statstutor

community project

encouraging academics to share statistics support resources

Statistical Methods 12. Multiple Linear Regression and Analysis of Covariance

Based on materials provided by Coventry University and Loughborough University under a National HE STEM Programme Practice Transfer Adopters grant

www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Workshop outline

- Multiple Linear Regression:
  - Two independent variables
  - Multicollinearity: VIF and tolerance
  - More than two independent variables:
    - Direct variable entry method
    - Backwards regression method
  - Robustness
- Analysis of Covariance

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Please note

- This workshop assumes knowledge of simple linear regression – see Workshop 11
- Some disciplines have a different culture in applying multiple linear regression without assumption checking – please seek guidance from your faculty
- Most people want to look for significance for deciding which variables to include in the model, not for the purpose of prediction

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Multiple Linear Regression with two independent variables

Model:

$$y = b_0 + b_1x + b_2z$$

Where:

- y is the dependent variable
- x and z are the independent variables
- b_0 is the intercept coefficient
- b_1 and b_2 are the slope coefficients

Goal: To minimise the sum of the squares of the errors

www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

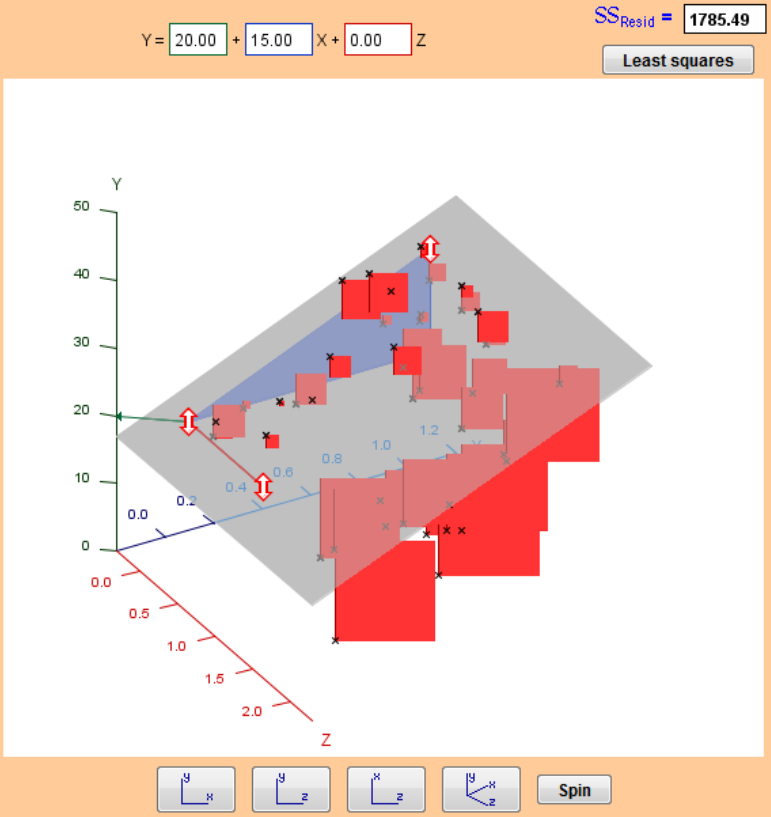
http://cast.massey.ac.nz/core/index.html?book=general, Section 6.3.7

Least squares estimation of y against x and z

The diagram below shows an artificial data set. A square is drawn beside each residual — its area is the squared residual. The total red area is therefore the sum of the squared residuals.

$$y_i = b_0 + b_1x_i + b_2z_i + e_i$$

Choose b_0, b_1 and b_2 such that $\sum_{i=1}^{n} e_i^2$ is minimised



The least squares plane can be moved by dragging the three arrows. Your aim should be to minimise the red area or, numerically, to minimise the residual sum of squares that is displayed at the top right of the diagram.

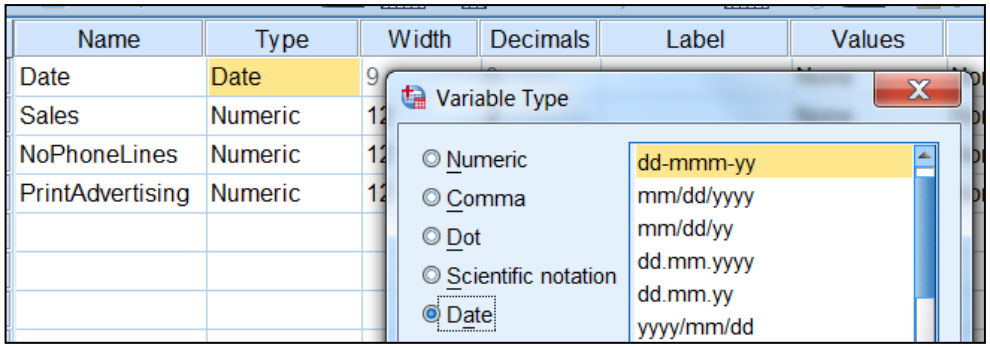
www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Example 1: Monthly sales figures for women's clothing

- 120 monthly sales figures for a catalogue-based mail ordering company from January 1989 to December 1998
- Independent variables:
  - Number of phone lines open for ordering
  - Amount spent on print advertising
- Open Sheet1 of the Excel file CatalogueData.xlsx associated with this workshop
- Turn it into an SPSS data file – for the *Date* field, use the data type "Date" and the format "dd-mmm-yy"



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

The regression analysis process

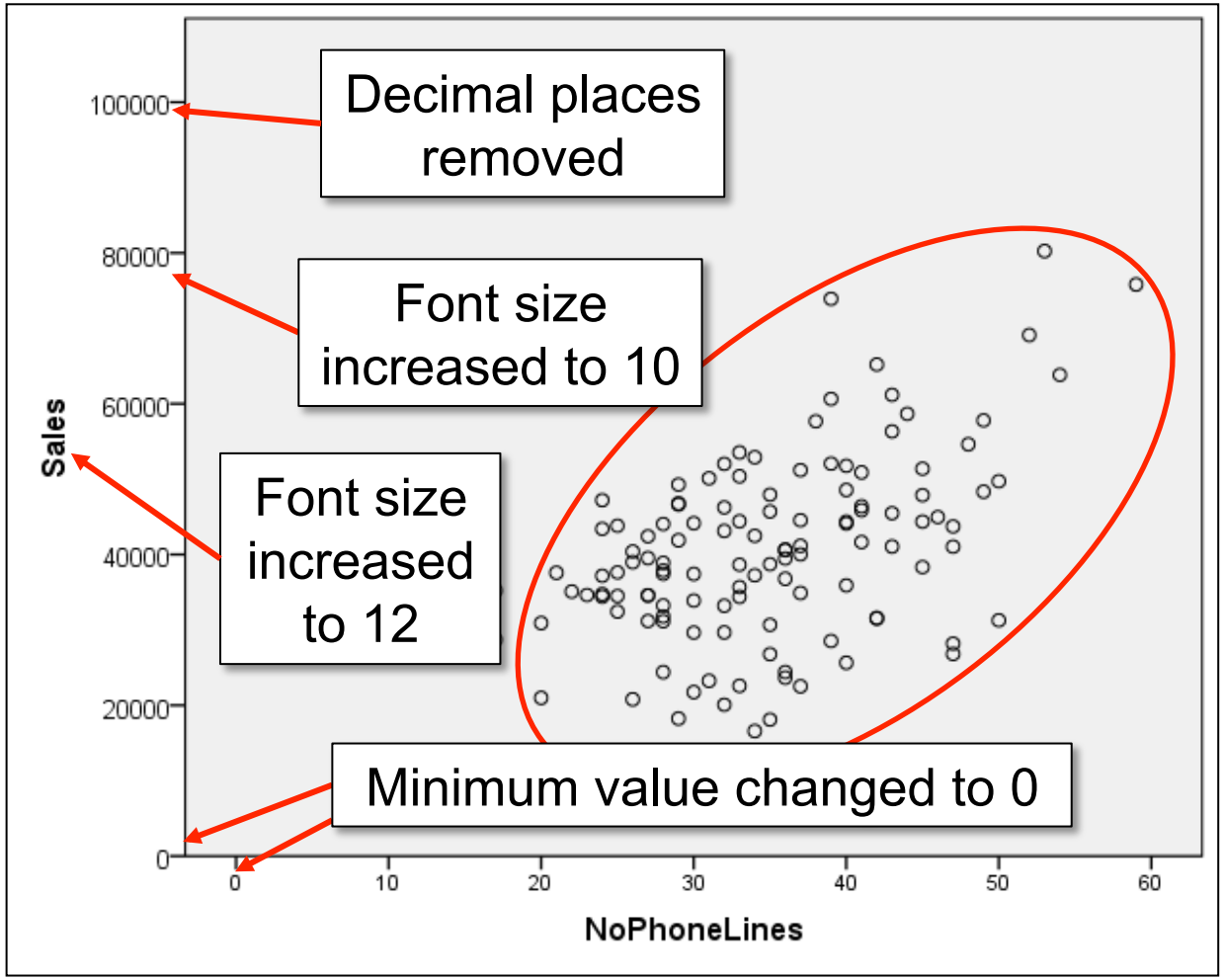
- **Step 1**: Get to know your data
- **Step 2**: Formulate a model and check the assumptions
- **Step 3**: Fit the model to the data
- **Step 4**: Report, interpret, and use the model

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 1A: Scatter plot of *Sales* against *NoPhoneLines*

- Relationship appears to be linear
- Variance in *Sales* appears to be constant for different values of *NoPhoneLines*
- 'Cigar shaped' data set

