DISCOVERING STATISTICS USING SPSS
THIRD EDITION

Multiple Regression

Lecture 07

ANDY FIELD

Aims

- Understand When To Use Multiple Regression.
- Understand the multiple regression equation and what the betas represent.
- Understand Different Methods of Regression
  - Hierarchical
  - Stepwise
  - Forced Entry
- Understand How to do a Multiple Regression on PASW/SPSS
- Understand how to Interpret multiple regression.
- Understand the Assumptions of Multiple Regression and how to test them

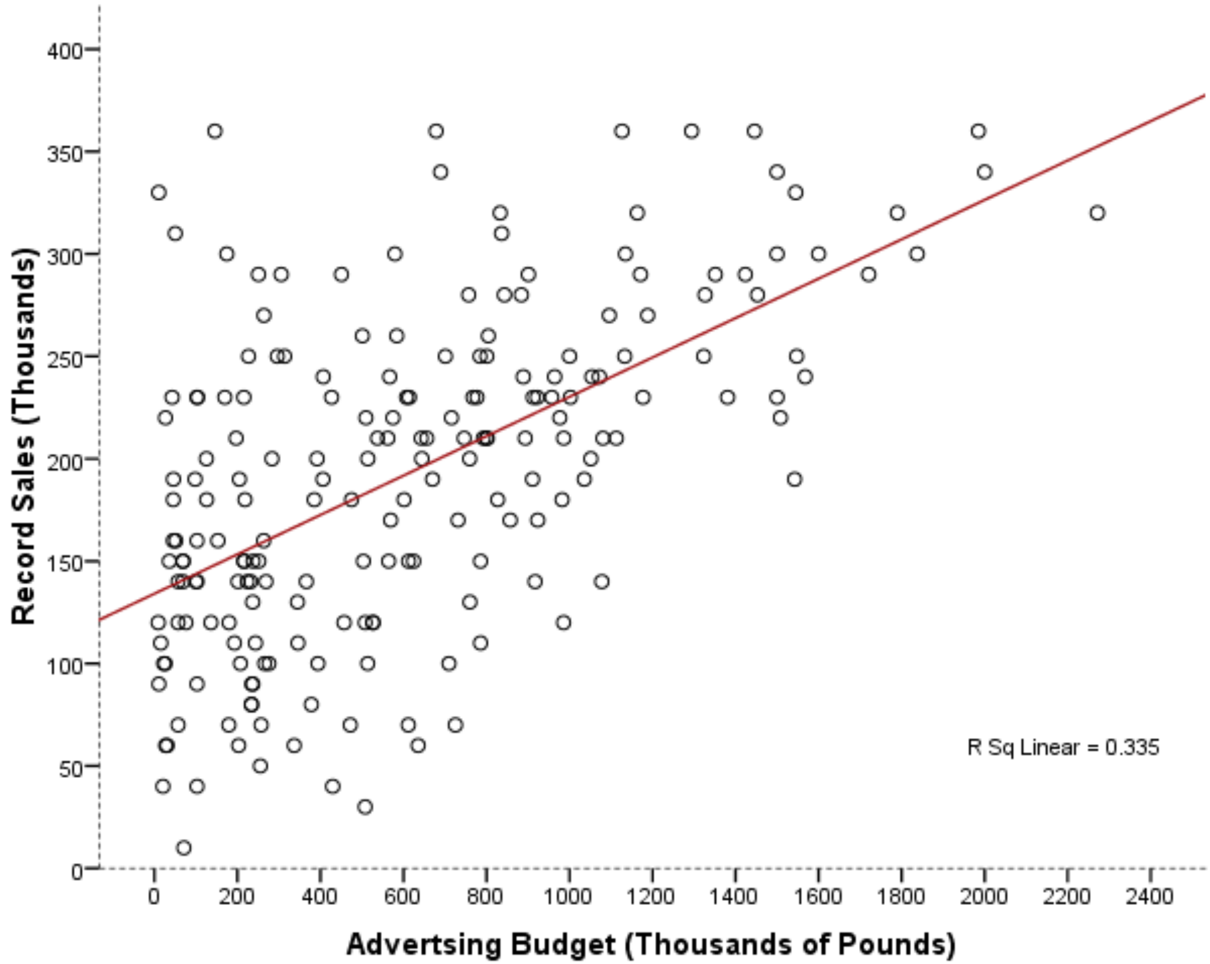
What is Multiple Regression?

- Linear Regression is a model to predict the value of one variable from another.
- Multiple Regression is a natural extension of this model:
  - We use it to predict values of an outcome from *several* predictors.
  - It is a hypothetical model of the relationship between several variables.

Regression: An Example

- A record company boss was interested in predicting record sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and Downloads) in the week after release
- Predictor variables
  - The amount (in £s) spent promoting the record before release (see last lecture)
  - Number of plays on the radio (new variable)

The Model with One Predictor



Multiple Regression as an Equation

- With multiple regression the relationship is described using a variation of the equation of a straight line.

