Linear Regression

Chapter 8

ANDY FIELD

Aims

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

What is Regression?

- A way of predicting the value of one variable from another.
  - It is a hypothetical model of the relationship between two variables.
  - The model used is a linear one.
  - Therefore, we describe the relationship using the equation of a straight line.

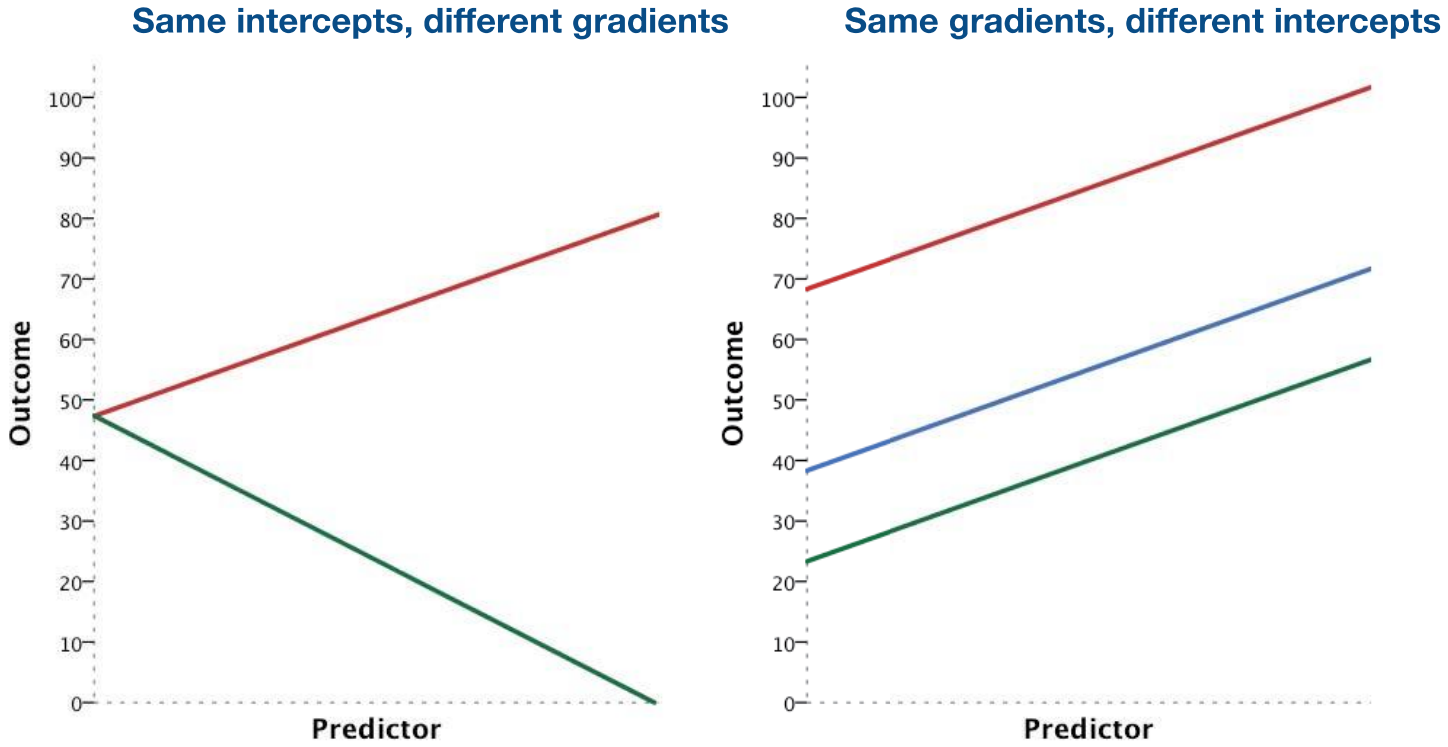
Describing a Straight Line

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

- b_i
  - Regression coefficient for the predictor
  - Gradient (slope) of the regression line
  - Direction/Strength of Relationship
- b_0
  - Intercept (value of Y when X = 0)
  - Point at which the regression line crosses the Y-axis (ordinate)

Intercepts and Gradients

FIGURE 8.2 Lines that share the same intercept but have different gradients, and lines with the same gradients but different intercepts