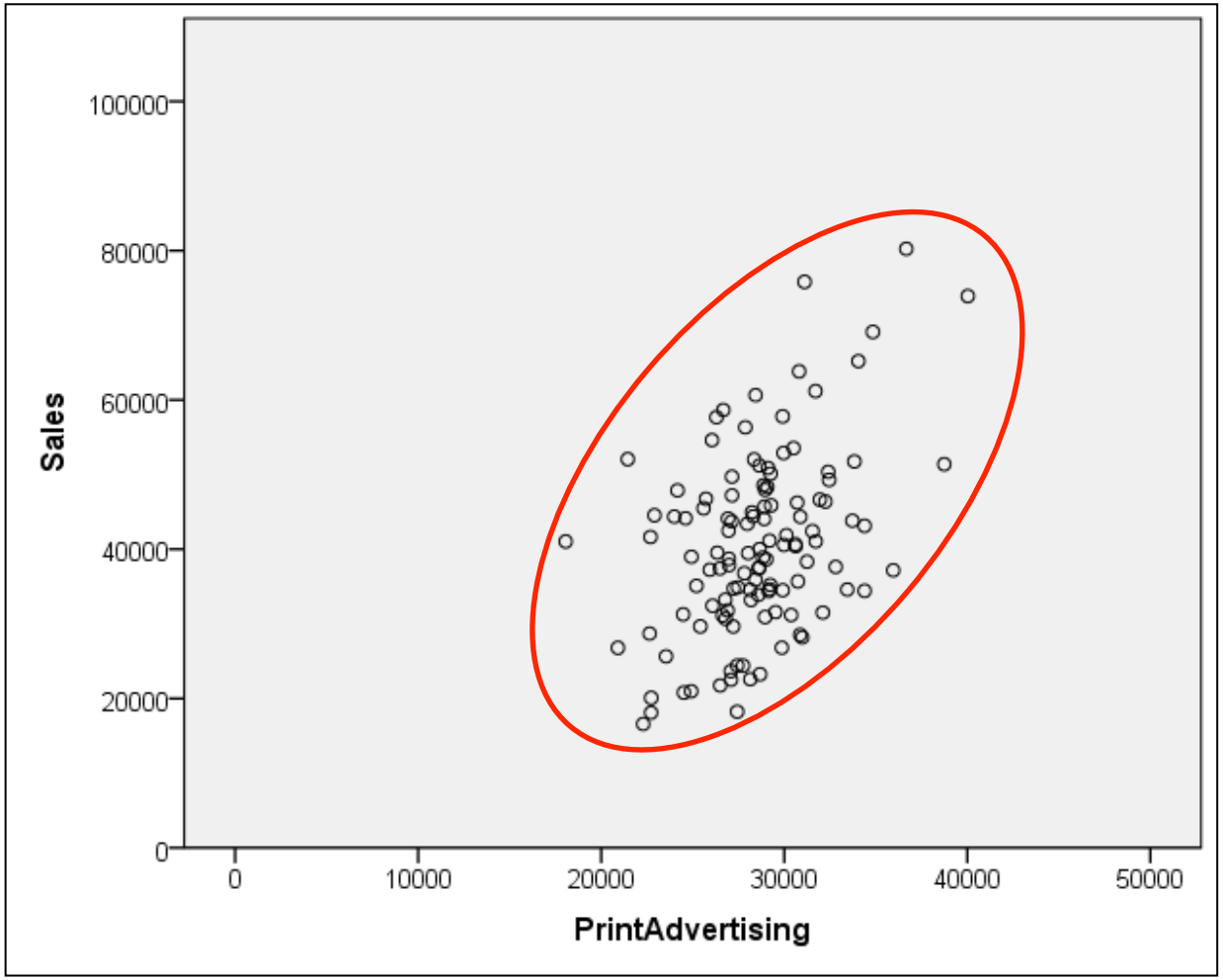


Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 1B: Scatter plot of *Sales* against *PrintAdvertising*

- Relationship appears to be linear
- Variance in *Sales* appears to be constant for different values of *PrintAdvertising*
- 'Cigar shaped' data set



www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 2: Formulate a model

$$Sales = b_0 + b_1 \times NoPhoneLines + b_2 \times PrintAdvertising$$

Note: A model is always an approximation to the data

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 2: Assumptions of Multiple Linear Regression

- The observations of the dependent variable are independent, e.g. they are not time or sequence dependent
- The independent variables are normally distributed or binary
- The dependent variable is normally distributed for each value of each predictor (independent) variable
- The variability of the outcome variable is the same for each value of the predictor variable
- The dependent variable varies linearly as the independent variables vary
- All seem to be OK for our data set (as indicated by the 'cigar shaped' scatter plots)

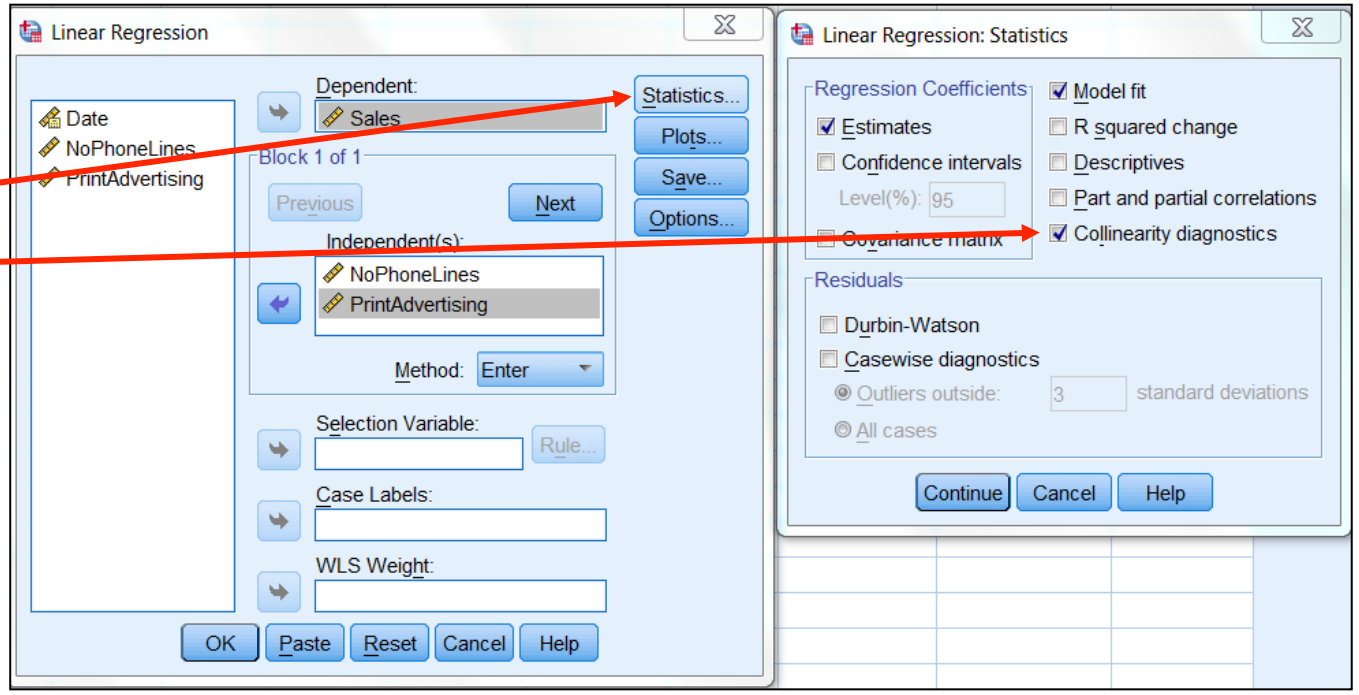
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 3: Fit the model to the data

- Analyze > Regression > Linear
- Add *Sales* as the dependent variable
- Add *NoPhoneLines* and *PrintAdvertising* as the independent variables

Select Statistics… and choose Collinearity diagnostics



www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Adjusted R Square = 0.383 ⇒ model explains 38.3% of the variation in *Sales*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.628[a]	.394	.383	9576.92890

a. Predictors: (Constant), PrintAdvertising, NoPhoneLines

b_0 = -21366

b_1 = 653.55

b_2 = 1.374

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-21365.807	7800.071		-2.739	.007		
	NoPhoneLines	653.552	104.165	.454	6.274	.000	.992	1.009
	PrintAdvertising	1.374	.252	.394	5.449	.000	.992	1.009

a. Dependent Variable: Sales

b_0, b_1 and b_2 all significantly different from 0

Tolerance coefficients slightly less than 1, VIF slightly more than 1 (see later)

www.**statstutor**.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 4: Fitted model

$$Sales = -21366 + 653.55 \times NoPhoneLines + 1.374 \times PrintAdvertising$$

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Multicollinearity

- Multicollinearity means there are high correlations between the predictor (independent) variables
- Thus two or more predictor variables carry similar information about the outcome variable
- With multicollinearity there is very high uncertainty for the regression coefficients
- Multicollinearity can be assessed informally by looking at the bivariate correlations between the independent variables (any $r > 0.8$ indicates a possible problem)
- There are two formal measures of multicollinearity:
  - Variance Inflation Factor (VIF)
  - Tolerance

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Variance Inflation Factor

- Based on fitting predictor variables to other predictor variables (i.e. *NoPhoneLines* to *PrintAdvertising* for our example) and calculating R^2
- Values of VIF (Variance Inflation Factor):
  - VIF < 5: don't worry
  - 5 < VIF < 10: multicollinearity may be a problem, be cautious
  - VIF > 10: multicollinearity is definitely a problem and will adversely affect results

Source: Myers, R. (1990) *Classical and modern regression with applications*. 2nd ed. Boston, MA: Duxbury

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Tolerance

- The percentage of the variance in a predictor variable that cannot be explained by the other predictors
- In our example the tolerance for each variable was 0.992, meaning 99.2% of the variance in both predictor variables cannot be explained by the other variables
- Tolerance is the reciprocal of VIF – so you only need to look at VIF
- For our example, VIF was 1.009 for both variables so it was not a problem

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Multiple Linear Regression with >2 independent variables

Model:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m$$

Where:

- y is the dependent variable
- $x_1, x_2, \ldots, x_m$ are the independent variables
- b_0 is the intercept coefficient
- $b_1, b_2, \ldots, b_m$ are the slope coefficients

Goal: To minimise the sum of the squares of the errors

Rule of thumb: For a sample size of n, use no more than $\sqrt{n}$ independent variables

www.**stats**tutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Issues with >2 variables

- Much more likely to have multicollinearity, or variables with non-significant coefficients
- Impossible to know in advance which variables are best removed

⇒ Use a systematic variable selection method in SPSS to determine an adequate model

