

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 or REGRESSION commands, as in previous chapters, and both provide R^2 and adjusted R^2 . Here is the *Peru* dataset analysed with GLM:

SPSS COMMANDS FOR BOX 11.1(a) Multiple regression for blood pressure	
Syntax	<pre>glm SYSTOL with YEARS WEIGHT /design YEARS WEIGHT.</pre>
Menu route	Analyze > General Linear Model > Univariate SYSTOL → Dependent Variable YEARS WEIGHT → Covariate(s)

SPSS OUTPUT FOR BOX 11.1(a) **Multiple regression for blood pressure****General linear model****Tests of Between-Subjects Effects**

Dependent Variable: SYSTOL

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2748.279 ^a	2	1374.139	13.076	.000
Intercept	1063.392	1	1063.392	10.119	.003
YEARS	972.899	1	972.899	9.258	.004
WEIGHT	2698.295	1	2698.295	25.677	.000
Error	3783.157	36	105.088		
Total	639633.000	39			
Corrected Total	6531.436	38			

a. R Squared = .421 (Adjusted R Squared = .389)

Prediction intervals

A prediction interval refers to a specific point. In the main text, we predict the fitted value, with 95% confidence, for an individual of weight 87 kg, who migrated to a lower altitude 40 years earlier. SPSS does not provide this kind of prediction interval.

SPSS does, however, provide the information to calculate another kind of prediction interval, namely, a 95% prediction interval for the mean value of SYSTOL for given values of WEIGHT and YEARS. Both types of interval take into account the uncertainty in the estimation of intercept and slopes in the regression. However, when predicting an interval for the mean, those are all the uncertainties we take into account. When predicting an interval for an individual datapoint, the scatter around the best fitting regression equation is also taken into account. Thus the interval for the mean, which we are about to obtain from SPSS, will be narrower than the interval for an individual datapoint as shown in the main text.

SPSS provides the predicted value (\hat{y} the fitted value) for each datapoint, based on the values of its x-variables, and also provides a standard error for that predicted value ($SE_{\hat{y}}$ arising from uncertainty in the intercept and slope of the fitted line). Following the general method for confidence intervals as discussed in the revision section (R1.3), we multiply the standard error by a critical t-ratio for the error degrees of freedom. This uses the SPSS function 'IDF.T' with parameters of 0.975 for area to the left (as we want a 95% confidence interval with 2.5% on either side) and 36 for the error degrees of freedom. The boxes show how to carry out the instructions, obtaining the lower and upper limits for the prediction interval for each predicted value in two new columns.

SPSS COMMANDS FOR BOX 11.2

A prediction interval using a model with two explanatory variables

```
Syntax      glm SYSTOL with YEARS WEIGHT
            /save pred(SYSPRED) sepred(SYSPSE)
            /design YEARS WEIGHT.

            compute LPIS=SYSPRED-idf.t(0.975,36)*SYSPSE .
            compute UPIS=SYSPRED+idf.t(0.975,36)*SYSPSE .

            execute.
```

```
Menu route  Analyze > General Linear Model > Univariate
            SYSTOL → Dependent Variable
            YEARS WEIGHT → Covariate(s)
            Save
            Predicted Values:
             Unstandardized
             Standard error

            Transform > Compute
            LPIS → Target variable
            PRE_1-idf.t(0.975,36)*SEP_1 → Numeric expression

            Transform > Compute
            UPIS → Target variable
            PRE_1+idf.t(0.975,36)*SEP_1 → Numeric expression
```

In addition to the ANOVA table, four variables will be added to the SPSS datasheet. LPIS and UPIS give the lower and upper prediction intervals for the expected value of each datapoint. Looking at the last datapoint, with an expected value of 145.25, SPSS gives the prediction interval (for a predicted value) of 132.96 – 157.54. As we explained at the beginning of this section, this interval is narrower than the prediction interval (for a datapoint) given in the main text, of 121.10 – 169.40.