$$y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + \varepsilon_i$$

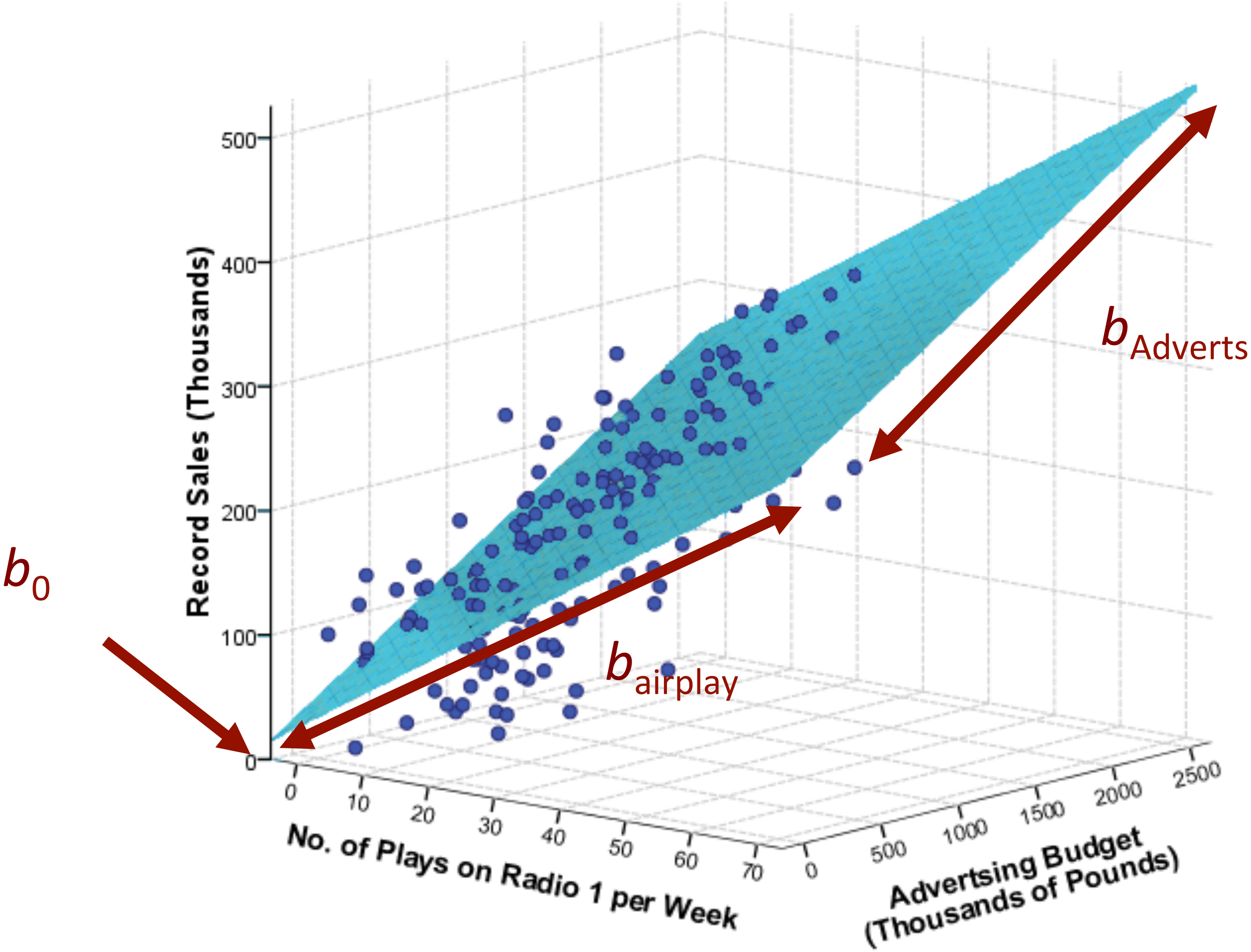
b_0

- b_0 is the intercept.
- The intercept is the value of the Y variable when all Xs = 0.
- This is the point at which the regression plane crosses the Y-axis (vertical).

Beta Values

- b_1 is the regression coefficient for variable 1.
- b_2 is the regression coefficient for variable 2.
- b_n is the regression coefficient for n^{th} variable.

The Model with Two Predictors



DISCOVERING STATISTICS USING SPSS
THIRD EDITION

ANDY FIELD

Methods of Regression

- Hierarchical:
  - Experimenter decides the order in which variables are entered into the model.
- Forced Entry:
  - All predictors are entered simultaneously.
- Stepwise:
  - Predictors are selected using their semi-partial correlation with the outcome.


Which method of regression should I use?

Hierarchical Regression

- Known predictors (based on past research) are entered into the regression model first.
- New predictors are then entered in a separate step/block.
- Experimenter makes the decisions.

Hierarchical Regression

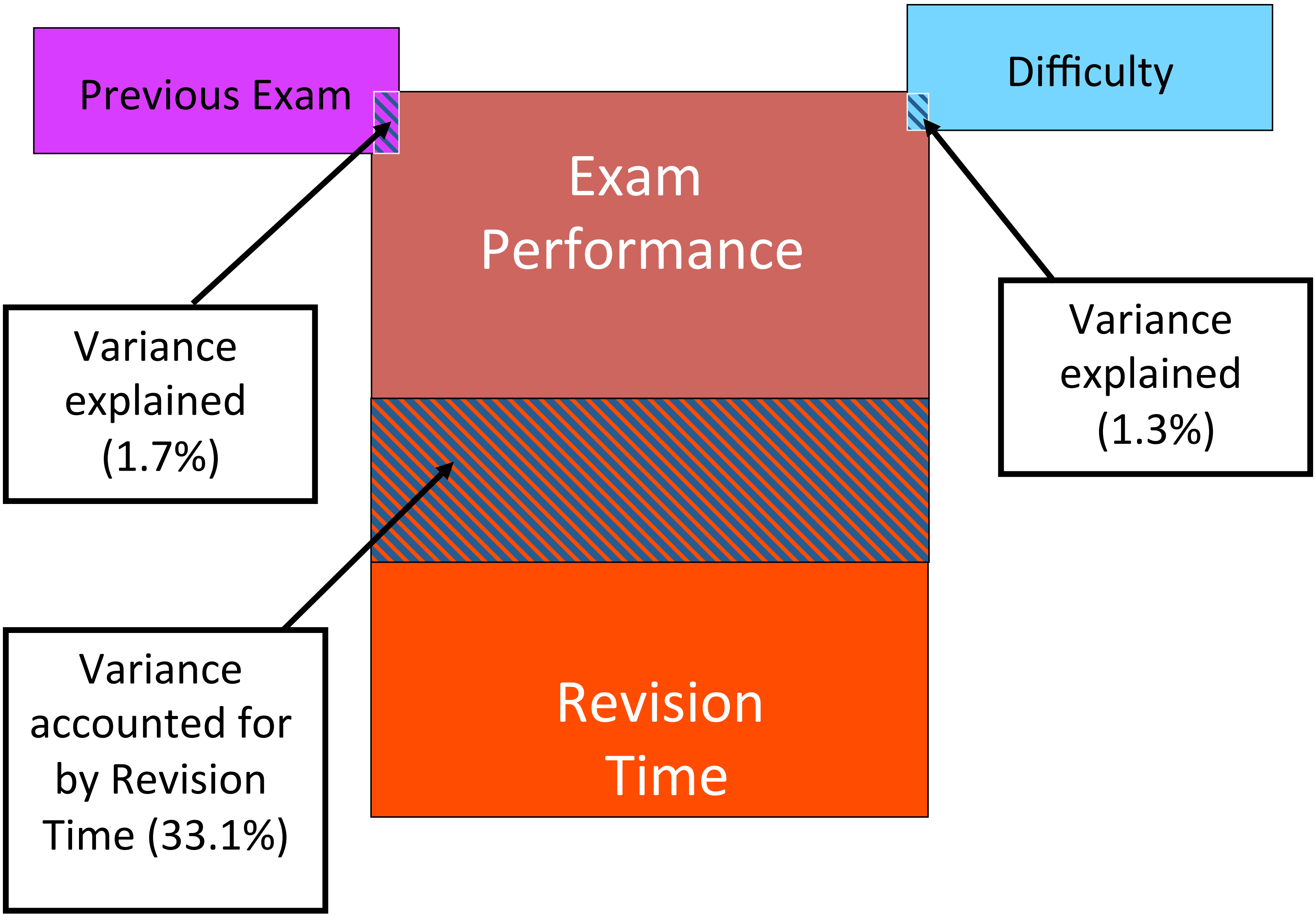
- It is the best method:
  - Based on theory testing.
  - You can see the unique predictive influence of a new variable on the outcome because known predictors are held constant in the model.
- Bad Point:
  - Relies on the experimenter knowing what they're doing!

Forced Entry Regression

- All variables are entered into the model simultaneously.
- The results obtained depend on the variables entered into the model.
  - It is important, therefore, to have good theoretical reasons for including a particular variable.

Stepwise Regression I

- Variables are entered into the model based on mathematical criteria.
- Computer selects variables in steps.
- Step 1
  - SPSS looks for the predictor that can explain the most variance in the outcome variable.