⇒ Justify any decision you make

- We recommend **backwards** removal of predictor variables:
  - All predictor variables initially included
  - Least significant variable removed
  - Repeat process until least significant variable is below a threshold (default is 0.1)

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Example 2: Monthly sales figures for women's clothing

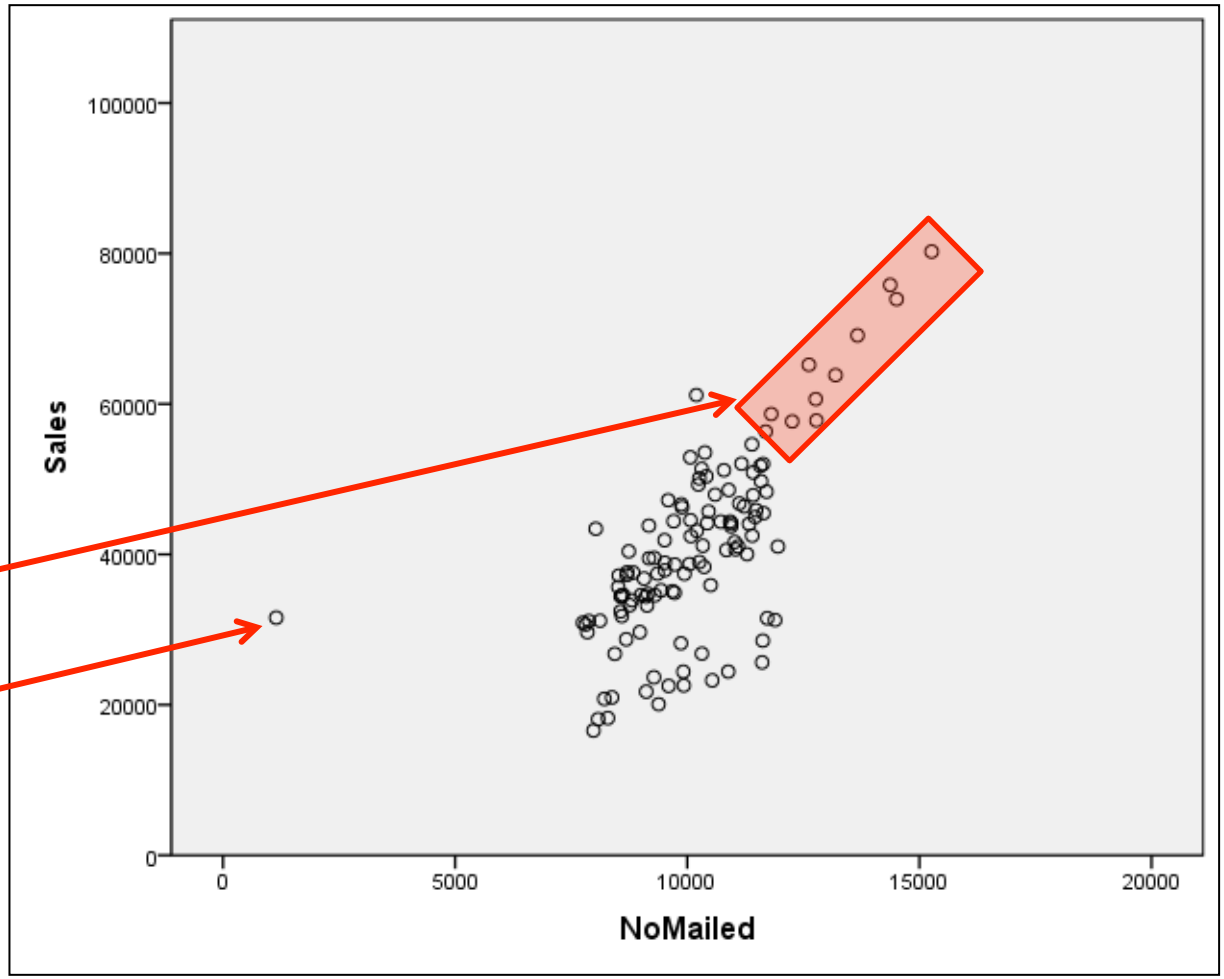
- Independent variables:
  - Number of phone lines open for ordering
  - Amount spent on print advertising
  - Number of catalogues mailed
  - Number of pages in catalogue
  - Number of customer service representatives
- Open Sheet2 of the Excel file Catalogue2Data.xlsx associated with this workshop
- Turn it into an SPSS data file

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 1A: *Sales* v. *NoMailed*

- Clear linear relationship
- May be heteroscedastic – variance in errors seems to depend on NoMailed
- This looks like an outlier – reason?



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Activity

- Create a new variable called *NoMailedGroup* by recoding *NoMailed* into a new variable
- Choose a suitable cut-off value
- *NoMailedGroup* = 1 below the cut-off value
- *NoMailedGroup* = 2 above the cut-off value
- Run a linear regression of Sales against *NoMailed* and choose Unstandardised residuals under Save…
- Run an independent samples t-test of the unstandardised residuals against *NoMailedGroup*

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

NoMailed cut-off = 12,500

Independent Samples Test

		Levene's Test for Equality of Variances			
		F	Sig.	t	df
Unstandardized Residual	Equal variances assumed	2.268	.135	-3.627	118
	Equal variances not assumed			-7.351	13.663

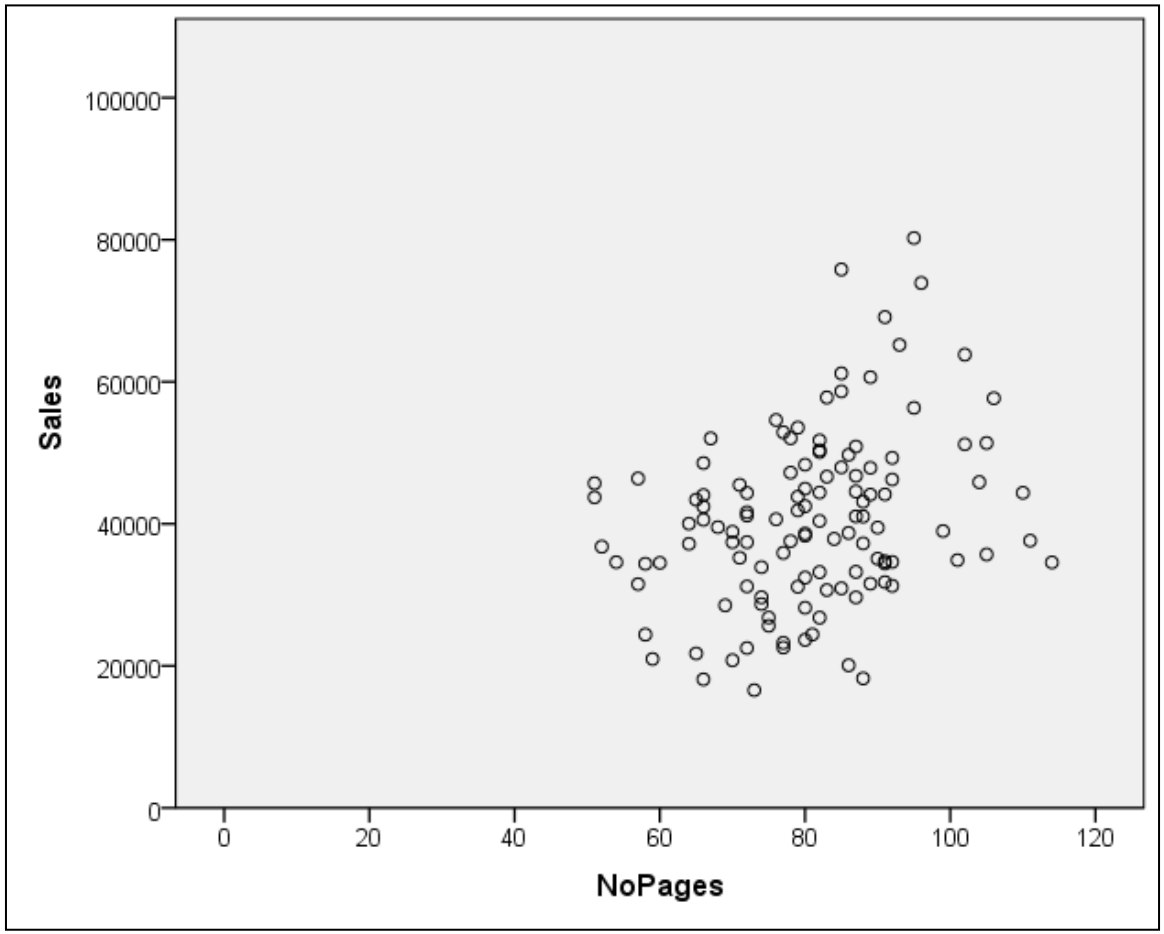
- Levene's test for equality of variances returns a non-significant result
- Only 8 data values in *NoMailedGroup* 2
- Probably OK to assume homoscedasticity
- There is no specific test for heteroscedasticity in SPSS