SPSS OUTPUT FOR BOX 11.2

Predicted values with standard errors using a model with two explanatory variables

YEARS	WEIGHT	SYSTOL	PRE_1	SEP_1	LPIS	UPIS
1	71	170	145.89	4.35	137.05	154.72
6	57	120	123.39	2.40	118.53	128.26
5	56	125	123.29	2.52	118.18	128.39
1	61	148	132.35	2.87	126.52	138.18
1	65	140	137.76	3.29	131.09	144.44
19	62	106	123.41	1.92	119.52	127.29
5	53	120	119.23	2.94	113.27	125.19
25	53	108	107.79	4.34	98.99	116.59
6	65	124	134.90	2.53	129.77	140.04
13	57	134	120.07	2.23	115.54	124.6
13	67	116	132.93	1.97	128.94	136.93
10	59	114	124.63	1.93	120.71	128.55
15	64	130	128.40	1.65	125.05	131.76
18	70	118	134.13	2.22	129.63	138.64
2	64	138	135.84	3.00	129.75	141.92
12	57	134	119.96	2.29	115.32	124.61
15	57	120	118.92	2.34	114.18	123.67
16	55	120	115.64	2.83	109.91	121.38
17	57	114	117.78	2.50	112.71	122.85
10	58	124	123.14	2.05	118.98	127.3
18	60	114	120.59	2.14	116.24	124.94
11	61	136	126.63	1.77	123.04	130.22
11	57	126	121.21	2.19	116.78	125.65
21	58	124	116.17	2.84	110.41	121.93
24	74	128	136.80	3.04	130.62	142.97
14	72	134	139.81	2.93	133.85	145.76
25	63	112	120.65	2.60	115.38	125.93
32	68	128	124.10	3.29	117.42	130.78
5	63	134	133.31	2.48	128.27	138.34
12	68	128	135.53	2.30	130.88	140.19
25	69	140	129.45	2.44	124.51	134.4
26	73	138	134.30	2.95	128.31	140.29
10	64	118	131.26	1.93	127.34	135.18
19	65	110	127.47	1.79	123.84	131.1
18	71	142	136.17	2.50	131.11	141.23
10	60	134	126.12	1.85	122.36	129.88
1	55	116	124.22	2.95	118.24	130.2
43	70	132	120.52	4.99	110.4	130.63
40	87	152	145.25	6.06	132.96	157.54

To predict an interval for a set of x -variables that does not appear in the dataset is slightly more complicated. We need to add the values for which we wish to predict an interval to the bottom of the columns for the existing x -variables, but leave the y -variable column missing for those new pseudo-datapoints. Then the same commands as shown in the analysis will omit the new datapoint(s) from the regression, because their y -values are missing, but will give predictions and prediction intervals for all the fitted values in the new rows.

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). The REGRESSION command we have already been using will also do stepwise regressions. (Note that in SPSS's terminology, this is a FORWARD regression. SPSS uses the term STEPWISE to indicate a more complicated procedure in which variables can be added or subtracted at each step.) The entry value of 0.05 is the default, so setting is unnecessary. We show the subcommand and subdialog by which this level is set in case you wish to use a different criterion yourself.

SPSS COMMANDS FOR BOX 11.6

Forwards stepwise regression of the whale watching dataset

Syntax regression
 /criteria=pin(.05)
 /dependent LRGWHAL
 /method=forward TRIPID YEAR MONTH DAY NPASS
 CLOUD8AM RAIN8AM VIS8AM RAIN VIS DURNTOT.

Menu route Analyze > Regression > Linear
 LRGWHAL → Dependent
 TRIPID YEAR MONTH DAY NPASS CLOUD8AM RAIN8AM
 VIS8AM RAIN VIS DURNTOT → Independent(s)
 Method:

 Use probability of F
 0.05 → Entry

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

Regression

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	VIS	.	Forward (Criterion: Probability- of-F-to-e nter <= .050)
2	VIS8AM	.	Forward (Criterion: Probability- of-F-to-e nter <= .050)
3	RAIN	.	Forward (Criterion: Probability- of-F-to-e nter <= .050)

a. Dependent Variable: LRGWHAL

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.676 ^a	.457	.454	*****
2	.683 ^b	.466	.462	*****
3	.691 ^c	.477	.470	*****

a. Predictors: (Constant), VIS

b. Predictors: (Constant), VIS, VIS8AM

c. Predictors: (Constant), VIS, VIS8AM, RAIN

(Contd.)

SPSS OUTPUT FOR BOX 11.6 (Contd.)

