Regression

Data files

- AlbumSales.sav
- Happiness.sav
- Homework: Mothering.sav

Introduction to Regression

- If two variables covary, we should be able to predict the value of one variable from another.
- Correlation only tells us how much two variables covary.
- In regression, we construct an equation that uses one or more variables (*the IV(s) or* ***predictor(s)***) to predict another variable (*the DV or* ***outcome***).
  - Predicting from one IV = **Simple Regression**
  - Predicting from multiple IVs = **Multiple Regression**

Simple Regression: The Model

- The general equation:

$$\text{Outcome}_i = (\text{model}) + \text{error}_i$$

- In regression the model is linear and we summarize a data set with a straight line.

- The regression line is determined through the **method of least squares**.

The Regression Line:

Model = "things that define the line we fit to the data"

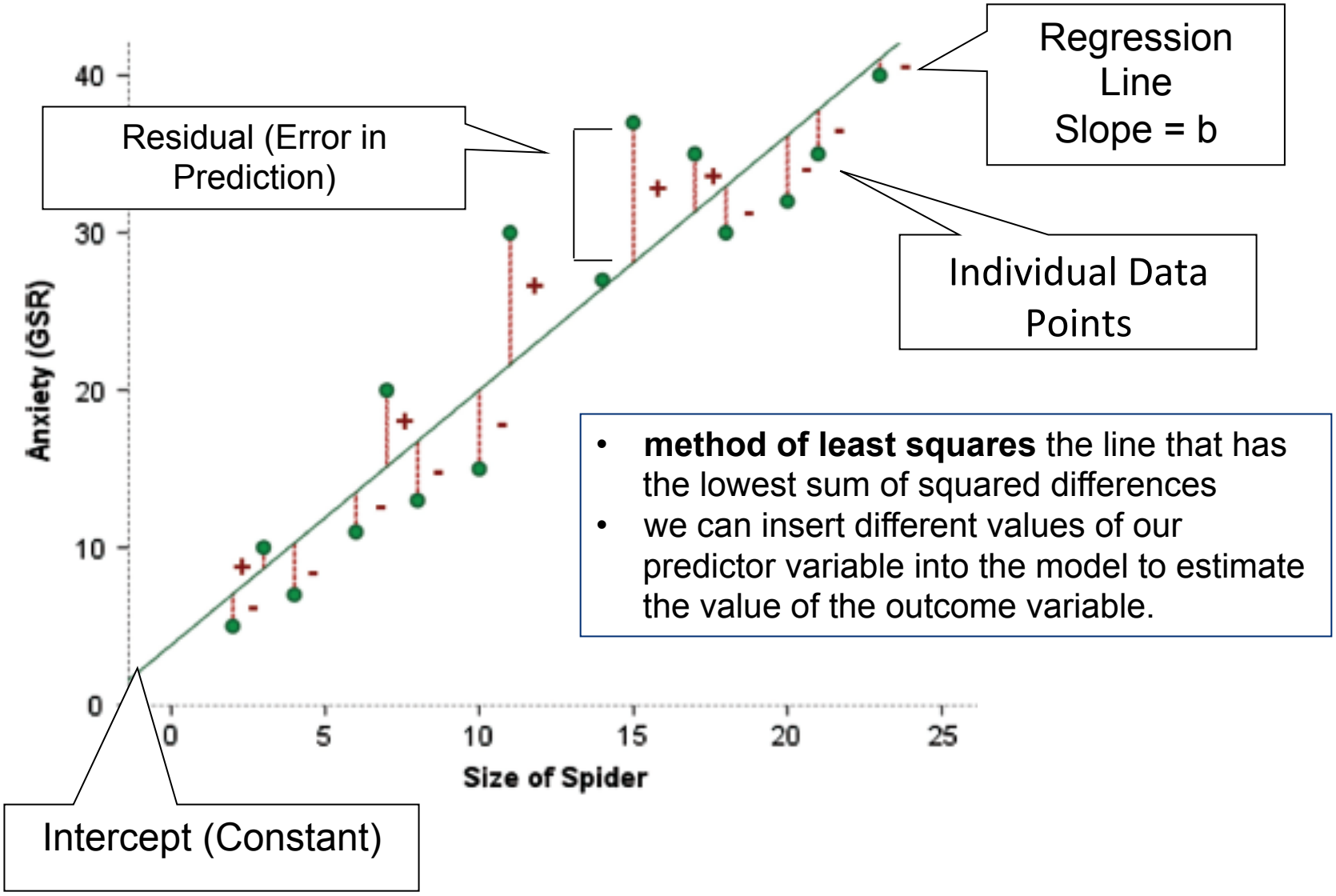
- Any straight line can be defined by:
  - The slope of the line (b)
  - The point at which the line crosses the vertical axis, termed the intercept of the line (a)