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 1B: *Sales* v. *NoPages*

- Seems to be a weak linear relationship
- Variance seems to be OK



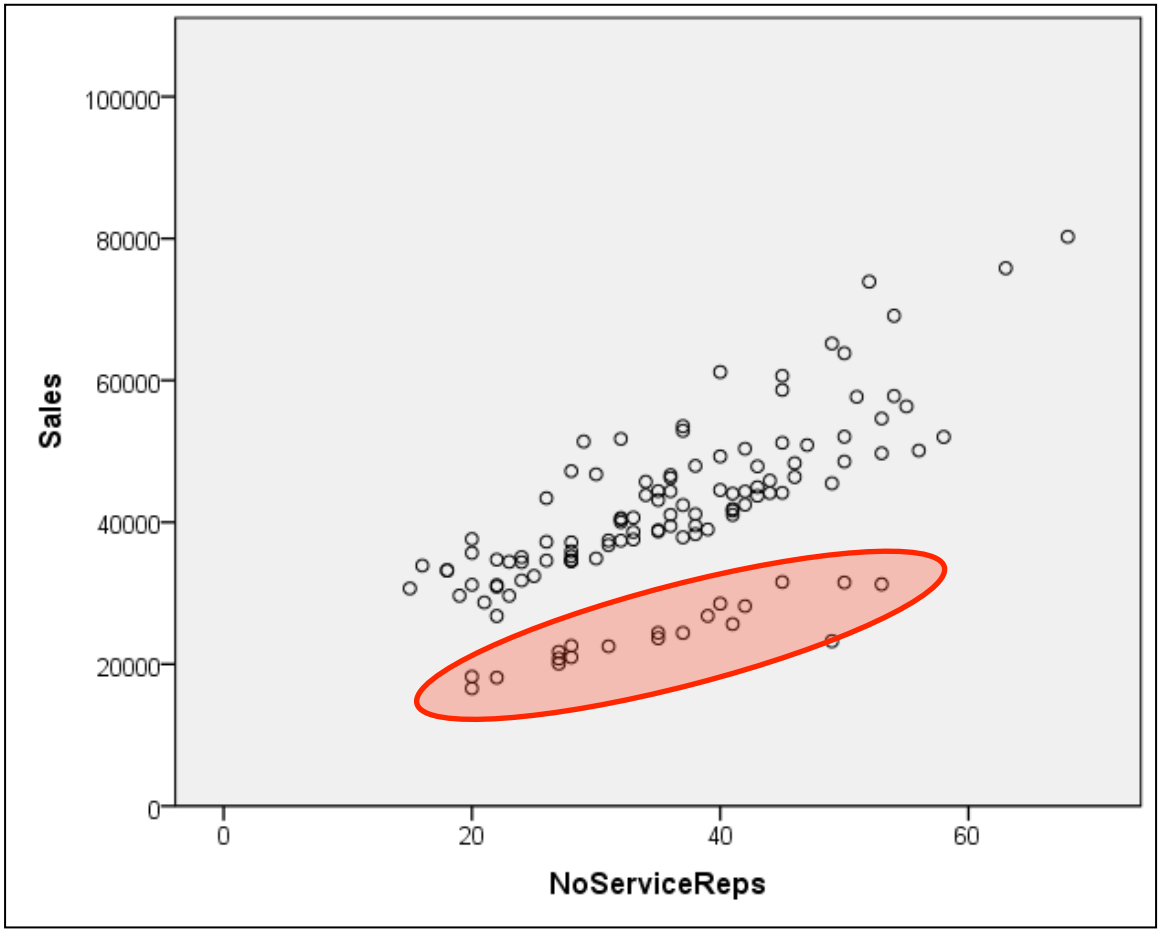
www.**stats**tutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 1C: *Sales* v. *NoServiceReps*

- Stronger linear relationship
- Data seems to be in two distinct groups – month dependent?



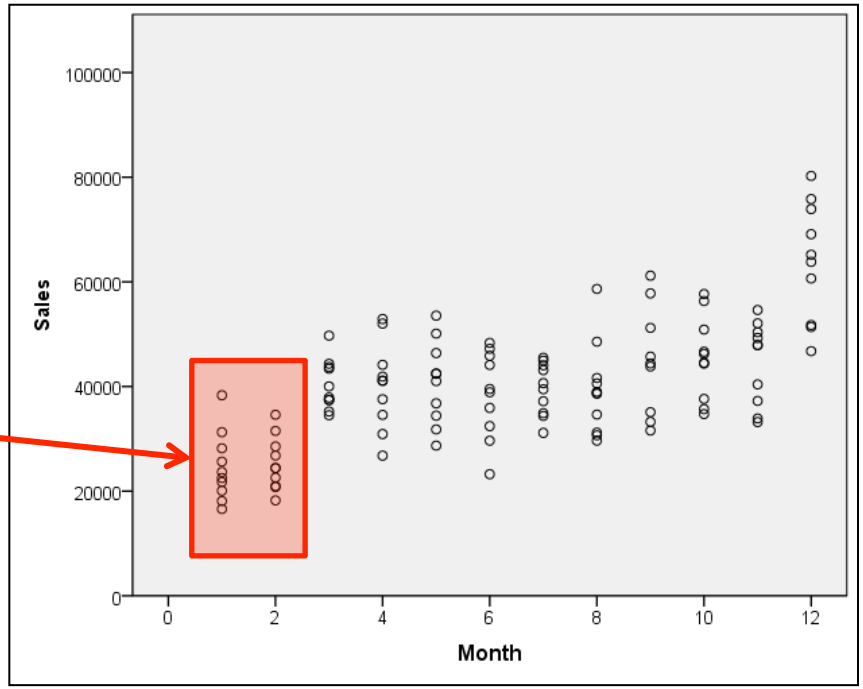
www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Activity

- Create a new scale variable called Month
- Enter 1 for a date in January, 2 for a date in February, etc.
- Create a scatter plot of *Sales* against *Month*
- Clearly lower in January and February