Previous Exam
Difficulty
Exam Performance
Variance explained (1.7%)
Variance explained (1.3%)
Variance accounted for by Revision Time (33.1%)
Revision Time

Correlations

		Exam Mark (%)	First Year Exam Grade	Difficulty of Question	Revision Time
Exam Mark (%)	Pearson Correlation	1.000	.130**	-.113*	.575**
	Sig. (2-tailed)	.	.005	.014	.000
	N	474	474	474	474
First Year Exam Grade	Pearson Correlation	.130**	1.000	[illegible]	.047
	Sig. (2-tailed)	.005	.	[illegible]	.303
	N	474	474	474	474
Difficulty of Question	Pearson Correlation	-.113*	-.012	1.000	-.236**
	Sig. (2-tailed)	.014	[illegible]	.	.000
	N	474	474	474	474
Revision Time	Pearson Correlation	.575**	.047	-.236**	1.000
	Sig. (2-tailed)	.000	.303	.000	.
	N	474	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

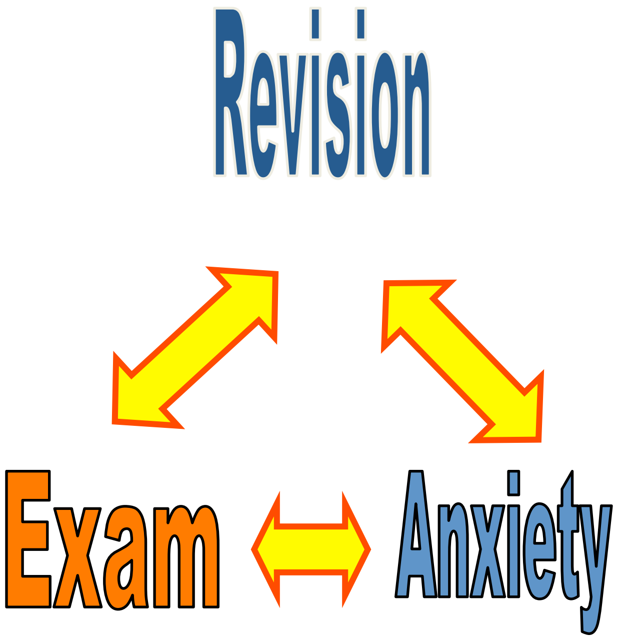
*. Correlation is significant at the 0.05 level (2-tailed).

Stepwise Regression II

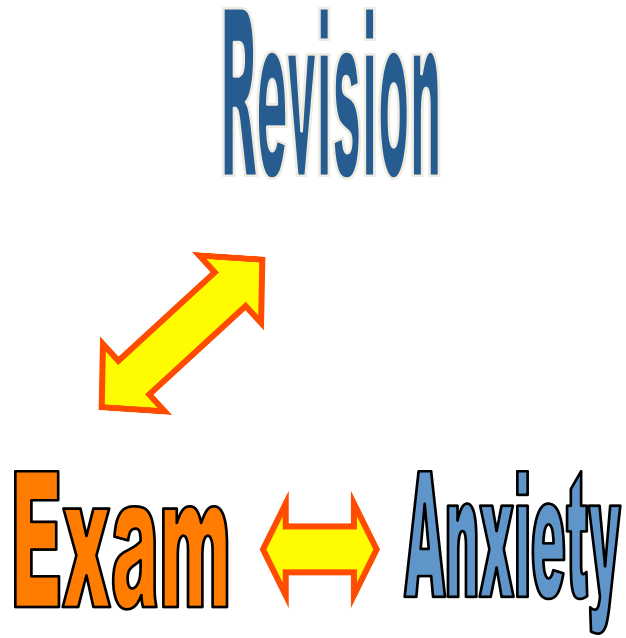
- Step 2:
  - Having selected the 1st predictor, a second one is chosen from the remaining predictors.
  - The *semi-partial correlation* is used as a criterion for selection.

Semi-Partial Correlation

- Partial correlation:
  - measures the relationship between two variables, controlling for the effect that a third variable has on them both.
- A semi-partial correlation:
  - Measures the relationship between two variables controlling for the effect that a third variable has on only one of the others.



Partial Correlation



Semi-Partial Correlation

Semi-Partial Correlation in Regression

- The semi-partial correlation
  - Measures the relationship between a predictor and the outcome, controlling for the relationship between that predictor and any others already in the model.
  - It measures the *unique contribution* of a predictor to explaining the variance of the outcome.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-10.863	3.069		-3.539	.000
	Revision Time	3.400	.222	.575	15.282	.000
2	(Constant)	-24.651	5.858		-4.208	.000
	Revision Time	3.371	.221	.570	15.241	.000
	First Year Exam Grade	.175	.063	.103	2.756	.006

a. Dependent Variable: Exam Mark (%)

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	First Year Exam Grade	.103[a]	2.756	.006	.126	.998
	Difficulty of Question	.024[a]	.624	.533	.029	.944
2	Difficulty of Question	.024[b]	.631	.528	.029	.944

