Plots

Test for Residual Normality: Use P-P plots and statistical tests (KS or SW) to test for normality





Correlation vs Regression

Main Differences: Pearson correlation is birectional. Regression is not.

Pearson correlation does not measure the slope of best-fit line. Regression does.

For example:
Correlation coefficient of +1 does not mean that for one unit increase in one variable there is one unit increase in the other.