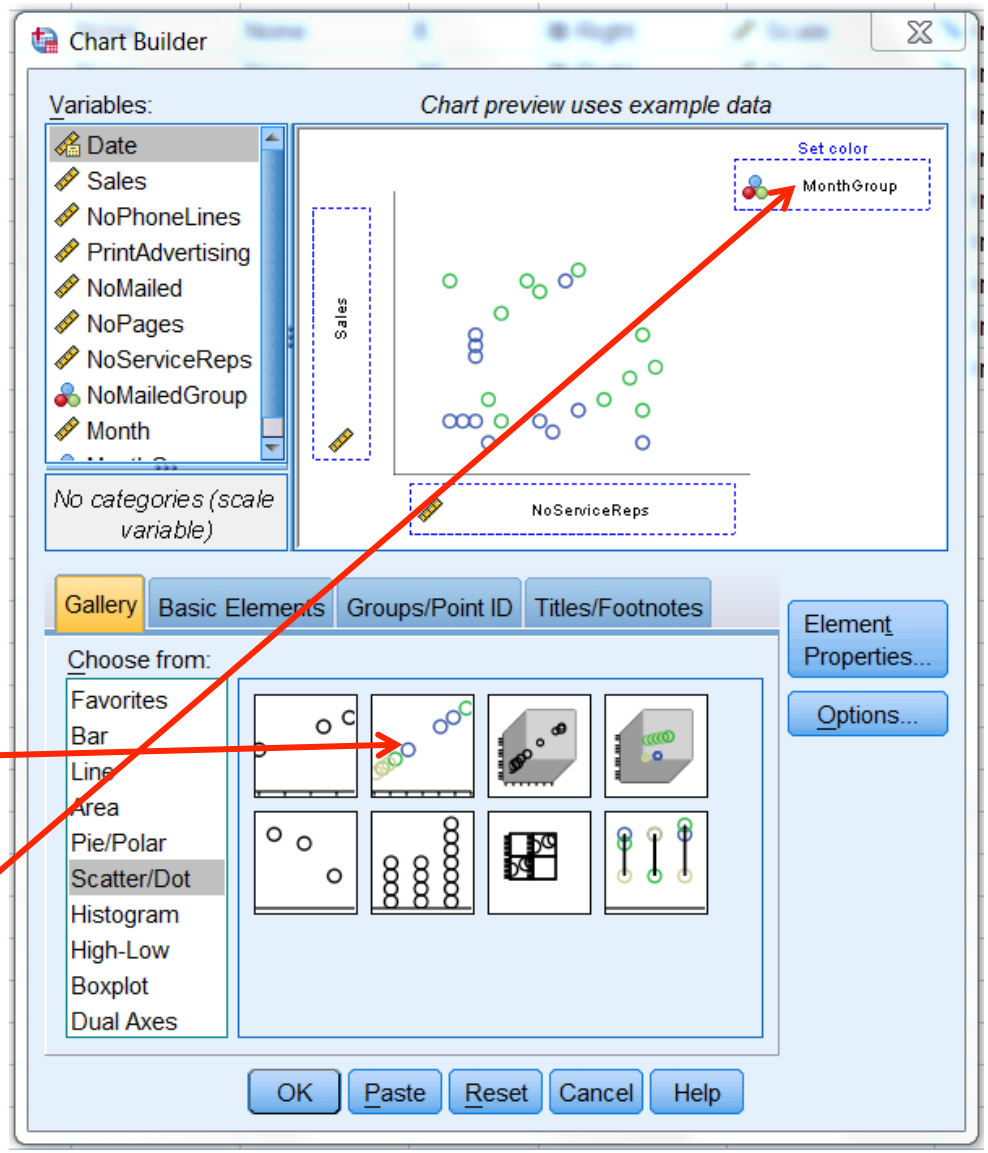


www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

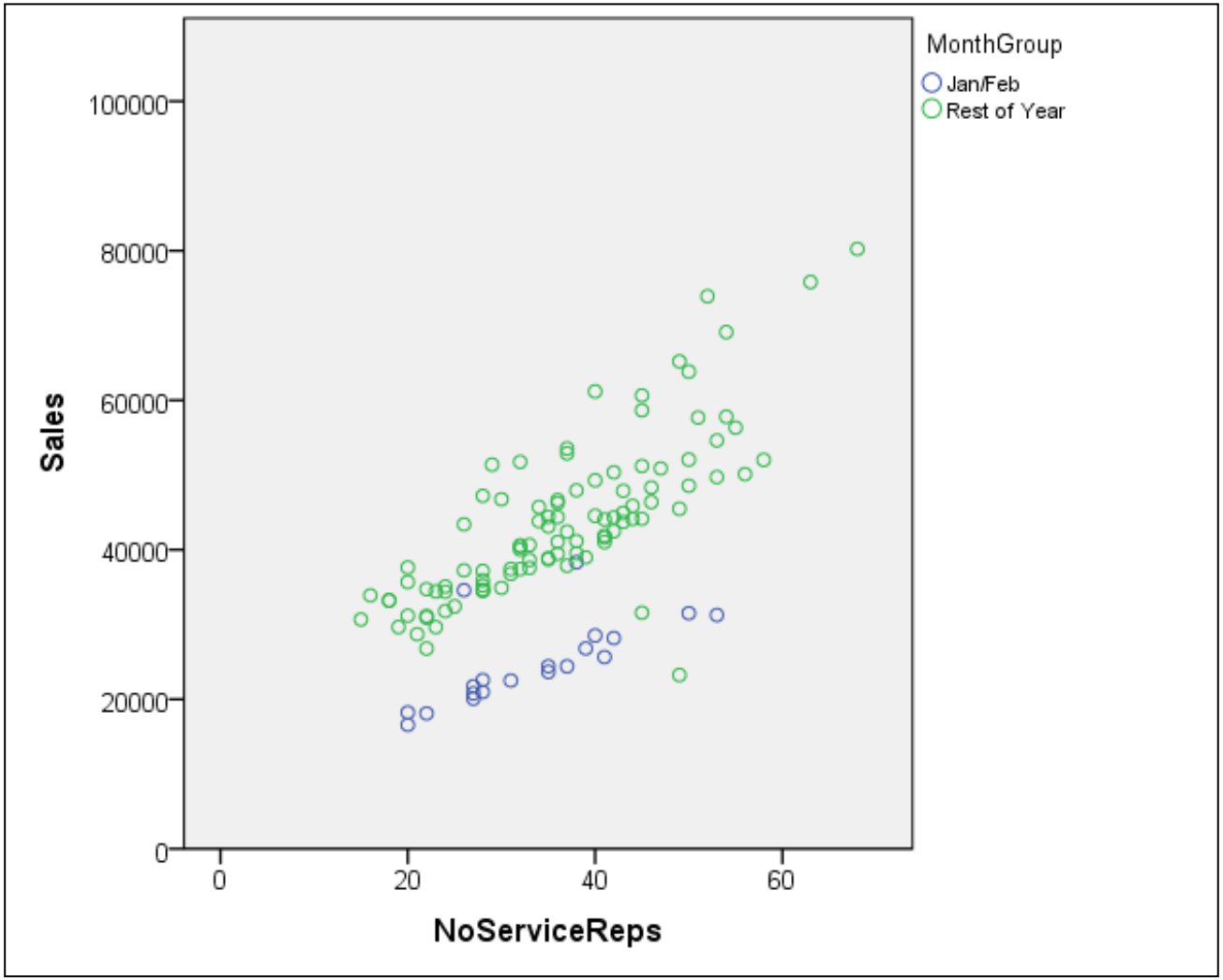
- Recode *Month* into a new variable called *MonthGroup* with 1 = Jan/Feb and 2 = otherwise
- Add values to *MonthGroup* to represent these groups
- Create a grouped scatter plot of *Sales* v. *NoServiceReps* with *MonthGroup* as the grouping variable



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

- Most of the lower *Sales* were from Jan/Feb
- We could add an additional variable to the model which is 1 for Jan/Feb and 0 for other months



- For the moment the months Jan/Feb are excluded
- Also see ANCOVA analysis later

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 2: Formulate a model

For the months March – December:

$$Sales = b_0 + b_1 \times NoPhoneLines + b_2 \times PrintAdvertising + b_3 \times NoMailed + b_4 \times NoPages + b_5 \times NoServiceReps$$

Peter Samuels
Birmingham City University

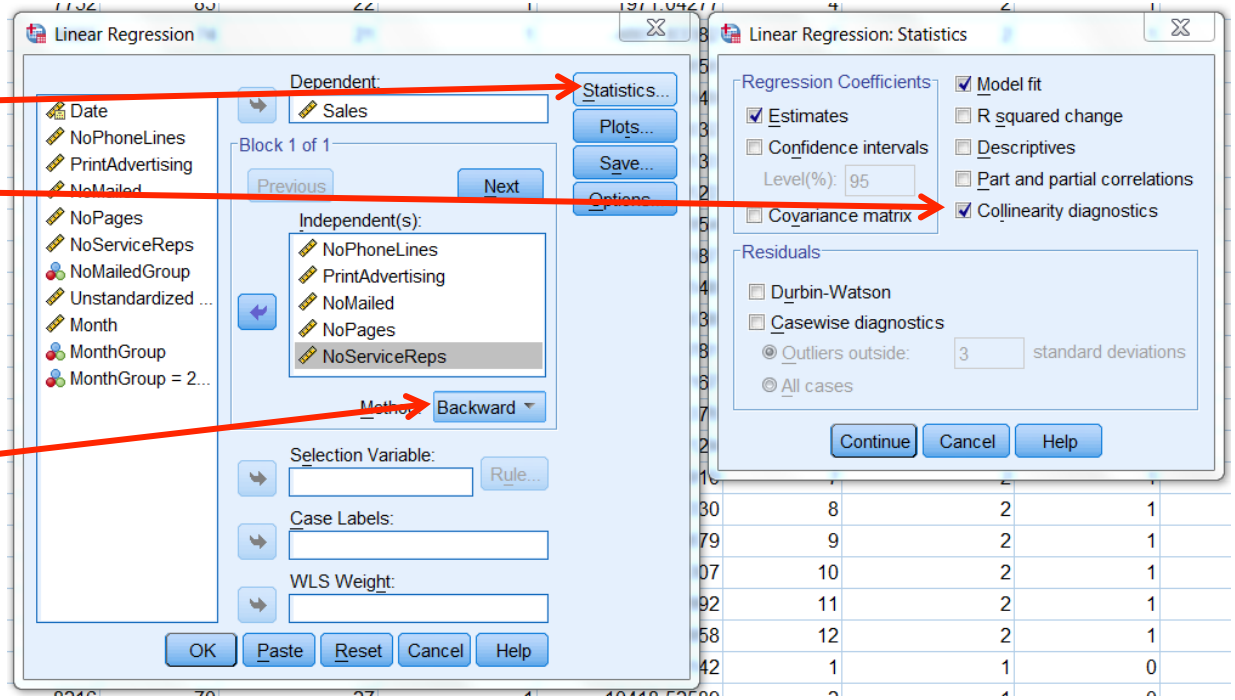
Reviewer: Ellen Marshall
University of Sheffield

Step 4: Fit the model

- Select only the cases with MonthGroup = 2
- Analyze > Regression > Linear
- Select *Sales* as the dependent variable and the other 5 variables and the independent variables

Under Statistics… select collinearity diagnostics

Select the method as Backward



www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

- Only one model required
- R^2_{adj} = 0.795 ⇒ model accounts for 79.5% of the variation in Sales

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.898[a]	.806	.795	4812.54378

a. Predictors: (Constant), NoServiceReps, NoPages, PrintAdvertising, NoPhoneLines, NoMailed
b. Dependent Variable: Sales

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-28309.791	5261.732		-5.380	.000		
	NoPhoneLines	144.430	79.671	.117	1.813	.073	.499	2.005
	PrintAdvertising	.797	.142	.261	5.632	.000	.960	1.041
	NoMailed	2.344	.395	.389	5.930	.000	.480	2.083
	NoPages	81.857	37.967	.100	2.156	.034	.968	1.033
	NoServiceReps	368.010	65.455	.388	5.622	.000	.435	2.301

a. Dependent Variable: Sales

- All variables included in initial model (backwards method)
- No variable removed because all the probability values < 0.1

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Activity

- Repeat the analysis with all the months included
- What affect does this have on the models?
- Analysis now has 2 models
- R^2_{adj} slightly higher in second model
- Both markedly lower than previous analysis

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.815[a]	.664	.650	7220.31815
2	.814[b]	.663	.652	7199.18267

a. Predictors: (Constant), NoServiceReps, NoPages, PrintAdvertising, NoPhoneLines, NoMailed
b. Predictors: (Constant), NoServiceReps, NoPages, PrintAdvertising, NoMailed
c. Dependent Variable: Sales

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-40897.136	7148.096		-5.721	.000		
	NoPhoneLines	-62.861	109.806	-.044	-.572	.568	.507	1.972
	PrintAdvertising	.980	.195	.281	5.035	.000	.946	1.057
	NoMailed	2.454	.577	.342	4.249	.000	.456	2.194
	NoPages	185.344	51.766	.198	3.580	.001	.963	1.039
	NoServiceReps	442.734	94.195	.397	4.700	.000	.412	2.425
2	(Constant)	-41328.937	7087.381		-5.831	.000		
	PrintAdvertising	.989	.193	.284	5.113	.000	.952	1.050
	NoMailed	2.364	.554	.329	4.264	.000	.491	2.035
	NoPages	181.638	51.210	.194	3.547	.001	.978	1.022
	NoServiceReps	420.165	85.297	.377	4.926	.000	.500	2.000

a. Dependent Variable: Sales

- *NoPhoneLines* was removed from Model 2 because its probability value in Model 1 > 0.1 (the absolute value of its standardised coefficient was low)
- Model 2 has very low probability values for all the variables

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Robustness exceptions

- Homoscedasticity is mandatory – otherwise use a nonparametric technique
- Linearity is mandatory – otherwise transform the independent variable or use another model, e.g. quadratic or polynomial
- Normality is "not necessary for the least-squares fitting of the regression model, but it is required in general for inference making." (e.g. calculating the p-values and confidence intervals of the coefficients) "…only **extreme departures** of the distribution of Y from normality yield spurious results."