a. Predictors in the Model: (Constant), Revision Time

b. Predictors in the Model: (Constant), Revision Time, First Year Exam Grade

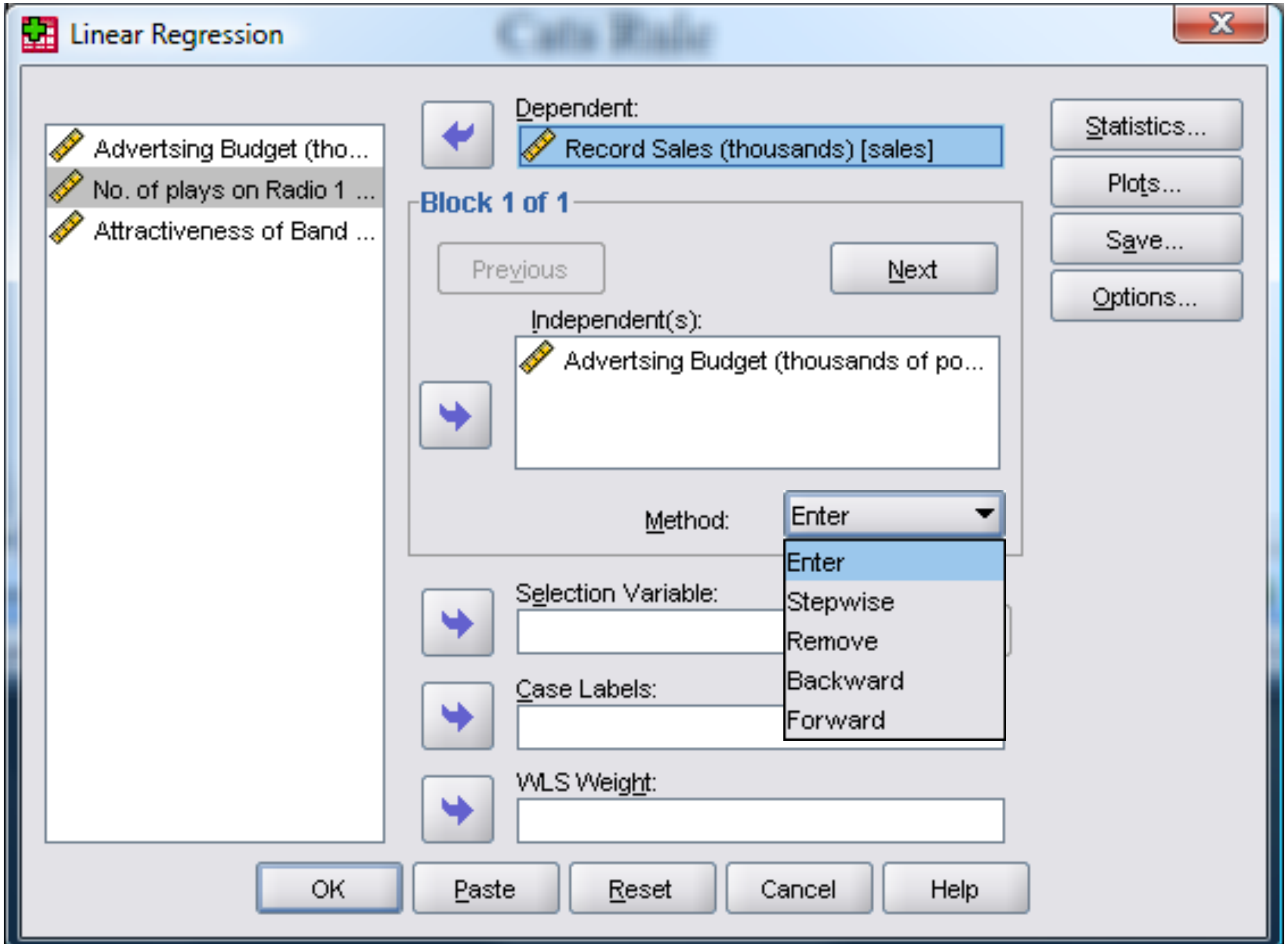
c. Dependent Variable: Exam Mark (%)

DISCOVERING STATISTICS USING SPSS
THIRD EDITION

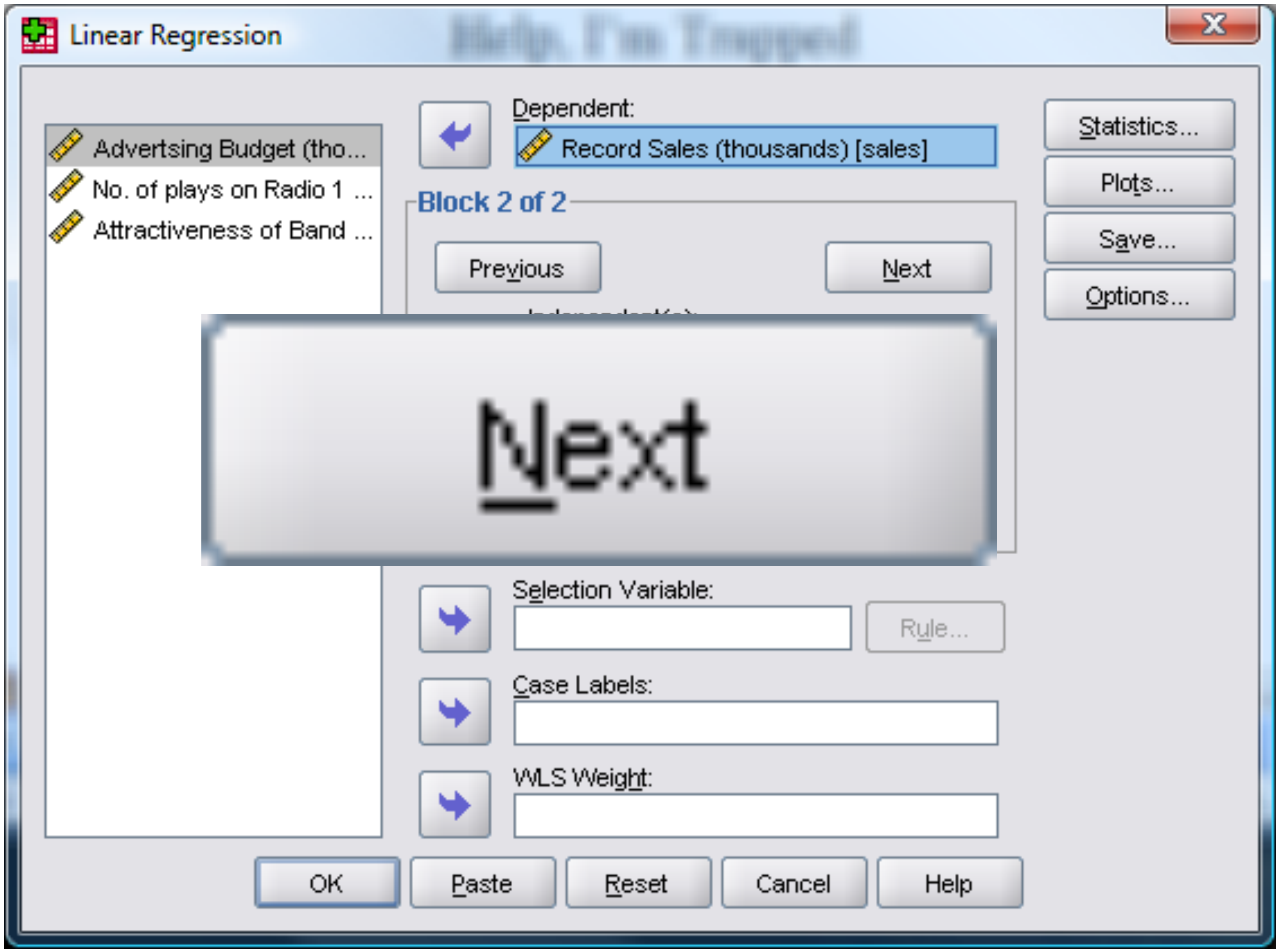
Problems with Stepwise Methods

- Rely on a mathematical criterion.
  - Variable selection may depend upon only slight differences in the Semi-partial correlation.
  - These slight numerical differences can lead to major theoretical differences.
- Should be used only for exploration