The general equation: $Outcome_i = (model) + error_i$

… becomes $Y_i = (\text{intercept} + bX_i) + \varepsilon_i$

- the intercept and b are termed **regression coefficients**
  - b tells us what the model looks like (it's shape)
  - the intercept tells us where the model is in geometric space
  - ε_i is the residual term and represents the difference between participant *i*'s predicted and obtained scores.

The line of best fit



- **method of least squares** the line that has the lowest sum of squared differences
- we can insert different values of our predictor variable into the model to estimate the value of the outcome variable.

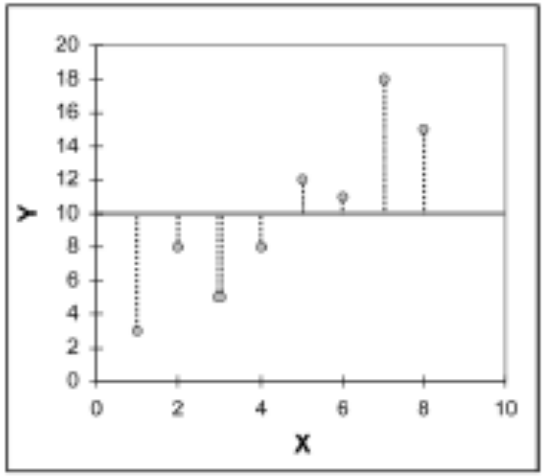
Assessing Goodness of Fit

- **goodness of fit:** improvement of fit compared to the **mean** (i.e., saying that the outcome = the mean whatever the value of the predictor).

- Let's consider an example (AlbumSales.sav):
  - A music mogul wants to know how many records her company will sell if she spends £100,000 on advertising.
  - In the absence of a model of the relationship between advertising and sales, the best guess would be the **mean** number of record sales -- regardless of amount of advertising.

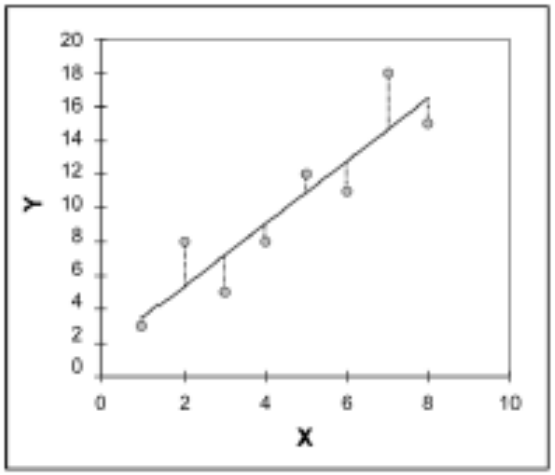
Assessing Goodness of Fit

SS_T



SS_T = the sum of squared differences between the observed data and the mean value of Y.

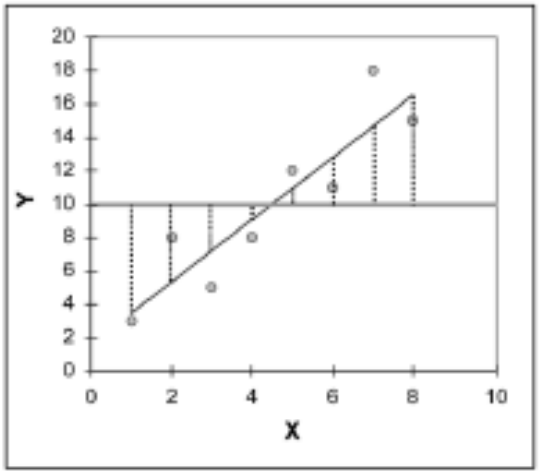
SS_R



The degree of inaccuracy (error) when the best model is fitted to the data.