The Method of Least Squares

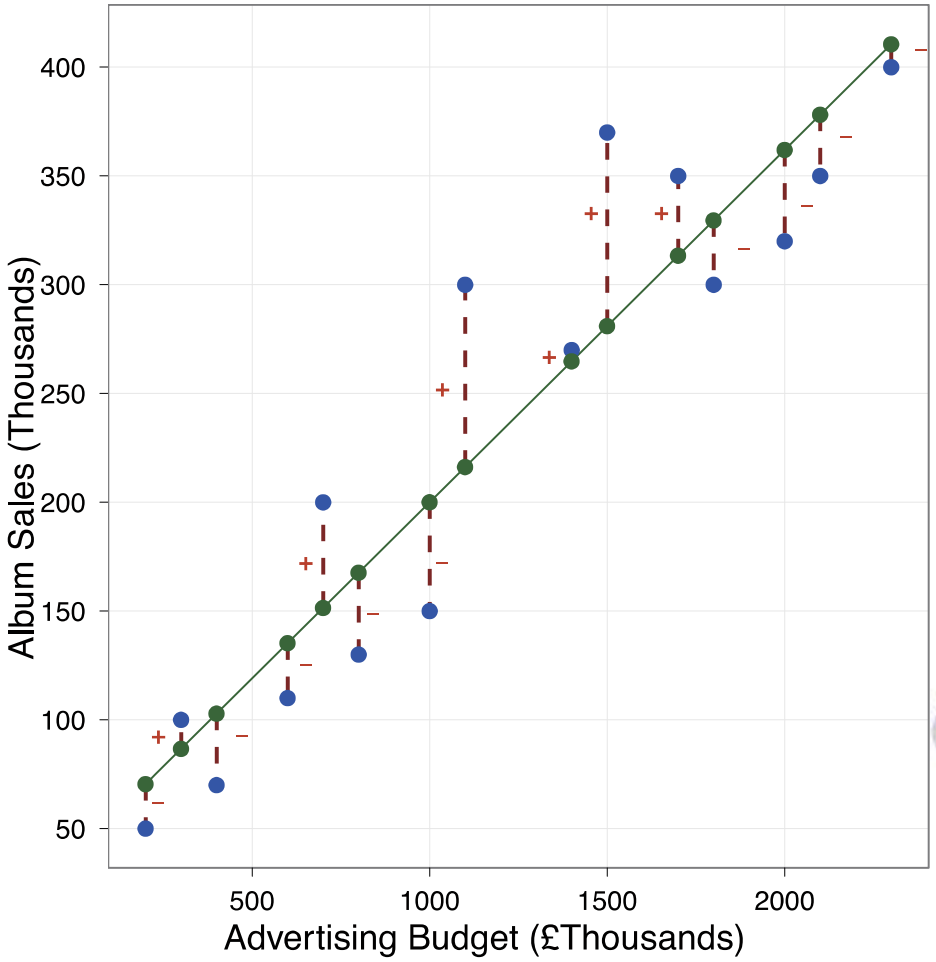
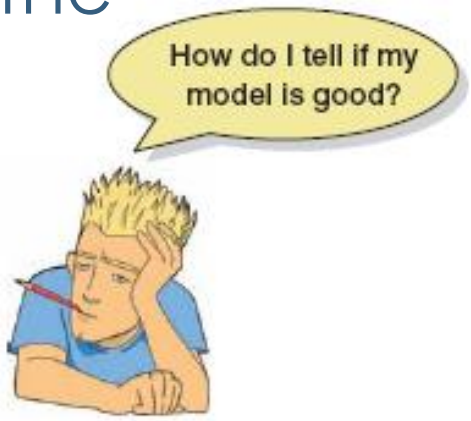