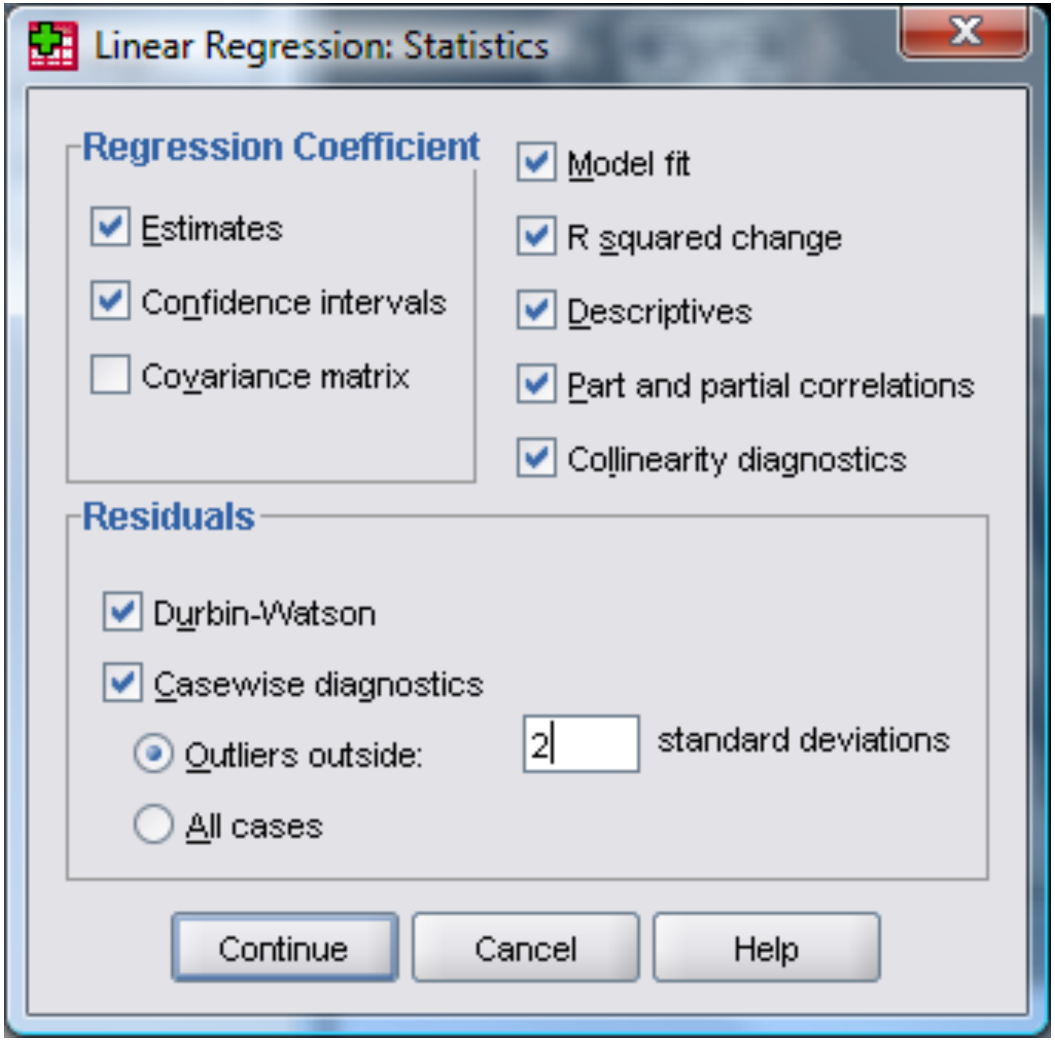
Doing Multiple Regression



Doing Multiple Regression

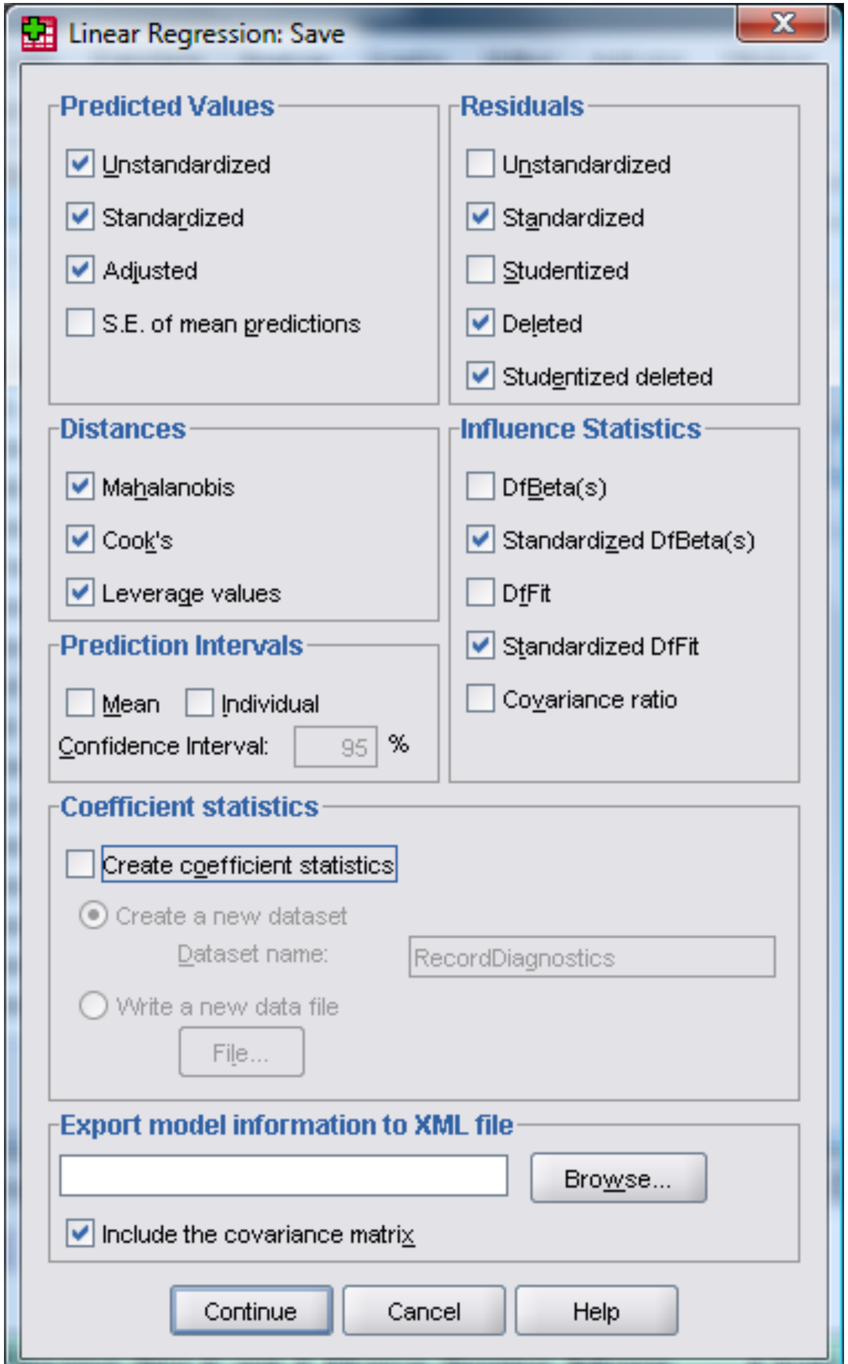


Regression Statistics



DISCOVERING STATISTICS USING SPSS
THIRD EDITION

Regression Diagnostics



ANDY FIELD

Output: Model Summary

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	65.9914	.335	99.587	1	198	.000	
2	.815[b]	.665	.660	47.0873	.330	96.447	2	196	.000	1.950

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

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

c. Dependent Variable: Record Sales (thousands)

R and R^2

- R
  - The correlation between the observed values of the outcome, and the values predicted by the model.
- R^2
  - Yhe proportion of variance accounted for by the model.
- Adj. R^2
  - An estimate of R^2 in the population (*shrinkage*).