SS_R = the sum of squared differences between the observed data and the regression line. (R: residual)

SS_M



The reduction in inaccuracy due to fitting the regression model.

SS_M = the sum of squared differences between the mean value of Y and the regression line.

Assessing Goodness of Fit

A large SS_M implies the regression model is much better than using the mean to predict the outcome variable.

How big is big? Assessed in two ways: (1) Via R^2 and (2) the *F-test* (assesses the ratio of systematic to unsystematic variance).

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

Represents the amount of variance in the outcome explained by the model relative to the total variance.

$$F = \frac{SS_M / df}{SS_R / df} = \frac{MS_M}{MS_R}$$

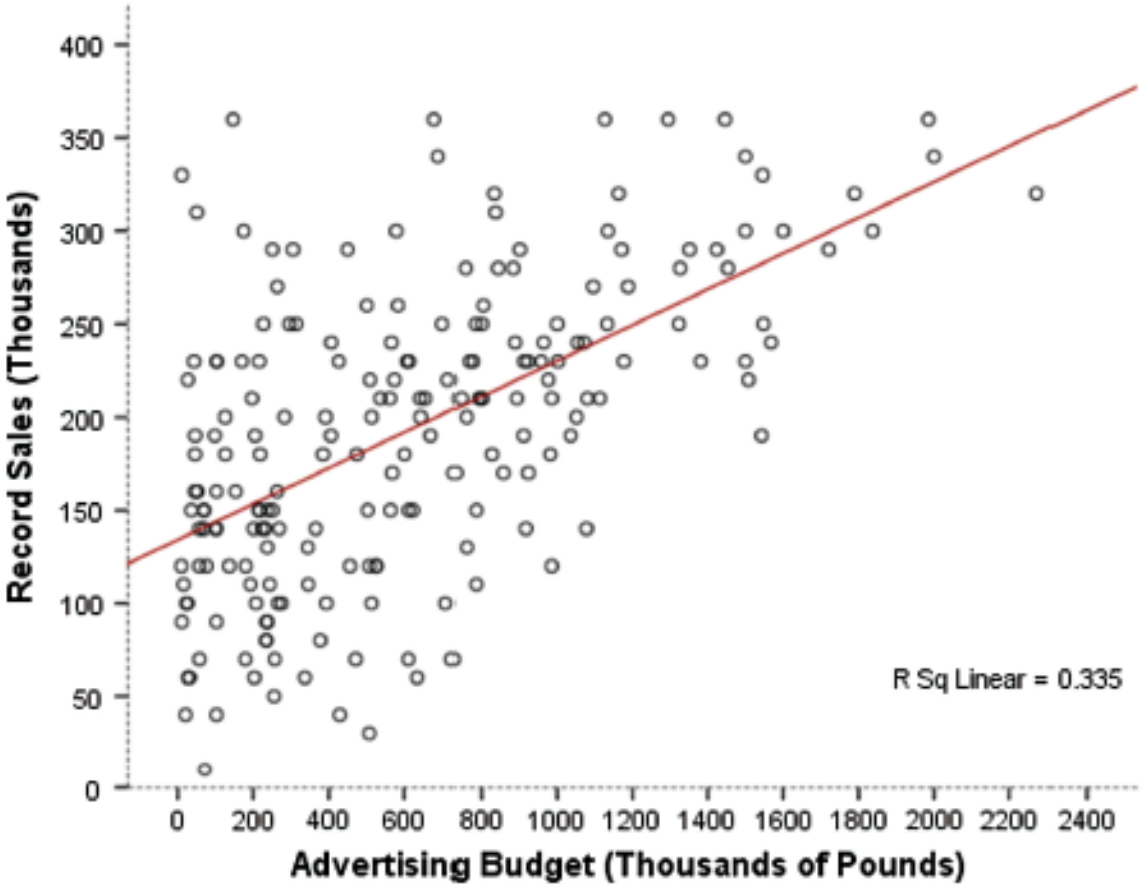
df for SS_M = number of predictors in the model

df for SS_R = number of participants - number of predictors - 1.