  Source: (Kleinbaum et al., 2008: 120)

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Application of robustness exceptions to activity

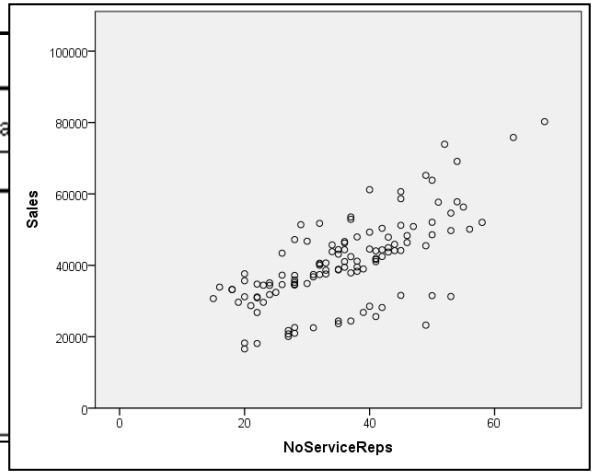
- *Sales* v. *NoServiceReps* was not normally distributed

⇒ The probability values of the coefficients may not be reliable

- However, the probability value was not borderline so the model can still be used
- Including January & February data just increases the 'noise'

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Sta	
		B	Std. Error	Beta			Tolerance	
1	(Constant)	-40897.136	7148.096		-5.72	.000		
	NoPhoneLines	-62.861	109.806	-.044	-.572	.568	.507	
	PrintAdvertising	.980	.195	.281	5.035	.000	.946	
	NoMailed	2.454	.577	.342	4.249	.000	.456	
	NoPages	185.344	51.766	.198	3.580	.001	.963	
	NoServiceReps	442.734	94.195	.397	4.700	.000	.412	



www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Example 3: Monthly catalogue sales figures for jewellery

- Independent variables:
  - Number of phone lines open for ordering
  - Amount spent on print advertising
  - Number of catalogues mailed
  - Number of pages in catalogue
  - Number of customer service representatives
- Open Sheet3 of the Excel file CatalogueData.xlsx associated with this workshop
- Turn it into an SPSS data file

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Activity

- Use *JewellerySales* as the dependent variable
- Fit the model with all 5 independent or predictor variables (direct or 'enter' method)
- Is multicollinearity important?
- Just re-fitting the model with only the variables with significant coefficients may not be sufficient
- Try also the backward variable selection method

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Five independent variable direct model

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-26389.042	5090.672		-5.184	.000		
	NoMailed	1.667	.411	.420	4.055	.000	.456	2.194
	NoPages	8.276	36.866	.016	.224	.823	.963	1.039
	NoPhoneLines	88.073	78.201	.111	1.126	.262	.507	1.972
	PrintAdvertising	.941	.139	.489	6.790	.000	.946	1.057
	NoServiceReps	-120.722	67.083	-.196	-1.800	.075	.412	2.425

a. Dependent Variable: JewellerySales

Three independent variable direct model

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-24901.816	4447.771		-5.599	.000		
	NoMailed	1.805	.393	.455	4.587	.000	.495	2.019
	PrintAdvertising	.931	.138	.483	6.763	.000	.954	1.048
	NoServiceReps	-88.945	60.766	-.145	-1.464	.146	.500	2.000

a. Dependent Variable: JewellerySales

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Results of backwards regression method

R^2_{adj} does not necessarily reduce when variables are removed from the model

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.664[a]	.441	.417	5142.10638
2	.664[b]	.441	.422	5120.83194
3	.659[c]	.434	.420	5128.87907
4	.651[d]	.424	.414	5153.85885

a. Predictors: (Constant), NoServiceReps, NoPages, PrintAdvertising, NoPhoneLines, NoMailed
b. Predictors: (Constant), NoServiceReps, PrintAdvertising, NoPhoneLines, NoMailed
c. Predictors: (Constant), NoServiceReps, PrintAdvertising, NoMailed
d. Predictors: (Constant), PrintAdvertising, NoMailed

Final model removes *NoServiceReps*

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

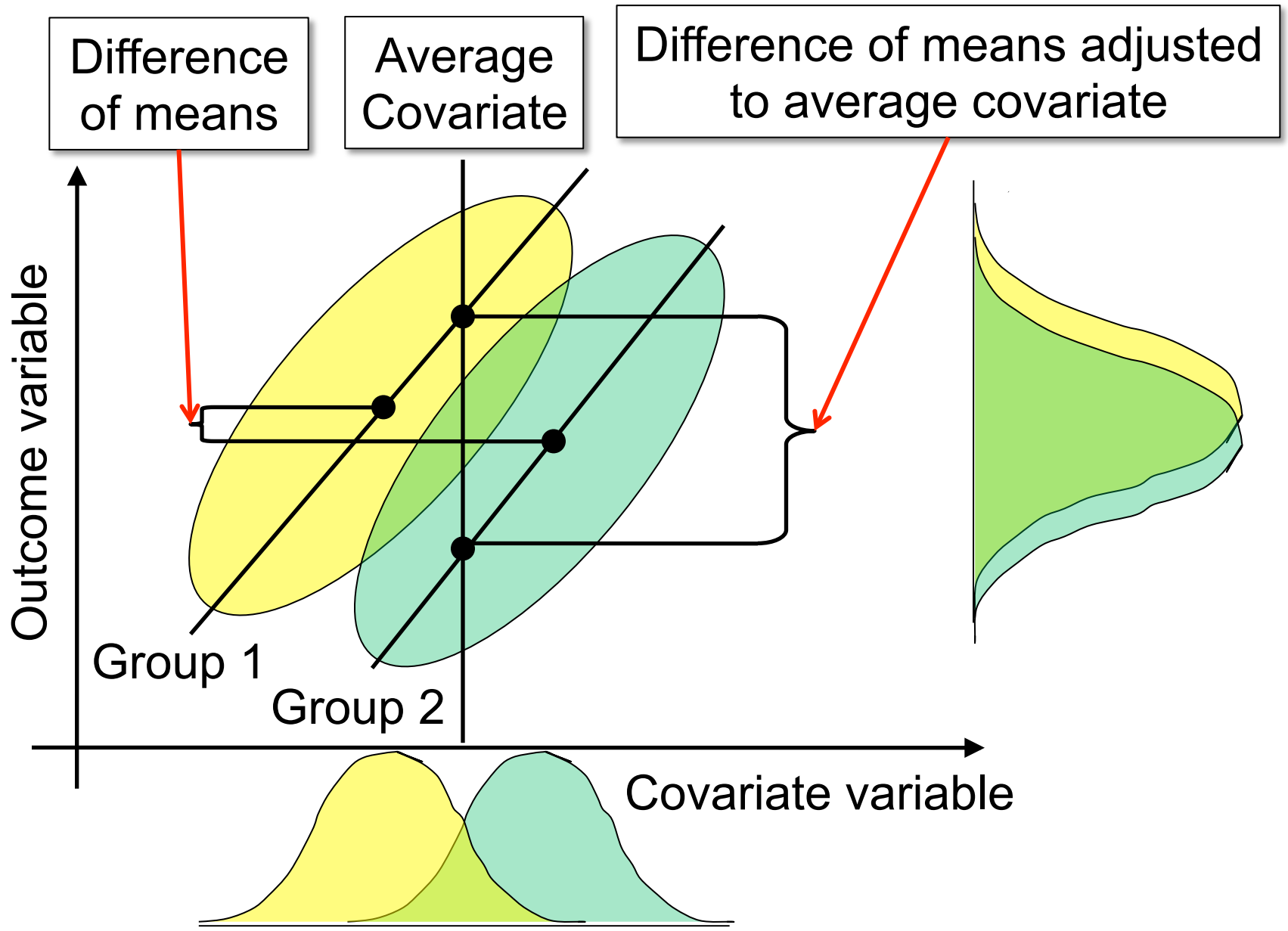
Analysis of Covariance (ANCOVA)

- Combines ANOVA and Linear Regression:
  - ANOVA: Explains outcome for different groups of the data
  - Linear Regression: Explains outcome with explanatory variables
  - ANCOVA: Does both simultaneously
- Increases the precision of the analysis
- Compares the means at the average values of the predictor variables
- The predictor (independent) variables must be correlated to the outcome (dependent) variable

www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield



www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Assumptions for ANCOVA

- Same as ANOVA and Linear Regression:
  - Independence of observations
  - Equality of variances
  - Normality of distribution
- In addition:
  - Equal regression slopes – if this assumption is in doubt, add an extra binary variable to the model to represent the two groups
- ANCOVA models can be very complex
- Seek advice if you feel ANCOVA is appropriate for your research

Peter Samuels
Birmingham City University

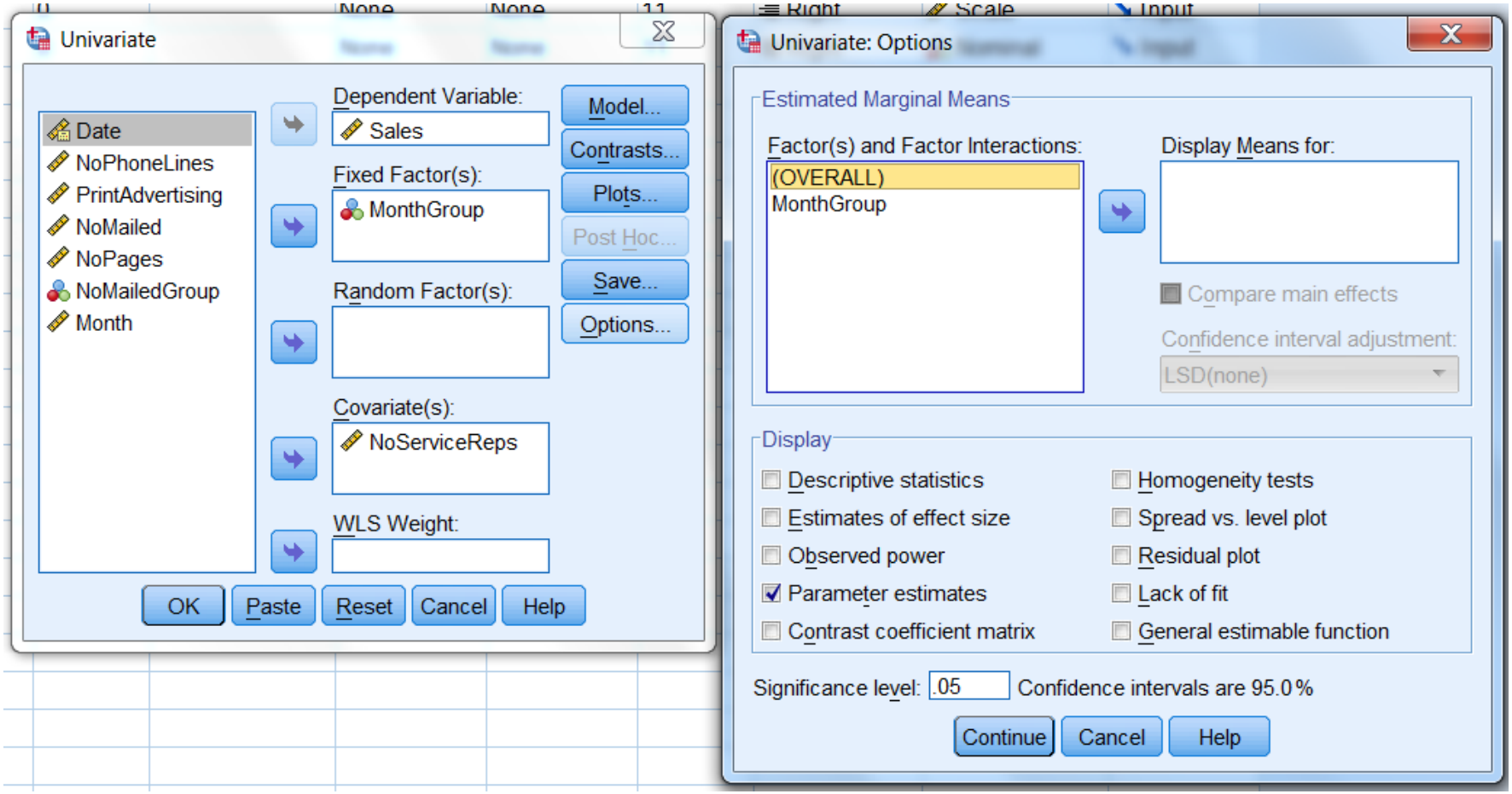
Reviewer: Ellen Marshall
University of Sheffield