ANOVA^d

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	61.166	1	61.166	193.353	.000 ^a
	Residual	72.759	230	.316		
	Total	133.925	231			
2	Regression	62.472	2	31.236	100.110	.000 ^b
	Residual	71.452	229	.312		
	Total	133.925	231			
3	Regression	63.925	3	21.308	69.406	.000 ^c
	Residual	69.999	228	.307		
	Total	133.925	231			

- a. Predictors: (Constant), VIS
 b. Predictors: (Constant), VIS, VIS8AM
 c. Predictors: (Constant), VIS, VIS8AM, RAIN
 d. Dependent Variable: LRGWHAL

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4.525	.061		-73.981	.000
	VIS	.125	.009	.676	13.905	.000
2	(Constant)	-4.555	.063		-72.849	.000
	VIS	.104	.014	.562	7.631	.000
	VIS8AM	2.850E-02	.014	.151	2.046	.042
3	(Constant)	-4.641	.074		-63.040	.000
	VIS	.106	.014	.570	7.792	.000
	VIS8AM	3.665E-02	.014	.194	2.560	.011
	RAIN	.146	.067	.115	2.176	.031

- a. Dependent Variable: LRGWHAL

(Contd.)

SPSS OUTPUT FOR BOX 11.6 (Contd.)

Excluded Variables^d

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics	
					Tolerance	
1	TRIPID	-.039 ^a	-.795	.427	-.052	.993
	YEAR	-.029 ^a	-.594	.553	-.039	.995
	MONTH	-.062 ^a	-1.272	.205	-.084	1.000
	DAY	.024 ^a	.492	.623	.032	.986
	NPASS	-.043 ^a	-.879	.380	-.058	1.000
	CLOUD8AM	.090 ^a	1.735	.084	.114	.863
	RAIN8AM	.075 ^a	1.286	.200	.085	.686
	VIS8AM	.151 ^a	2.046	.042	.134	.430
	RAIN	.080 ^a	1.541	.125	.101	.877
	DURNTOT	.054 ^a	1.101	.272	.073	.998
2	TRIPID	-.043 ^b	-.881	.379	-.058	.991
	YEAR	-.031 ^b	-.632	.528	-.042	.994
	MONTH	-.065 ^b	-1.346	.180	-.089	.999
	DAY	.021 ^b	.437	.663	.029	.985
	NPASS	-.040 ^b	-.832	.406	-.055	.999
	CLOUD8AM	.104 ^b	1.997	.047	.131	.852
	RAIN8AM	.097 ^b	1.642	.102	.108	.669
	RAIN	.115 ^b	2.176	.031	.143	.817
	DURNTOT	.045 ^b	.921	.358	.061	.989
	3	TRIPID	-.057 ^c	-1.181	.239	-.078
YEAR		-.042 ^c	-.872	.384	-.058	.983
MONTH		-.056 ^c	-1.176	.241	-.078	.992
DAY		.010 ^c	.211	.833	.014	.974
NPASS		-.050 ^c	-1.049	.295	-.069	.990
CLOUD8AM		.069 ^c	1.217	.225	.081	.709
RAIN8AM		.050 ^c	.757	.450	.050	.536
DURNTOT		.031 ^c	.640	.523	.042	.971

a. Predictors in the Model: (Constant), VIS

b. Predictors in the Model: (Constant), VIS, VIS8AM

Sometimes you may wish to force some variables to be included, while allowing stepwise procedures on other variables. We show here the commands for forcing MONTH and YEAR to be included while following a forwards stepwise procedure for the other variables. SPSS allows different 'blocks' of x-variables to have different methods associated with them.

SPSS COMMANDS SHOWING HOW TO FORCE SOME VARIABLES TO BE INCLUDED	
Syntax	<pre> regression /dependent LRGWHAL /method=enter YEAR MONTH /method=forward TRIPID DAY NPASS CLOUD8AM RAIN8AM VIS8AM RAIN VIS DURNTOT . </pre>
Menu route	<pre> Analyze > Regression > Linear LRGWHAL → Dependent YEAR MONTH → Independent(s) Method: <input type="button" value="Enter ▼"/> <input type="button" value="Next"/> TRIPID DAY NPASS CLOUD8AM RAIN8AM VIS8AM RAIN VIS DURNTOT → Independent(s) Method: <input type="button" value="Forward ▼"/> </pre>

Just the ANOVA table is reproduced from the output, below, illustrating that YEAR and MONTH are automatically included as the first two variables of the model, with any further explanatory variables being selected by SPSS as before.