Simple Regression Using SPSS:

Predicting Record Sales (Y) from Advertising Budget (X)

AlbumSales.sav



Interpreting a Simple Regression:

Overall Fit of the Model

Advertising expenditure accounts for 33.5% of the variation in record sales.

R^2 adjusted: adjusted for number of predictors. Interesting only for multiple regression.

The square root of the average of the squared deviations about the regression line

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.991

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

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), Advertising Budget (thousands of pounds)

b. Dependent Variable: Record Sales (thousands)

SS_M

SS_R

SS_T

MS_M

MS_R

The significant "F" test allows us to conclude that the regression model results in significantly better prediction of record sales than the mean value of record sales.

Interpreting a Simple Regression: Model Parameters

t-tests: are the intercept and b significantly different from 0

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

the Y intercept

b, the slope, or the change in the outcome associated with a unit change in the predictor

The ANOVA tells us whether the **overall** model results in a significantly good prediction of the outcome variable.

constant = 134.14. Tells us that when no money is spent on ads, the model predicts 134,140 records will be sold.

b = .096. The amount of change in the outcome associated with a unit change in the predictor. Thus, we can predict 96 extra record sales for every £1000 in advertising.

Interpreting a Simple Regression: Model Parameters

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: Record Sales (thousands)

Unstandardized Regression Weights

Y = intercept + bX

Intercept and Slope are in original units of X and Y and so aren't directly comparable

Standardized Regression Weights

Standardized regression weights tell us the number of standard deviations that the outcome will change as a result of one standard deviation change in the predictor.

Richards. (1982). Standardized versus Unstandardized Regression Weights. Applied Psychological Measurement, 6, 201-212.

Interpreting a Simple Regression: Using the Model

Since we've demonstrated the model significantly improves our ability to predict the outcome variable (record sales), we can plug in different values of the predictor variable(s).

$$\text{record sales}_i = \text{intercept} + b \times \text{advertising budget}_i$$

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

What could the record executive expect if she spent £500,000 on advertising? How about £1,000,000?

MULTIPLE REGRESSION

Multiple Regression

- AlbumSales.sav - additional predictors for predicting album sales:
    - advertising budget
    - radio play (hours)
    - attractiveness of artist (scale of 1 to 10)