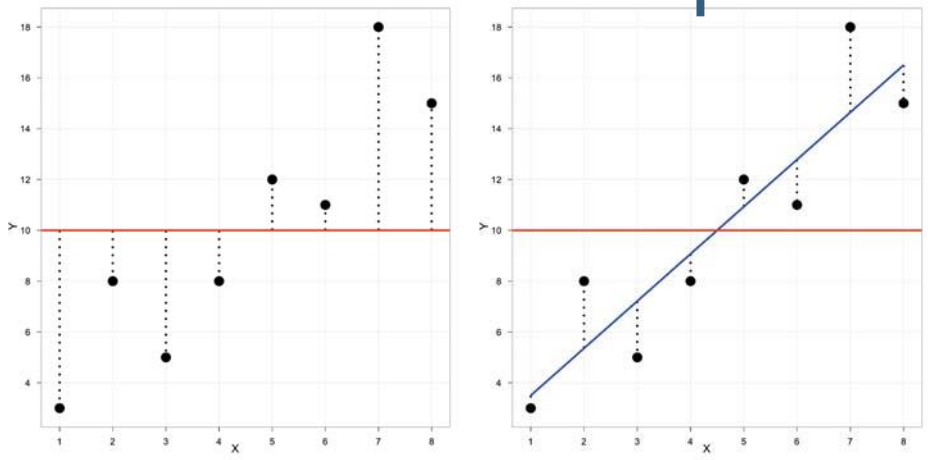
FIGURE 8.4 A scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

How Good is the Model?

- The regression line is only a model based on the data.
- This model might not reflect reality.
  - We need some way of testing how well the model fits the observed data.
  - How?

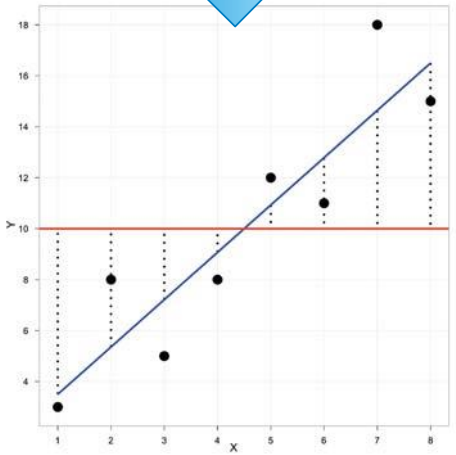


Sums of Squares



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

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



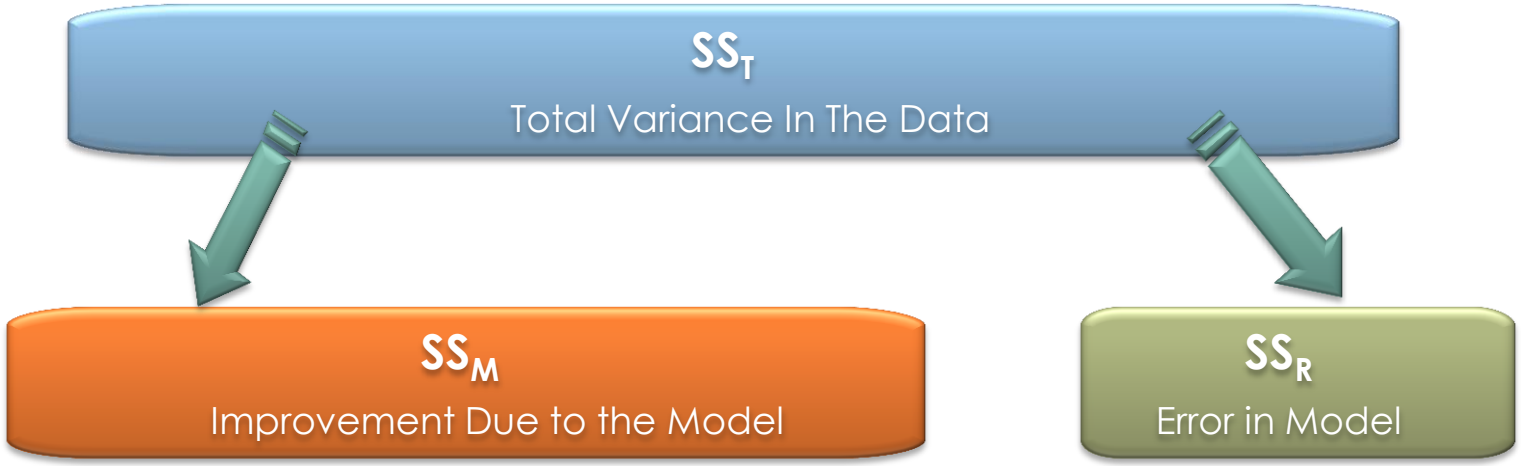
SS_M uses the differences between the mean value of *Y* and the regression line

FIGURE 8.5
Diagram showing from where the regression sums of squares derive

Summary

- SS_T
  - Total variability (variability between scores and the mean).
- SS_R
  - Residual/Error variability (variability between the regression model and the actual data).
- SS_M
  - Model variability (difference in variability between the model and the mean).