Example of ANCOVA

- Open the SPSS file Catalogue2 you created earlier
- We want to compare *Sales* in January and February against the rest of the year
- The independent variable is *NoServiceReps*
- Use Analyze > General Linear Model > Univariate:
  - Dependent variable: *Sales*
  - Fixed Factor: *MonthGroup*
  - Covariate: *NoServiceReps*
  - Under Options… choose Parameter estimates

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield



www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Model accounts for 71.4% of the total variance

Tests of Between-Subjects Effects

Dependent Variable:Sales

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1.273E10	2	6.366E9	149.835	.000
Intercept	9.772E8	1	9.772E8	23.001	.000
NoServiceReps	6.864E9	1	6.864E9	161.557	.000
MonthGroup	4.499E9	1	4.499E9	105.893	.000
Error	4.971E9	117	42483467.74		
Total	2.153E11	120			
Corrected Total	1.770E10	119			

a. R Squared = .719 (Adjusted R Squared = .714)

The *MonthGroup* =1 coefficient should be added to the intercept coefficient for this group

Parameter Estimates

Dependent Variable:Sales

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	18224.491	2108.417	8.644	.000	14048.882	22400.099
NoServiceReps	698.256	54.935	12.711	.000	589.460	807.053
[MonthGroup=1]	-16528.609	1606.210	-10.290	.000	-19709.624	-13347.594
[MonthGroup=2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

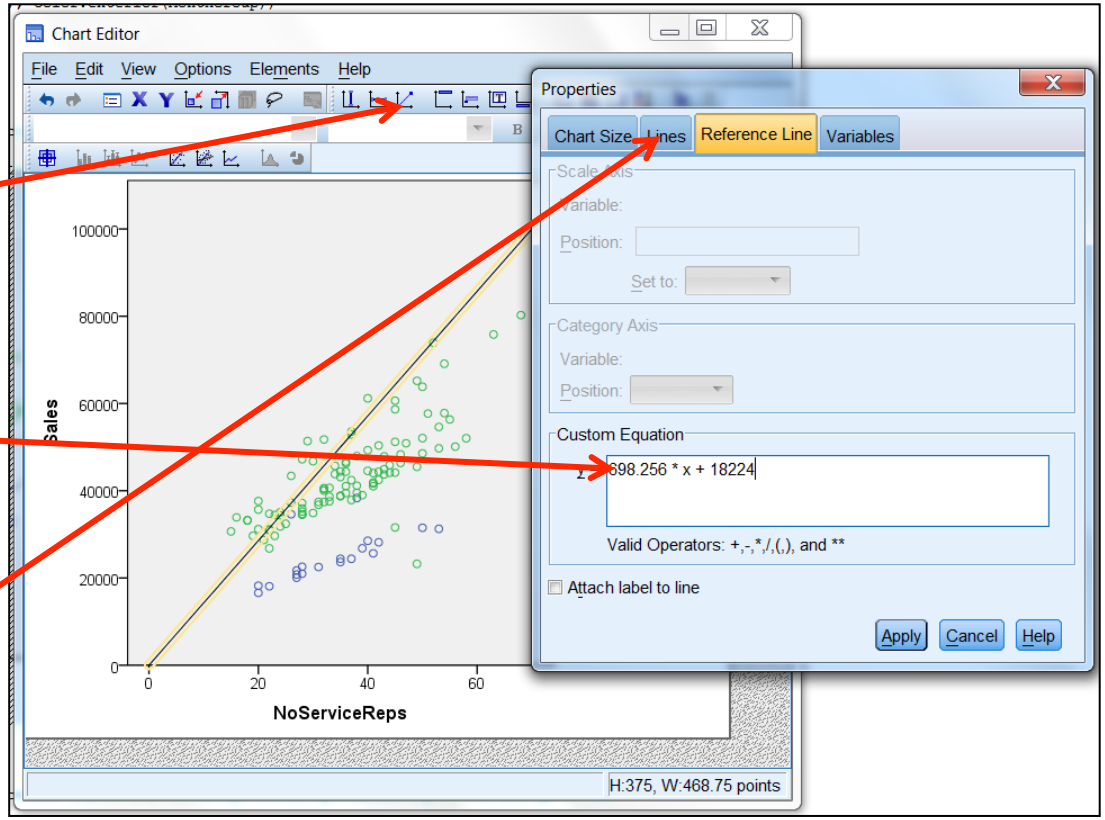
www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Add a fitted line to a grouped scatter plot

- Open the chart editor
- Select the Add a reference line from Equation tool
- Enter the Custom Equation from the ANCOVA output
- Change the line colour using the Lines tab
- Repeat for the other month group by subtracting the ANCOVA coefficient from the constant



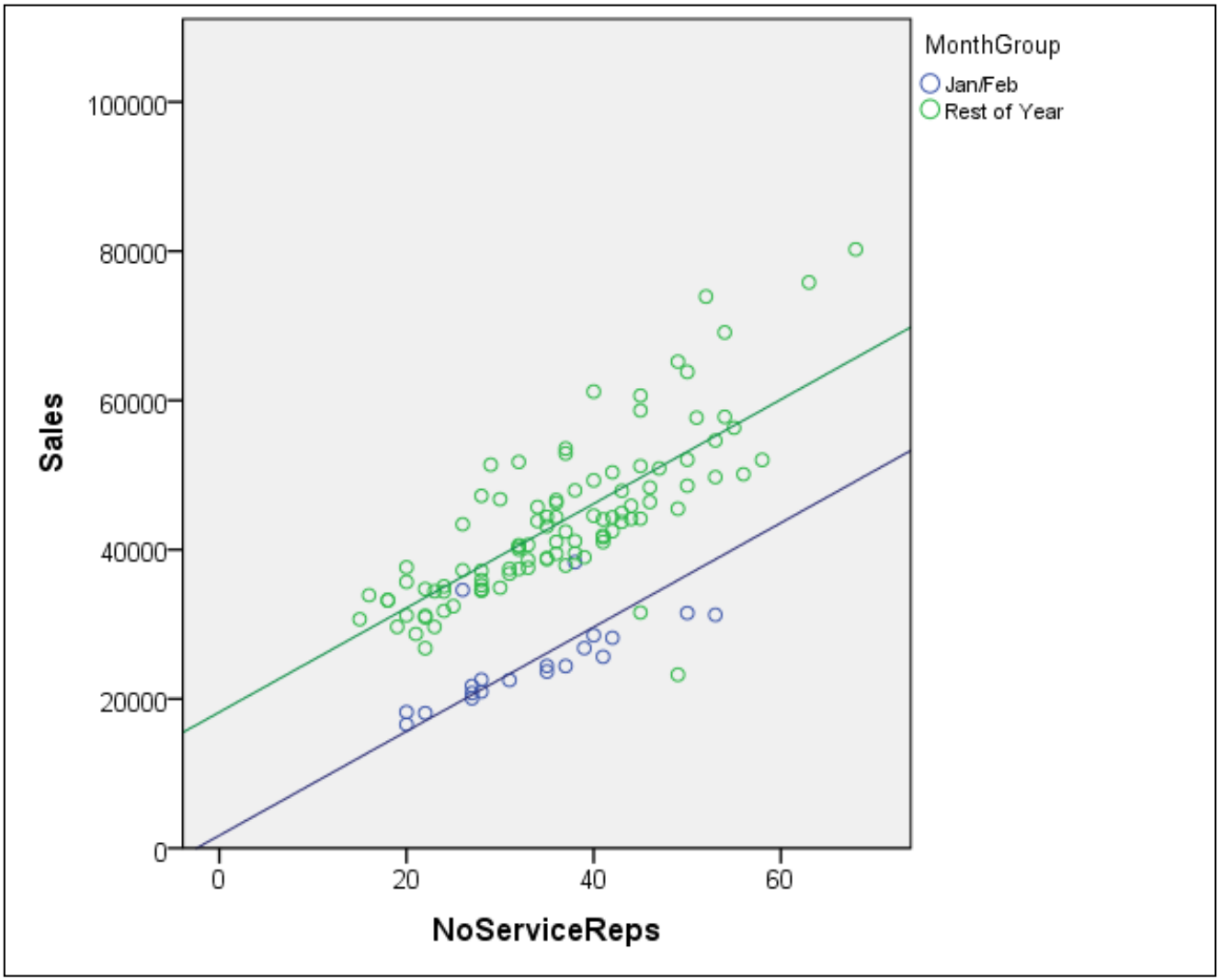
www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Scatter plot with fitted lines