Multiple Regression

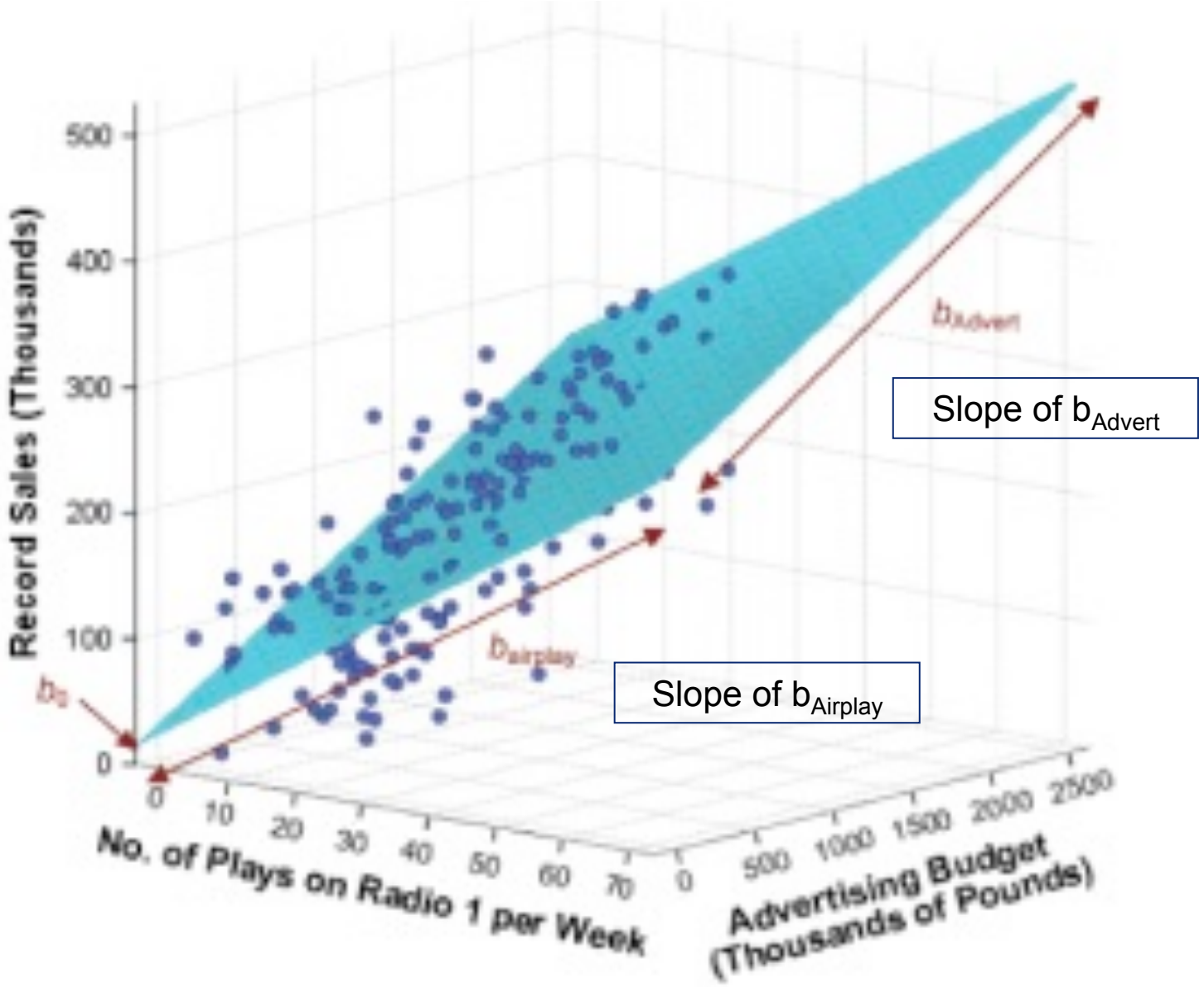
- 1 Continuous DV (Outcome Variable)
- 2 or more Quantitative IVs (Predictors)
- General Form of the Equation:

$$outcome_i = (model) + error_i$$

$$Y_{pred} = (intercept + b_1X_1 + b_2X_2 + \ldots + b_nX_n) + \varepsilon_I$$

$$record\ sales_{pred} = intercept + b_1 ad\ budget_i + b_2 airplay_i + \varepsilon$$

record sales, advertising budget and radio play



Partitioning the Variance:

Sums of Squares, R, and R^2

SS_T Represents the total amount of differences between the observed values and the mean value of the outcome variable.

SS_R Represents the degree of inaccuracy when the best model is fitted to the data. SS_R uses the differences between the observed data and the regression line.