Output: ANOVA

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	1295952.0	199			
2	Regression	861377.418	3	287125.806	129.498	.000[b]
	Residual	434574.582	196	2217.217		
	Total	1295952.0	199			

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

b. Predictors: (Constant), Advertising Budget (thousands of pounds), Attractiveness of Band, No. of Plays on Radio 1 per Week

c. Dependent Variable: Record Sales (thousands)

Analysis of Variance: ANOVA

- The F-test
  - looks at whether the variance explained by the model (SS_M) is significantly greater than the error within the model (SS_R).
  - It tells us whether using the regression model is significantly better at predicting values of the outcome than using the mean.

Output: betas

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	134.140	7.537		17.799	.000	119.278	149.002
	Advertising 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
	Advertising 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)

How to Interpret Beta Values

- Beta values:
  - the change in the outcome associated with a unit change in the predictor.
- Standardised beta values:
  - tell us the same but expressed as standard deviations.

Beta Values

- $b_1 = 0.087$.
  - So, as advertising increases by £1, record sales increase by 0.087 units.
- $b_2 = 3589$.
  - So, each time (per week) a song is played on radio 1 its sales increase by 3589 units.

Constructing a Model

$$y = b_0 + b_1 X_1 + b_2 X_2$$

$$\text{Sales} = 41124 + 0.087\,\text{Adverts} + 3589\,\text{plays}$$

$$\text{£1 Million Advertising, 15 plays}$$

$$\begin{aligned}
\text{Sales} &= 41124 + (0.087 \times 1{,}000{,}000) + (3589 \times 15) \\
&= 41124 + 87000 + 53835 \\
&= 181959
\end{aligned}$$

Standardised Beta Values

- $\beta_1 = 0.523$
  - As advertising increases by 1 standard deviation, record sales increase by 0.523 of a standard deviation.
- $\beta_2 = 0.546$
  - When the number of plays on radio increases by 1 s.d. its sales increase by 0.546 standard deviations.

Interpreting Standardised Betas

- As advertising increases by £485,655, record sales increase by $0.523 \times 80{,}699 = 42{,}206$.

- If the number of plays on radio 1 per week increases by 12, record sales increase by $0.546 \times 80{,}699 = 44{,}062$.

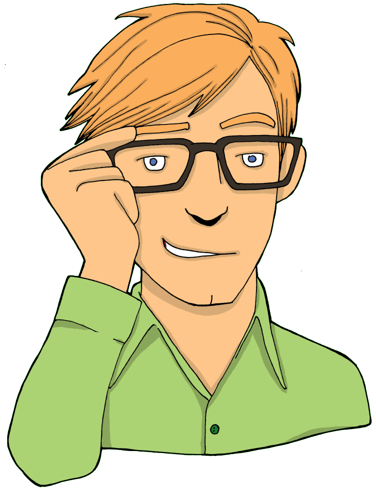
Reporting the Model

TABLE 7.2 How to report multiple regression

	B	*SE B*	β
Step 1			
Constant	134.14	7.54	
Advertising Budget	0.10	0.01	.58*
Step 2			
Constant	–26.61	17.35	
Advertising Budget	0.09	0.01	.51*
Plays on BBC Radio 1	3.37	0.28	.51*
Attractiveness	11.09	2.44	.19*

How well does the Model fit the data?

- There are two ways to assess the accuracy of the model in the sample:
- Residual Statistics
  - Standardized Residuals
- Influential cases
  - Cook's distance



Standardized Residuals

- In an average sample, 95% of standardized residuals should lie between ± 2.
- 99% of standardized residuals should lie between ± 2.5.
- Outliers
  - Any case for which the absolute value of the standardized residual is 3 or more, is likely to be an outlier.

Cook’s Distance

- Measures the influence of a single case on the model as a whole.
- Weisberg (1982):
  - Absolute values greater than 1 may be cause for concern.

Generalization

- When we run regression, we hope to be able to generalize the sample model to the entire population.
- To do this, several assumptions must be met.
- Violating these assumptions stops us generalizing conclusions to our target population.

Straightforward Assumptions

- Variable Type:
  - Outcome must be continuous
  - Predictors can be continuous or dichotomous.
- Non-Zero Variance:
  - Predictors must not have zero variance.
- Linearity:
  - The relationship we model is, in reality, linear.
- Independence:
  - All values of the outcome should come from a different person.

The More Tricky Assumptions

- No Multicollinearity:
  - Predictors must not be highly correlated.
- Homoscedasticity:
  - For each value of the predictors the variance of the error term should be constant.
- Independent Errors:
  - For any pair of observations, the error terms should be uncorrelated.
- Normally-distributed Errors

Multicollinearity

- Multicollinearity exists if predictors are highly correlated.
- This assumption can be checked with collinearity diagnostics.

Coefficients[a]

Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	Advertsing Budget (thousands of pounds)	.578	.578	.578	1.000	1.000
2	Advertsing Budget (thousands of pounds)	.578	.659	.507	.986	1.015
	No. of plays on Radio 1 per week	.599	.655	.501	.959	1.043
	Attractiveness of Band	.326	.309	.188	.963	1.038

a. Dependent Variable: Record Sales (thousands)

- Tolerance should be more than 0.2 (Menard, 1995)

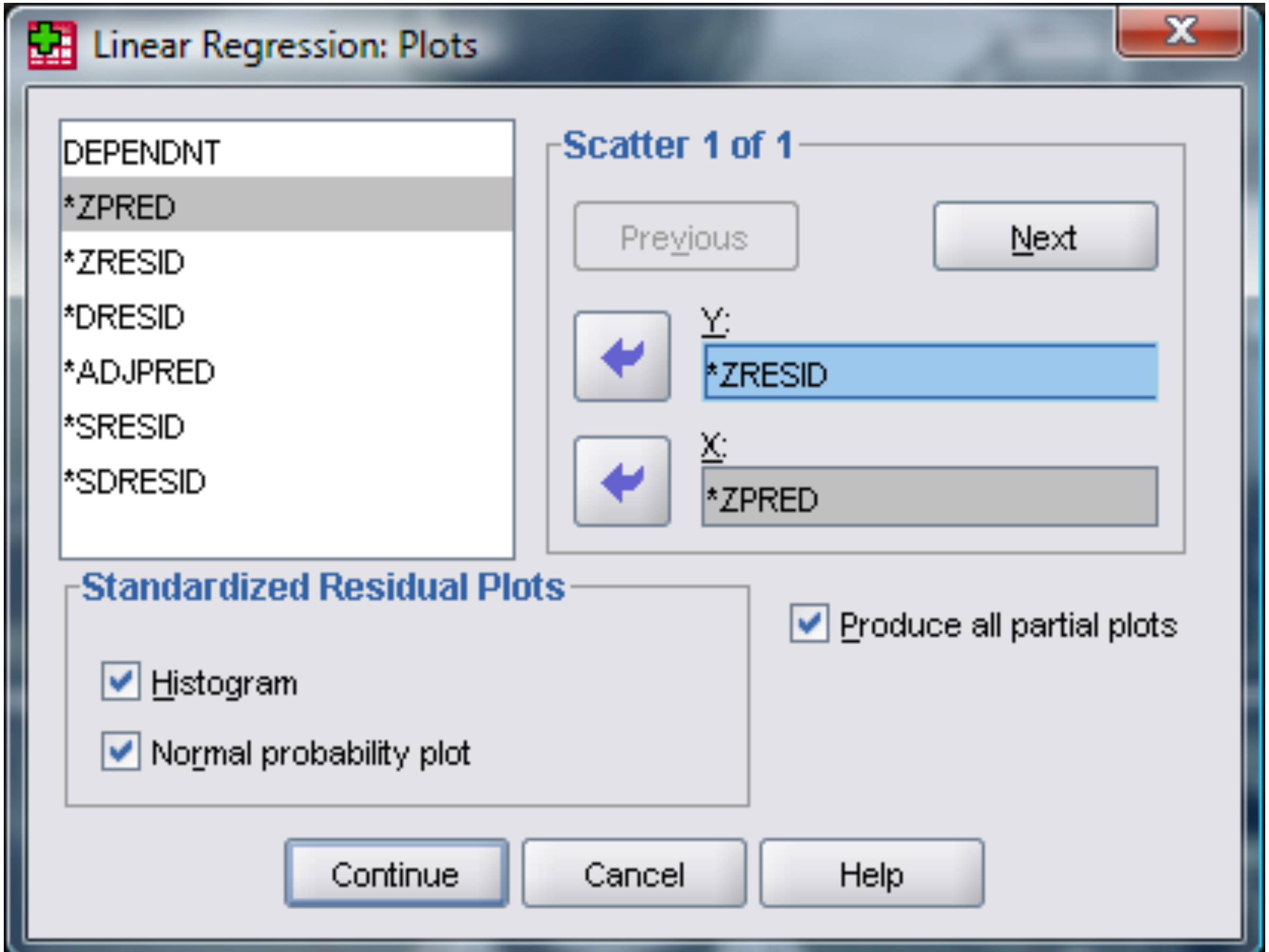
- VIF should be less than 10 (Myers, 1990)

Checking Assumptions about Errors

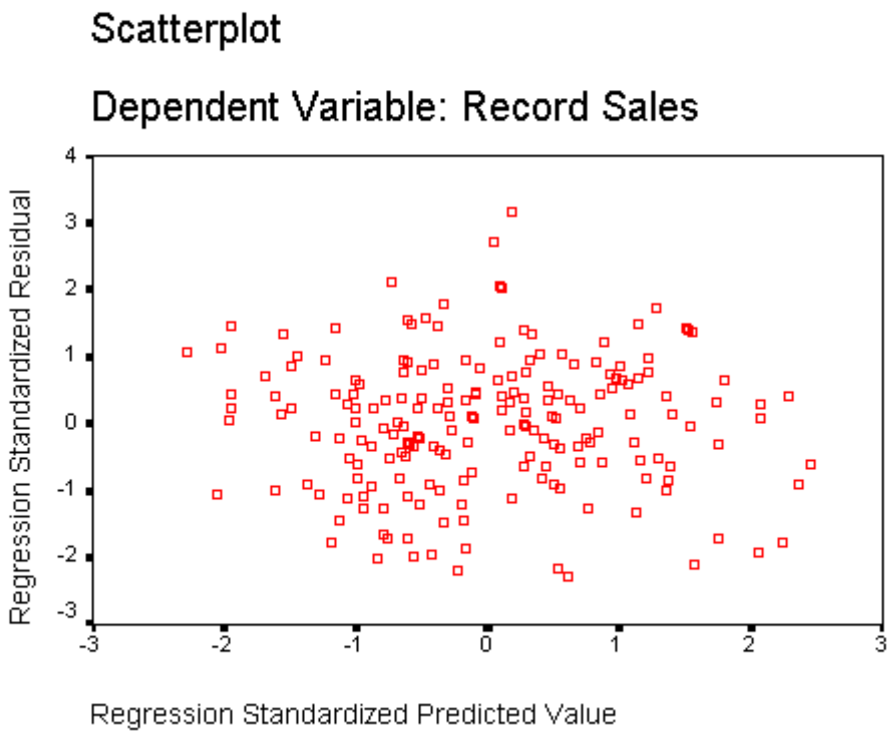
- Homoscedacity/Independence of Errors:
  - Plot ZRESID against ZPRED.

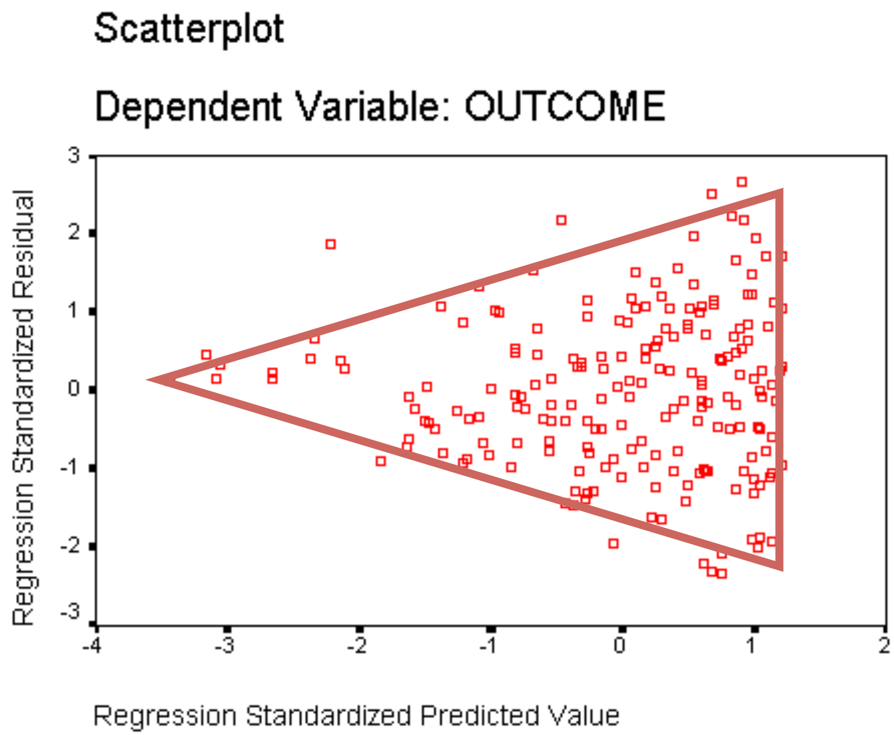
- Normality of Errors:
  - Normal probability plot.

Regression Plots



Homoscedasticity: ZRESID vs. ZPRED

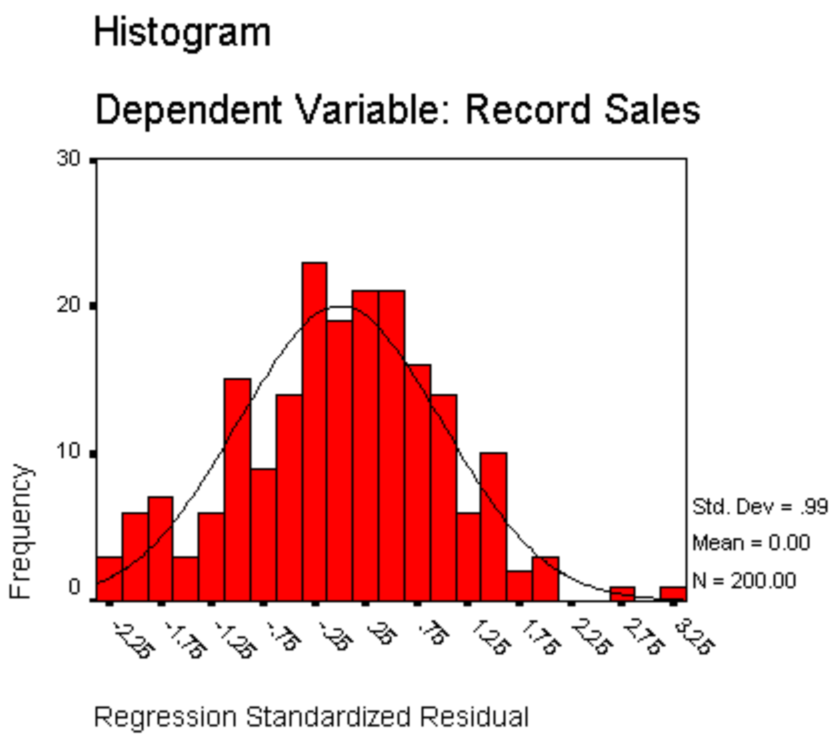


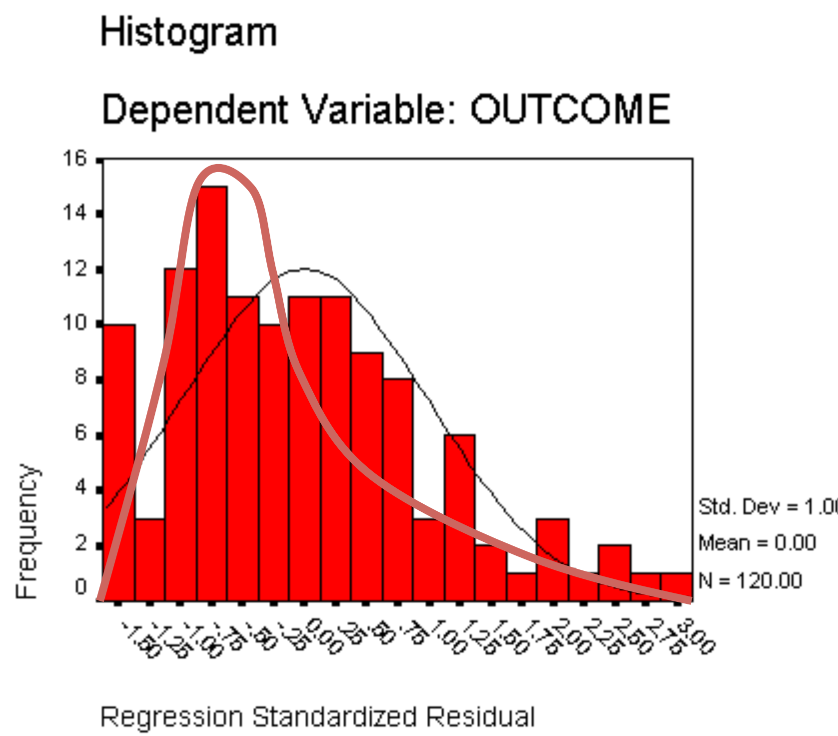


Good

Bad

Normality of Errors: Histograms

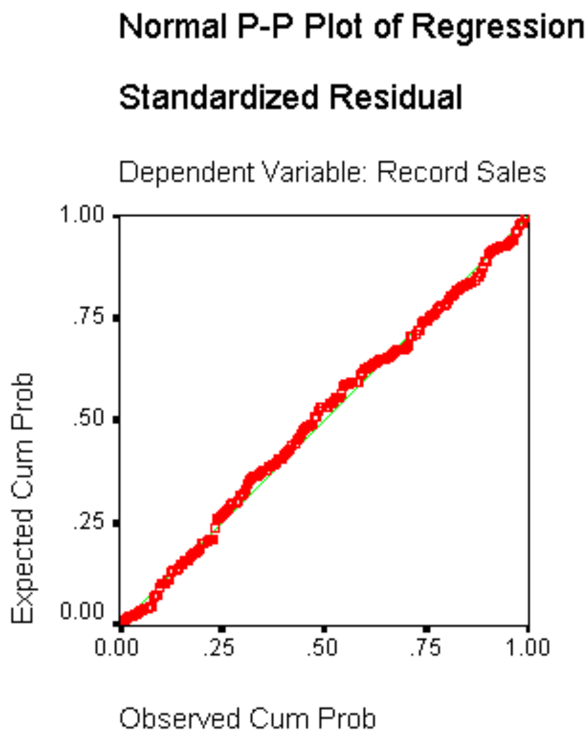




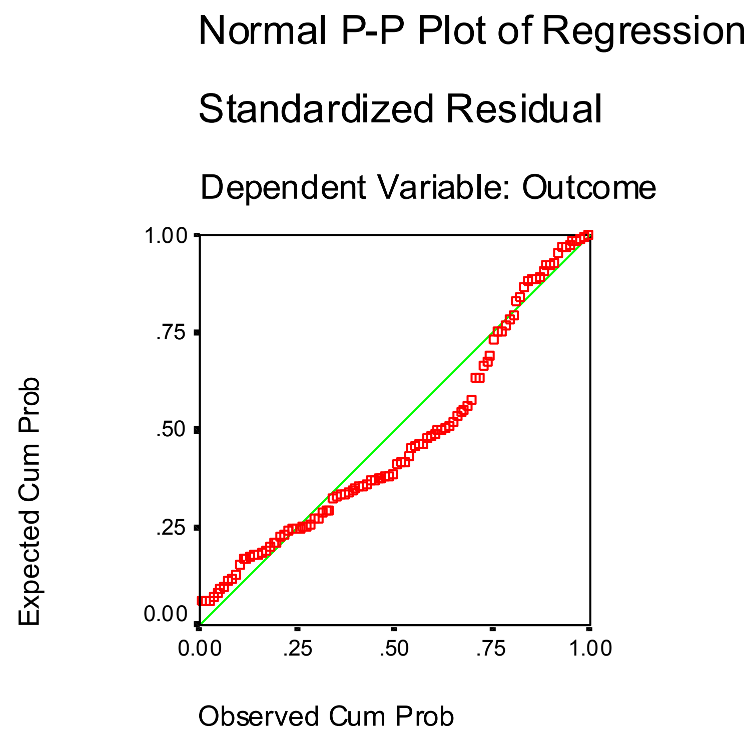
Good

Bad

Normality of Errors: Normal Probability Plot



Good



Bad