Testing the Model: ANOVA



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

Testing the Model: ANOVA

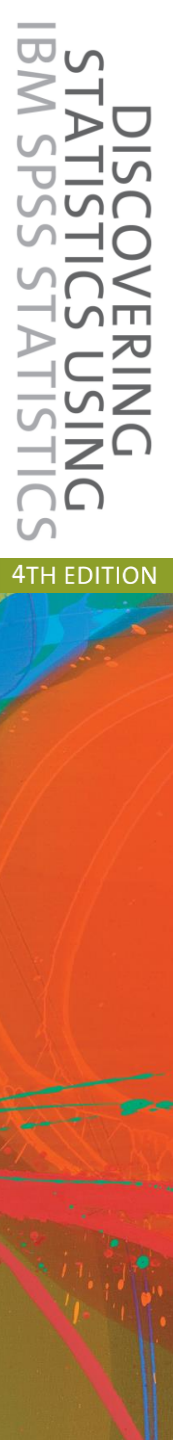
- Mean Squared Error
  - Sums of Squares are total values.
  - They can be expressed as averages.
  - These are called Mean Squares, MS

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

Testing the Model: R^2

- R^2
  - The proportion of variance accounted for by the regression model.
  - The Pearson Correlation Coefficient Squared

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



Regression: An Example

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

Step One: Graph the Data

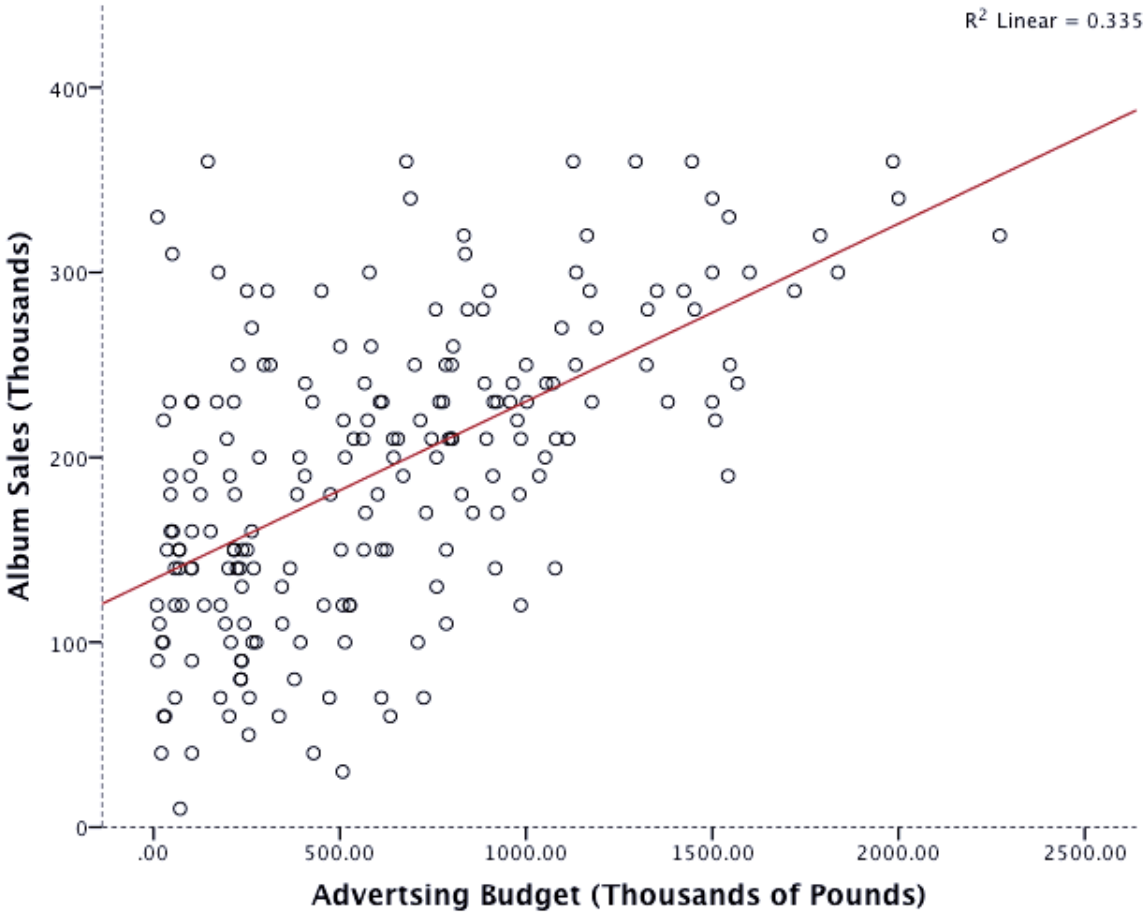


FIGURE 8.12 Scatterplot showing the relationship between album sales and the amount spent promoting the album

Regression Using IBM SPSS

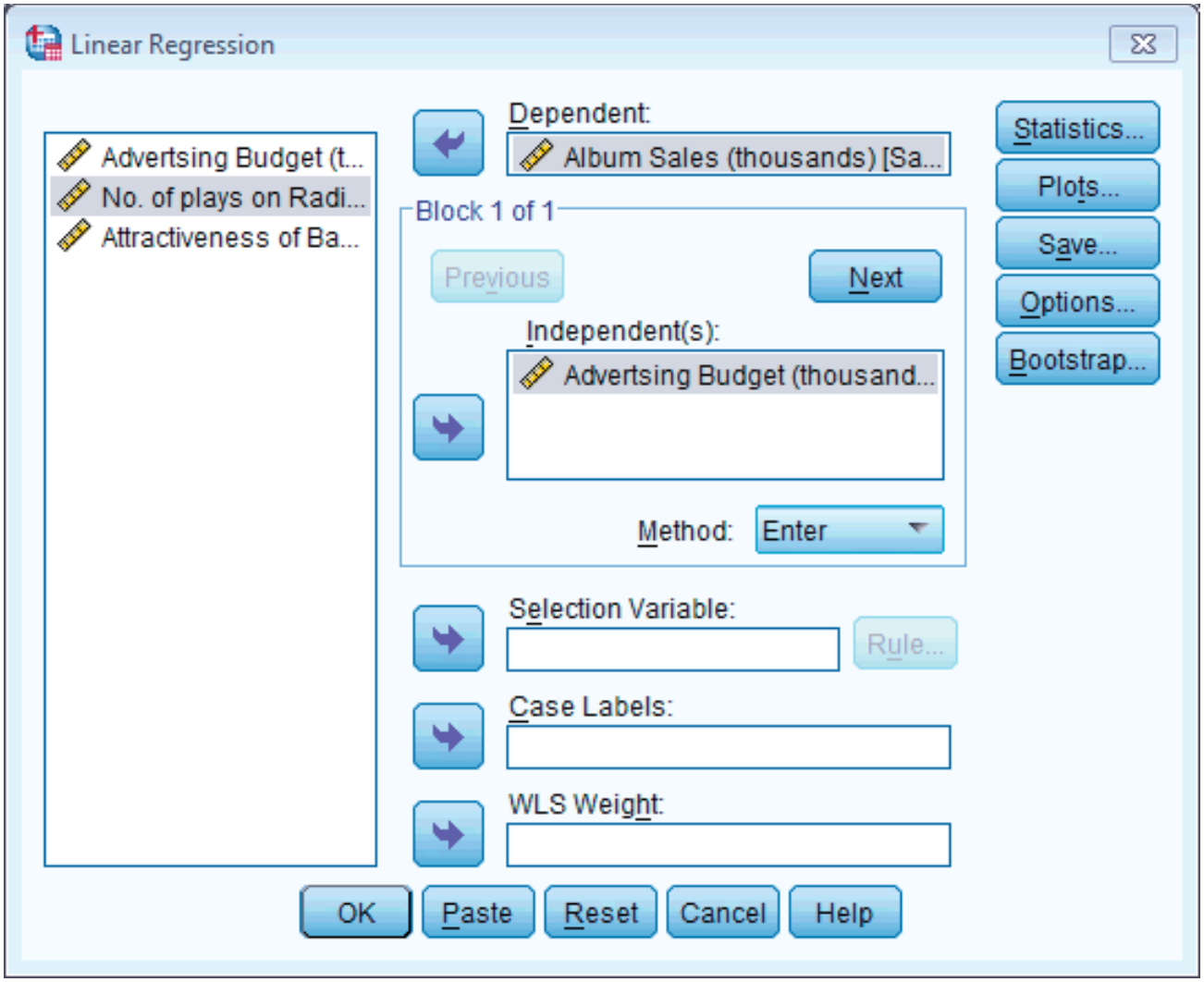


FIGURE 8.13
Main dialog box for regression

Output: Model Summary

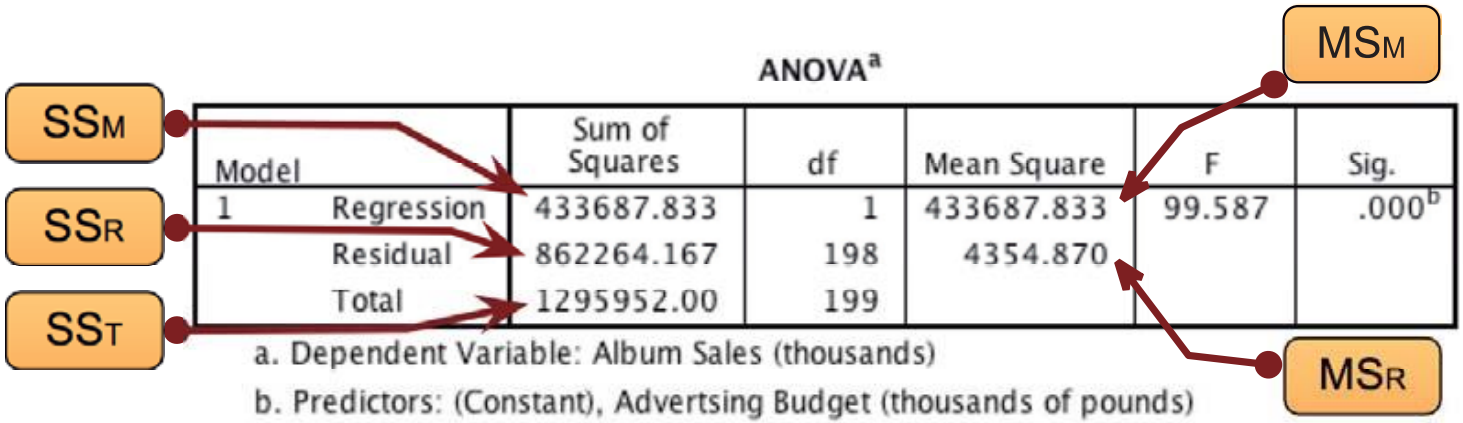
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.9914

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

Output: ANOVA

OUTPUT 8.2



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[b]
	Residual	862264.167	198	4354.870		
	Total	1295952.00	199			

a. Dependent Variable: Album Sales (thousands)

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

SPSS Output: Model Parameters

OUTPUT 8.3

Coefficients[a]

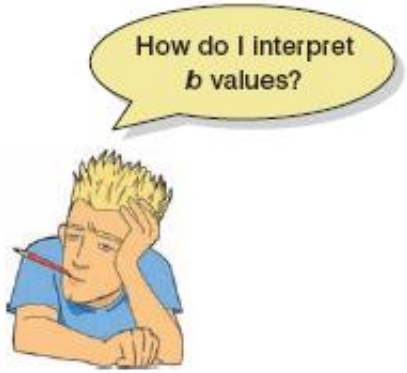
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134.140	7.537		17.799	.000
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000

a. Dependent Variable: Album Sales (thousands)

Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	134.140	.356	8.214	.001	117.993	151.258
	Advertsing Budget (thousands of pounds)	.096	.000	.009	.001	.080	.113

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples



Using The Model

$$\text{album sales}_i = b_0 + b_1 \text{advertising budget}_i$$

$$= 134.14 + \left(0.096 \times \text{ advertising budget}_i\right)$$

$$\text{album sales}_i = 134.14 + \left(0.096 \times \text{ advertising budget}_i\right)$$

$$= 134.14 + \left(0.096 \times 100\right)$$

$$= 143.74$$