SPSS OUTPUT SHOWING HOW TO FORCE SOME VARIABLES TO BE INCLUDED

ANOVA^e

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.544	2	.272	.467	.627 ^a
	Residual	133.380	229	.582		
	Total	133.925	231			
2	Regression	61.772	3	20.591	65.067	.000 ^b
	Residual	72.152	228	.316		
	Total	133.925	231			
3	Regression	63.143	4	15.786	50.626	.000 ^c
	Residual	70.781	227	.312		
	Total	133.925	231			
4	Regression	64.557	5	12.911	42.065	.000 ^d
	Residual	69.368	226	.307		
	Total	133.925	231			

a. Predictors: (Constant), MONTH, YEAR

b. Predictors: (Constant), MONTH, YEAR, VIS

c. Predictors: (Constant), MONTH, YEAR, VIS, VIS8AM

d. Predictors: (Constant), MONTH, YEAR, VIS, VIS8AM, RAIN

e. Dependent Variable: LRGWHAL

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.

SPSS COMMANDS FOR BOX 11.10(a) Cat flea analysis with minimal model	
Syntax	glm LOGFLEAS by TRTMT /design TRTMT.
Menu route	Analyze > General Linear Model > Univariate LOGFLEAS → Dependent Variable TRTMT → Fixed Factor(s)

SPSS OUTPUT FOR BOX 11.10(a) Cat flea analysis with minimal model					
General linear model					
Between-Subjects Factors					
	N				
TRTMT 1	40				
2	49				
Tests of Between-Subjects Effects					
Dependent Variable: LOGFLEAS					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1.612 ^a	1	1.612	1.678	.199
Intercept	1435.703	1	1435.703	1494.245	.000
TRTMT	1.612	1	1.612	1.678	.199
Error	83.591	87	.961		
Total	1525.926	89			
Corrected Total	85.203	88			
a. R Squared = .019 (Adjusted R Squared = .008)					

SPSS COMMANDS FOR BOX 11.10(b) Cat flea analysis with full model for FLEAS	
Syntax	glm FLEAS with HAIRL NCATS by TRTMT CARPET /design TRTMT HAIRL NCATS CARPET.
Menu route	Analyze > General Linear Model > Univariate FLEAS → Dependent Variable TRTMT CARPET → Fixed Factor(s) HAIRL NCATS → Covariate(s)

SPSS OUTPUT FOR BOX 11.10(b) Cat flea analysis with full model for FLEAS					
General linear model					
Between-Subjects Factors					
	N				
TRTMT 1	40				
2	49				
CARPET 1	54				
2	35				
Tests of Between-Subjects Effects					
Dependent Variable: FLEAS					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	240319.673 ^a	4	60079.918	8.870	.000
Intercept	143.177	1	143.177	.021	.885
TRTMT	18437.993	1	18437.993	2.722	.103
HAIRL	3.779	1	3.779	.001	.981
NCATS	188994.214	1	188994.214	27.903	.000
CARPET	32497.752	1	32497.752	4.798	.031
Error	568947.226	84	6773.181		
Total	1513702.000	89			
Corrected Total	809266.899	88			
a. R Squared = .297 (Adjusted R Squared = .263)					

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

Syntax glm LOGFLEAS with HAIRL NCATS by TRTMT CARPET
 /design TRTMT HAIRL NCATS CARPET.

Menu route Analyze > General Linear Model > Univariate
 LOGFLEAS → Dependent Variable
 TRTMT CARPET → Fixed Factor(s)
 HAIRL NCATS → Covariate(s)

SPSS OUTPUT FOR BOX 11.10(c) **Cat flea analysis with full model for LOGFLEAS****General linear model****Between-Subjects Factors**

		N
TRTMT	1	40
	2	49
CARPET	1	54
	2	35

Tests of Between-Subjects Effects

Dependent Variable: LOGFLEAS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	31.609 ^a	4	7.902	12.385	.000
Intercept	25.101	1	25.101	39.341	.000
TRTMT	5.759	1	5.759	9.027	.004
HAIRL	.164	1	.164	.257	.613
NCATS	22.592	1	22.592	35.408	.000
CARPET	6.260	1	6.260	9.812	.002
Error	53.594	84	.638		
Total	1525.926	89			
Corrected Total	85.203	88			

a. R Squared = .371 (Adjusted R Squared = .341)

For the final model...see the answers to exercises.

Multiplicity of p-values

The SPSS output for this exercise may be found in the answers to exercises.