Clearly an improvement on Simple Linear Regression but the best fit slope for the two groups is slightly different



www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

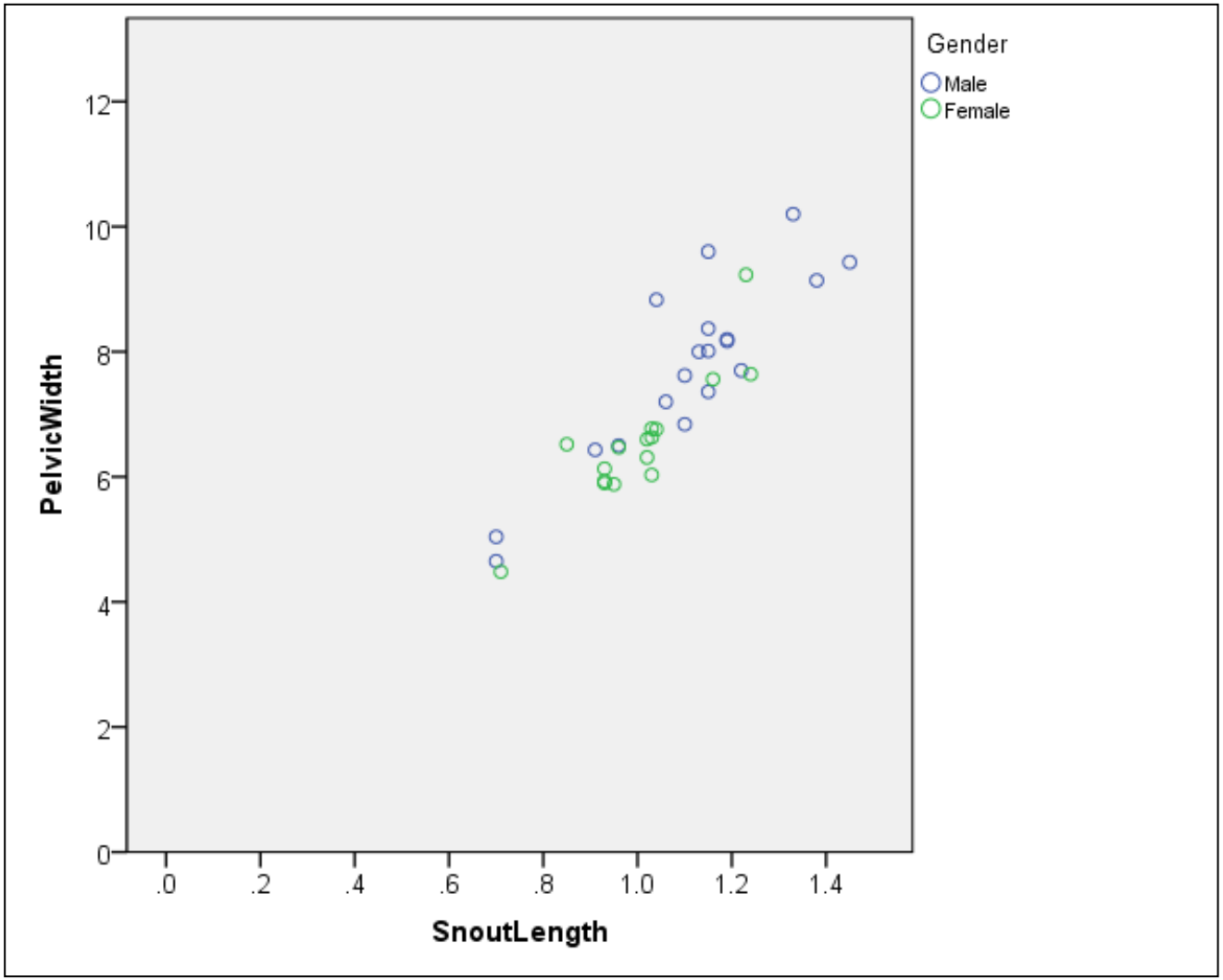
Activity

- Open the file AlligatorData.xlsx associated with this Workshop
- The data file contains pelvic canal width, snout-vent length and the gender of 35 alligators
- Create a scatterplot for *PelvicWidth* against *SnoutLength* for the different *Gender* groups
- Run an ACOVA model for *PelvicWidth* against *SnoutLength* for the different *Gender* groups
- Plot the ANCOVA model on the scatterplot

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Scatterplot of alligator data



www.**stats**tutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

ANCOVA for alligator data

Tests of Between-Subjects Effects

Dependent Variable:PelvicWidth

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	53.887[a]	2	26.943	72.108	.000
Intercept	.018	1	.018	.049	.826
SnoutLength	41.388	1	41.388	110.765	.000
Gender	2.016	1	2.016	5.395	.027
Error	11.957	32	.374		
Total	1882.116	35			
Corrected Total	65.844	34			

a. R Squared = .818 (Adjusted R Squared = .807)

Parameter Estimates

Dependent Variable:PelvicWidth

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	.402	.712	.564	.576	-1.049	1.853
SnoutLength	6.631	.630	10.524	.000	5.348	7.915
[Gender=female]	-.506	.218	-2.323	.027	-.949	-.062
[Gender=male]	0[a]	.	.	.	.	.

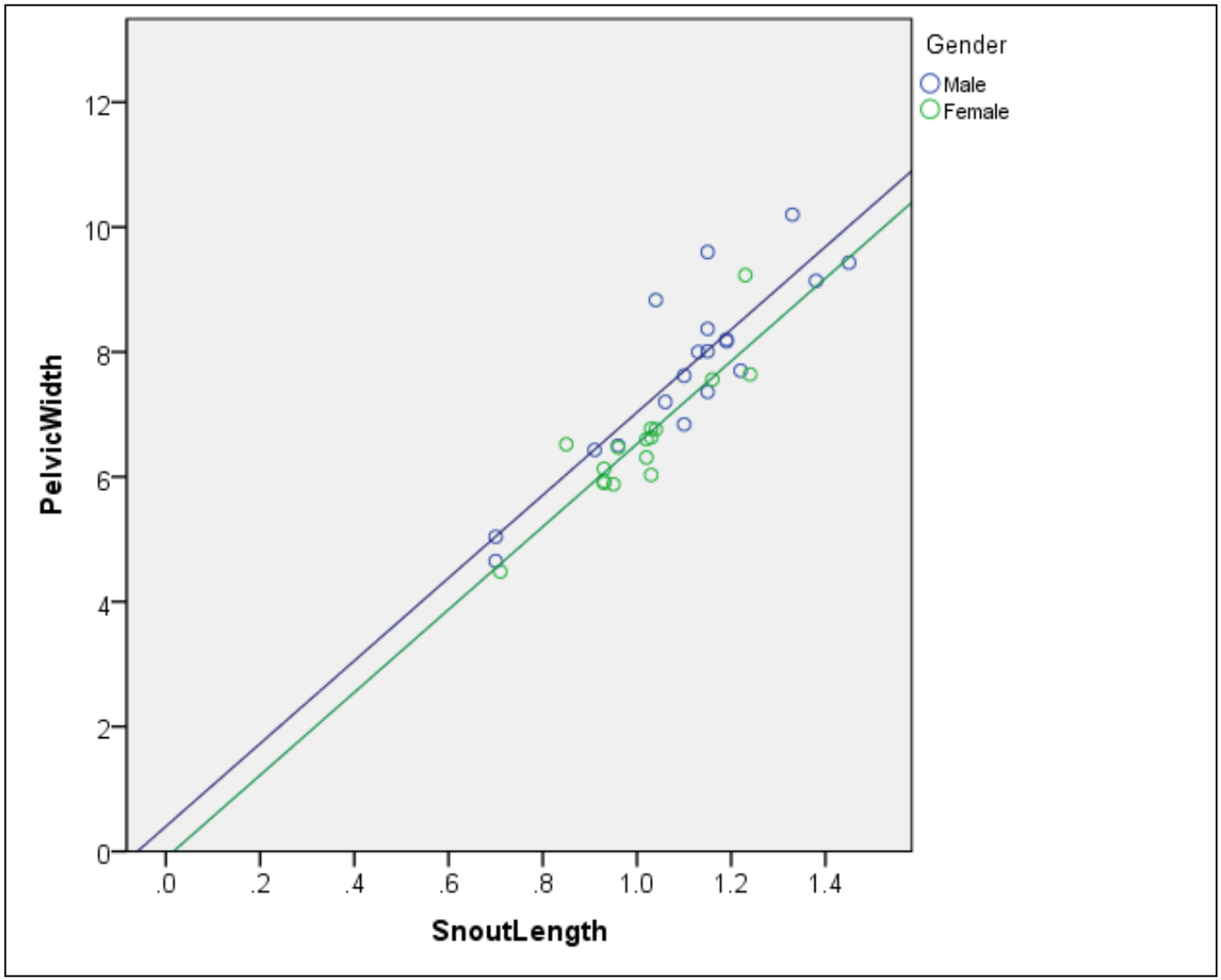
a. This parameter is set to zero because it is redundant.

www.**statstutor**.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Plot of ANCOVA on scatterplot



www.**stats**tutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Recap

- We have introduced Multiple Linear Regression by adding one predictor variable then several predictor variables to Simple Linear Regression
- New diagnostics were required in order to address the possibility of multicollinearity, namely VIF or Tolerance
- Better to use a systematic method for introducing or removing variables, such as Backwards
- Robustness arguments are similar to those in Simple Linear Regression
- Analysis of Covariance (ANCOVA) is a combination of ANOVA and Linear Regression

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Bibliography

Bovas, A. & Ledolter, J. (2006) *Introduction to Regression Modelling.* Belmont, CA: Thomson Brooks/Cole.

Field, A. (2013) *Discovering Statistics using SPSS: (And sex and drugs and rock 'n' roll)*, 4th ed., London: SAGE, Sections 8.5 - 8.7 and Chapter 12.

Kleinbaum, D., Kupper L., Muller, K. and Nizam, A. (2008) *Applied Regression Analysis and Other Multivariable Methods*. 4th ed. Belmont, CA: Thomson Brooks/Cole.

Kutner, M., Nachtsheim, C. and Neter, J. (2004) *Applied Linear Regression Models*. 4th ed., Irwin: McGraw-Hill.

Myers, R. (1990) *Classical and Modern Regression with Applications*. 2nd ed. Boston, MA: Duxbury.

statstutor (n. d.) *Multiple Regression resources*. Available at: http://www.statstutor.ac.uk/topics/regression-and-model-building/multiple-regression/ [Accessed 8/01/14].

Stirling, W. D. (2013) *Welcome to the General CAST e-book*. Available at: http://cast.massey.ac.nz/core/index.html?book=general [Accessed 8/01/14], Section 6.3.7.

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield