

11

Model selection II: Datasets with several explanatory variables

11.1 Economy of variables in the context of multiple regression

R-squared and adjusted R-squared

The models fitted in this section can be fitted using the `glm` command as in previous chapters. However, if all explanatory variables are continuous, using the `regression` command will also provide you with the R^2 and adjusted R^2 directly, as shown below with the *Peru* dataset:

MINITAB COMMANDS FOR BOX 11.1(a) Multiple regression for blood pressure

Commands `regress SYSTOL 2 YEARS WEIGHT;`
`brief 1.`

Menu route Stat > Regression > Regression
SYSTOL → Response
YEARS WEIGHT → Predictors

Results...

⊙ Regression equation, table of coefficients, s,
R-squared, and basic analysis of variance.

MINITAB OUTPUT FOR BOX 11.1 (a) Multiple regression for blood pressure

Regression Analysis: SYSTOL versus YEARS, WEIGHT

The regression equation is

SYSTOL = 50.3 - 0.572 YEARS + 1.35 WEIGHT

Predictor	Coef	SE Coef	T	P
Constant	50.32	15.82	3.18	0.003
YEARS	-0.5718	0.1879	-3.04	0.004
WEIGHT	1.3541	0.2672	5.07	0.000

S = 10.25 R-Sq = 42.1% R-Sq(adj) = 38.9%

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	2	2748.3	1374.1	13.08	0.000
Residual Error	36	3783.2	105.1		
Total	38	6531.4			

Prediction intervals

A prediction interval refers to a specific point. In this example, we aim to predict the fitted value, with 95% confidence, for an individual of weight 87 kg, who migrated to a lower altitude 40 years earlier. This is done as an extra option of the previous analysis, producing additional output.

MINITAB COMMANDS FOR BOX 11.2 A prediction interval using a model with two explanatory variables	
Commands	<pre>Regress SYSTOL 2 YEARS WEIGHT; predict 40 87; brief 1.</pre>
Menu route	Stat > Regression > Regression SYSTOL → Response YEARS WEIGHT → Predictors <input type="button" value="Results..."/> <input checked="" type="radio"/> Regression equation, table of coefficients, s, R-squared, and basic analysis of variance <input type="button" value="Options..."/> 40 87 → Prediction intervals for new observations

MINITAB OUTPUT FOR BOX 11.2 A prediction interval using a model with two explanatory variables				
Predicted Values for New Observations				
New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	145.25	6.06	(132.96, 157.54)	(121.10, 169.40) X
X denotes a row with X values away from the centre				
Values of Predictors for New Observations				
New Obs	YEARS	WEIGHT		
1	40.0	87.0		

No new commands are then introduced until Section 11.3.

11.3 Automated model selection procedures

We have chosen the values of 0.05 for alpha to enter and remove. Alpha is the critical p -value below which you choose to reject the null hypothesis—thus by setting it at 0.05, we are adopting the convention we normally follow. In a multiple regression analysis, we may wish to make our conditions more stringent (see main text). Note that because our list of potential explanatory variables overruns the line, the & sign is used to carry the model over to the following line.

MINITAB COMMANDS FOR BOX 11.6 Forwards stepwise regression of the whale watching dataset

Commands stepwise LRGWHAL 11 TRIPID YEAR MONTH DAY NPASS COULD8AM RAIN8AM &
 VIS8AM RAIN DURNTOT VIS;
 aenter 0.05;
 aremove 0.05.

Menu route Stat > Regression > Stepwise
 LRGWHAL → Response
 TRIPID YEAR MONTH DAY NPASS COULD8AM
 RAIN8AM VIS8AM RAIN DURNTOT VIS → Predictors

 0.05 → Alpha to enter
 0.05 → Alpha to remove

MINITAB OUTPUT FOR BOX 11.6 Forwards stepwise regression of the whale watching dataset

Stepwise Regression: LRGWHAL versus TRIPID, YEAR, ...

Alpha-to-Enter: 0.05 Alpha-to-Remove: 0.05

Response is LRGWHAL on 11 predictors, with N = 232

Step	1	2	3
Constant	-4.525	-4.555	-4.641
VIS	0.1252	0.1041	0.1056
T-Value	13.91	7.63	7.79
P-Value	0.000	0.000	0.000
VIS8AM		0.029	0.037
T-Value		2.05	2.56
P-Value		0.042	0.011
RAIN			0.146
T-Value			2.18
P-Value			0.031

S	0.562	0.559	0.554
R-Sq	45.67	46.65	47.73
R-Sq (adj)	45.44	46.18	47.04
C-p	8.2	6.0	3.2

In the main stepwise window, there is also a pane for ‘predictors to include in every model’. If previous studies have indicated that specific variables are important, then these would be entered into this pane (see main text).

The alternative, and rather more satisfactory, way of analysing this dataset is interactively with general linear models, as explained in the main text.

11.6 Exercises

Finding the best treatment for cat fleas

Three models are fitted in the main text.

MINITAB COMMANDS FOR BOX 11.10 (a) Cat flea analysis with minimal model	
Commands	<pre>glm LOGFLEAS = TRTMT; brief 1 .</pre>
Menu route	Stat > ANOVA > General Linear Model LOGFLEAS → Response TRTMT → Model <div style="border: 1px solid black; display: inline-block; padding: 2px;">Results...</div> ☉ Analysis of variance table

MINITAB OUTPUT FOR BOX 11.10 (a) Cat flea analysis with minimal model						
General Linear Model: LOGFLEAS versus TRTMT						
Factor	Type	Levels	Values			
TRTMT	fixed	2	1 2			
Analysis of Variance for LOGFLEAS, using Adjusted SS for Tests						
Source	DF	Seq SS	Adj SS	Adj MS	F	P
TRTMT	1	1.6118	1.6118	1.6118	1.68	0.199
Error	87	83.5915	83.5915	0.9608		
Total	88	85.2033				

MINITAB COMMANDS FOR BOX 11.10 (b) **Cat flea analysis with full model for FLEAS**

Commands `glm FLEAS = TRTMT + HAIRL + NCATS + CARPET;`
 `covariates HAIRL NCATS;`
 `brief 1 .`

Menu route `Stat > ANOVA > General Linear Model`
 `FLEAS → Response`
 `TRTMT + HAIRL + NCATS + CARPET → Model`

Analysis of variance table

`HAIRL NCATS → Covariates`

MINITAB OUTPUT FOR BOX 11.10 (b) **Cat flea analysis with full model for FLEAS**

General Linear Model: FLEAS versus TRTMT, CARPET

Factor	Type	Levels	Values
TRTMT	fixed	2	1 2
CARPET	fixed	2	1 2

Analysis of Variance for FLEAS, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
TRTMT	1	1412	18438	18438	2.72	0.103
HAIRL	1	17463	4	4	0.00	0.981
NCATS	1	188947	188994	188994	27.90	0.000
CARPET	1	32498	32498	32498	4.80	0.031
Error	84	568947	568947	6773		
Total	88	809267				

MINITAB COMMANDS FOR BOX 11.10 (c) **Cat flea analysis with full model for LOGFLEAS**

Commands `glm LOGFLEAS = TRTMT + HAIRL + NCATS + CARPET;`
 `covariates HAIRL NCATS;`
 `brief 1 .`

Menu route `Stat > ANOVA > General Linear Model`
 `LOGFLEAS → Response`
 `TRTMT + HAIRL + NCATS + CARPET → Model`

Analysis of variance table

`HAIRL NCATS → Covariates`

MINITAB COMMANDS FOR BOX 11.10(c) Cat flea analysis with full model for LOGFLEAS						
General Linear Model: LOGFLEAS versus TRTMT, CARPET						
Factor	Type	Levels	Values			
TRTMT	fixed	2	1 2			
CARPET	fixed	2	1 2			
Analysis of Variance for LOGFLEAS, using Adjusted SS for Tests						
Source	DF	Seq SS	Adj SS	Adj MS	F	P
TRTMT	1	1.6118	5.7593	5.7593	9.03	0.004
HAIRL	1	1.1523	0.1642	0.1642	0.26	0.613
NCATS	1	22.5843	22.5915	22.5915	35.41	0.000
CARPET	1	6.2605	6.2605	6.2605	9.81	0.002
Error	84	53.5944	53.5944	0.6380		
Total	88	85.2033				

For the final model ... see the answers to exercises in Chapter 14.

Multiplicity of p-values

The Minitab output for this exercise may be found in the answers to exercises in Chapter 14.