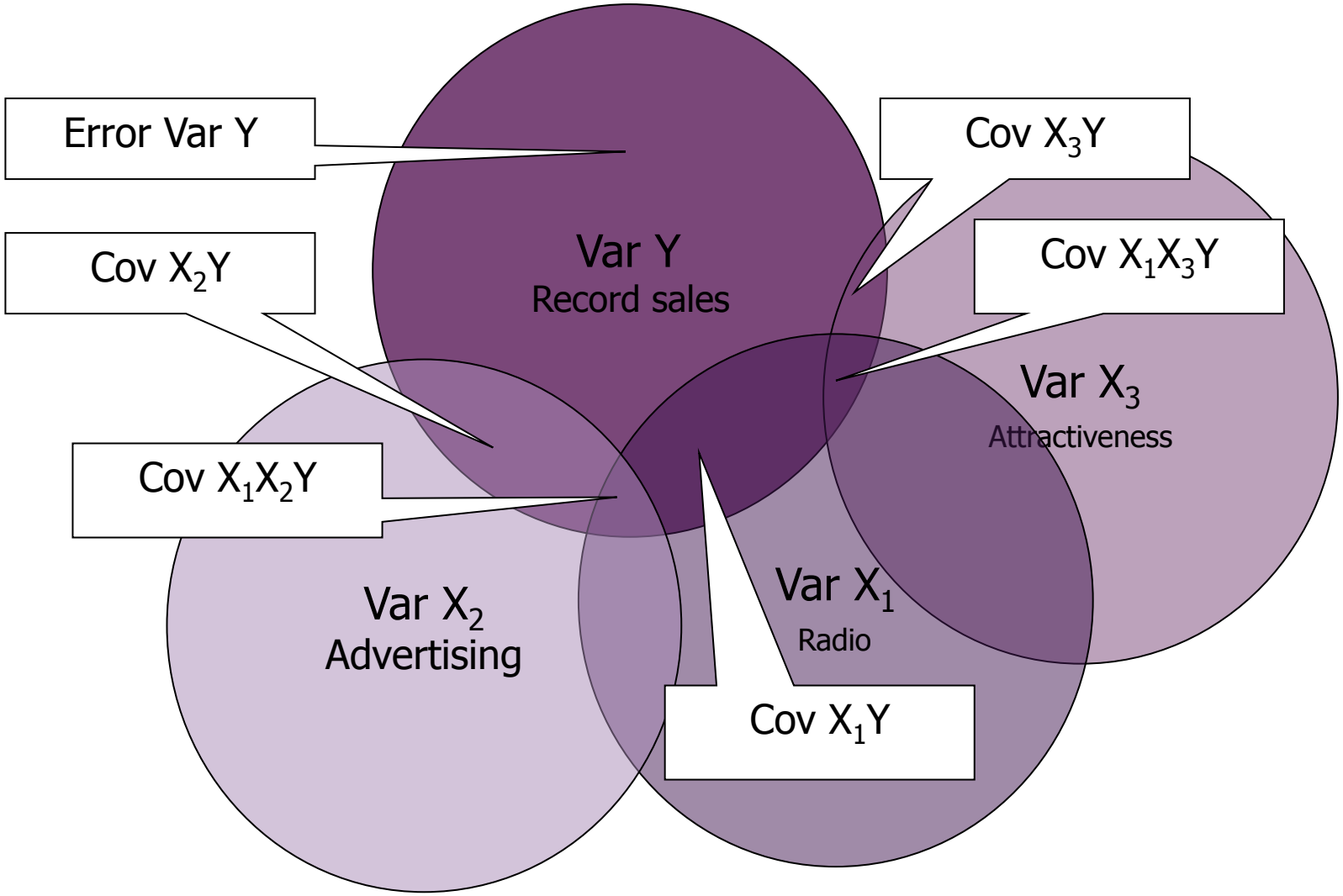
SS_M Shows the reduction in inaccuracy resulting from fitting the regression model to the data. SS_M uses the differences between the values of Y predicted by the model (the regression line) and the mean. A large SS_M implies the regression model predicts the outcome variable better than the mean.

R The correlation between the observed values of Y (outcome variable) and values of Y predicted by the multiple regression model. It is a gauge of how well the model predicts the observed data. R^2 is the amount of variation in the outcome variable accounted for by the model.

Variance Partitioning

Variance in the outcome variable is due to action of all predictors plus some error:

Covariation



Partial Statistics

- Partial correlations describe the *independent effect* of the predictor on the outcome, controlling for the effects of all other predictors

Part (semi-Partial) Statistics

- Part (semi-partial) r
  - Effect of other predictors are NOT held constant.
  - Semi-partial r's indicate the *marginal (additional/unique) effect* of a particular predictor on the outcome.

Methods of Regression:

Predictor Selection and Model Entry Rules

- Selecting Predictors
  - More is not better! Select the most important ones based on past research findings.
- Entering variables into the Model
  - When predictors are uncorrelated order makes no difference.
  - Rare to have completely uncorrelated variables, so method of entry becomes crucial.

Methods of Regression

- Forced entry (Enter)
  - All predictors forced into model simultaneously.
- Hierarchical (blockwise entry)
  - Predictors selected and entered by researcher based on knowledge of their relative importance in predicting the outcome – most important first.
- Stepwise (mathematically determined entry)
  - Forward method
  - Backward method
  - Stepwise method

Forced Entry (*Enter*)

- All predictors forced into the model simultaneously.
- Default option
- Method most appropriate for testing theory
  (Studenmund Cassidy, 1987)

Hierarchical / Blockwise Entry

- Researcher decides order.
- Known predictors usually entered first, in order of their importance in predicting the outcome.
- Additional predictors are added in further blocks.

Stepwise Entry: Forward Method

Procedure

1. SPSS selects predictor with the highest simple correlation with the outcome variable.
2. Subsequent predictors selected on the basis of the size of their semi-partial correlation with the outcome variable.
3. Process repeated until all predictors that contribute significant unique variance to the model have been included in the model.

Stepwise Entry: Backward Method

Procedure

1. SPSS places all predictors in the model and then computes the contribution of each one.
2. Predictors with less than a given level of contribution are removed. (In SPSS the default probability to eliminate a variable is called pout = p ≥ 0.10. (*probability out*).
3. SPSS re-estimates the regression equation with the remaining predictor variables. Process repeats until all the predictors in the equation are statistically significant, and all outside the equation are not.

Stepwise Entry: Stepwise Method

Procedure

1. Same as the Forward method, except that each time a predictor is added to the equation, a removal test is made of the least useful predictor.

2. The regression equation is constantly reassessed to see whether any redundant predictors can be removed.

Checking Assumptions: basics

- **Variable Types:** predictors must be quantitative or binary categorical; outcome must be quantitative.
- **Non-zero variance**: Predictors must have some variation.

Checking assumptions: using scatter plots

- **No perfect collinearity:** Predictors should not correlate too highly.
  - Easiest to test by drawing scatterplots with all possible pairs of predictors
  - Can be tested with the VIF (variance inflation factor). Values over 8 are worrisome. (Regression > Linear > Statistics > Collinearity diagnostics)
- **Normally distributed errors**
  - Save residuals and test normality (Regression > Linear > Save)
  - identify outliers

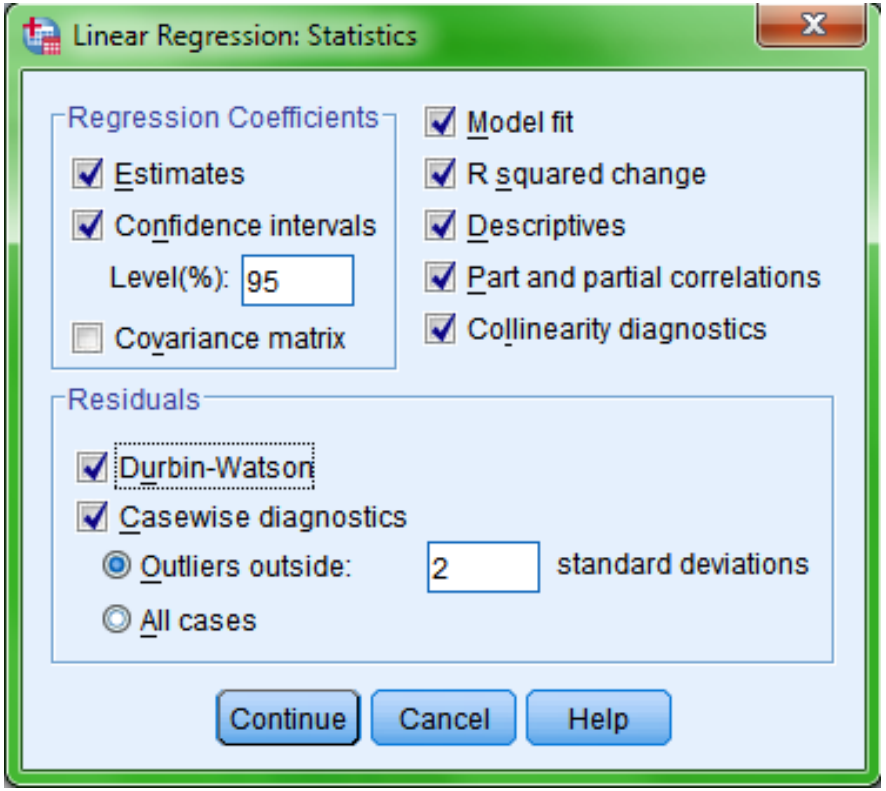
Checking assumptions: using scatter plots

- **No outliers or extreme residuals**: outliers and very large residuals can distort the results
  - use scatter plot; save standardised residuals (Regression > Linear > Save) and identify outliers
  - Or use Casewise diagnostics, which finds cases with unusual residuals
  - Use Standardised DFFit to identify influential cases: shows the difference (in SDs) between the regression coefficients with and without a case (Regression > Linear > Save) – a value larger than 1 is suspicious
  - if there are outliers or influential cases, it is safest to use bootstrapping
  - or the outlier can be deleted/replaced and the regression rerun

For Time-Series data

- **Independent errors**: The residual terms for any two observations should be independent (uncorrelated)
  - Tested with the **Durbin-Watson test**, which ranges from 0 to 4. Value of 2 means residuals are uncorrelated. Values greater than 2 indicate a negative correlation; values below 2 indicate a positive correlation. (Regression > Linear > Statistics)

Estimates: Provides estimated coefficients of the regression model, test statistics and their significance.
Confidence Intervals: Useful tool for assessing likely value of the regression coefficients in the population.
Model Fit: Omnibus test of the model's ability to predict the DV.
R-squared Change: Change in R^2 resulting from inclusion of a new predictor.



Collinearity diagnostics: To what extent the predictors correlate. (VIF < 8)
Durbin-Watson: Tests the assumption of independent errors. (should be 2)
Case-wise diagnostics: Lists cases with unusual residuals

Bootstrapping

- Confidence intervals
- Neutralises the effects of outliers, etc.

Multiple Regression: Model Summary

Increases only if the addition of the predictor has a higher-than-chance effect

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.578[a]	.335	.331	65.991	.335	99.587	1	198	.000	
2	.815[b]	.665	.660	47.087	.330	96.447	2	196	.000	1.950

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

b. Predictors: (Constant), Advertsing Budget (thousands of pounds), Attractiveness of Band, No. of plays on Radio 1 per week

c. Dependent Variable: Record Sales (thousands)

Should be close to 2; less than 1 or greater than 3 poses a problem.

ANOVA[c]

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			
2	Regression	861377.418	3	287125.806	129.498	.000[b]
	Residual	434574.582	196	2217.217		
	Total	1295952.000	199			

We have two models, with different IVs

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

b. Predictors: (Constant), Advertsing Budget (thousands of pounds), Attractiveness of Band, No. of plays on Radio 1 per week

c. Dependent Variable: Record Sales (thousands)

Shows whether the IVs combined have a significant contribution in the model

Multiple Regression: Model Parameters

Note the difference between unstandardised and standardised coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	134.140	7.537		17.799	.000	119.278	149.002
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000	.077	.115
2	(Constant)	-26.613	17.350		-1.534	.127	-60.830	7.604
	Advertsing Budget (thousands of pounds)	.085	.007	.511	12.261	.000	.071	.099
	No. of plays on Radio 1 per week	3.367	.278	.512	12.123	.000	2.820	3.915
	Attractiveness of Band	11.086	2.438	.192	4.548	.000	6.279	15.894

a. Dependent Variable: Record Sales (thousands)

The t-tests show whether each variable has a significant contribution when the other IVs are controlled for

Writing up the results

Table + description:

	b	Beta	t	p
Step 1				
Constant				
Advert				
Step 2				
Constant				
Advert				
Radio				
Attract				

A linear regression model was built to predict record sales from advertising budget, hours of radio play of songs and attractiveness of the band. Advertising budget was entered first followed by the remaining two predictors. The first model with just advertising budget entered gave a significant result ($R^2 = .335$, $F(1, 198) = 99.59$, $p < .001$). The second model with all three predictors was also significant ($R^2 = .665$, $F(3, 196) = 129.5$, $p < .001$) with each of the predictors having a significant contribution. The contributions of advertising budget and radio play were the highest (about .5 SD increase in record sales), while the attractiveness of the band had a smaller effect (.19 SD increase in record sales). The three predictors together explained about 66% of the variance in record sales.

Exercise: happiness

- Correlations, outliers, etc.
- Which predictors first?
- Run the analysis.

Homework:
Mothering.sav

A study was carried out to explore the relationship between self-confidence in being a mother and maternal care as a child (Leerkes and Crockenberg 1999). It was also hypothesised that self-esteem would have an effect.

Potential predictor variables measured were:

Maternal Care (high score = better care)
Self-esteem (high score = higher self-esteem)

Enter Maternal Care first and Self-esteem in the next model.

Run the analysis. Write up the results.