

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 `proc glm` 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:

SAS COMMANDS FOR BOX 11.1 (A) Multiple regression for blood pressure	
Commands	<pre>proc reg data=gandh.Chapter11; model SYSTOL = YEARS WEIGHT; run;</pre>
Menu route	Statistics > Regression > Linear SYSTOL → Dependent YEARS WEIGHT → Explanatory

SAS OUTPUT FOR BOX 11.1 (A) Multiple regression for blood pressure					
The REG Procedure					
Model: MODEL1					
Dependent Variable: SYSTOL					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2748.27852	1374.13926	13.08	<.0001
Error	36	3783.15738	105.08770		
Corrected Total	38	6531.43590			
	Root MSE	10.25123	R-Square	0.4208	
	Dependent Mean	127.41026	Adj R-Sq	0.3886	
	Coeff Var	8.04584			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	50.31913	15.81839	3.18	0.0030
YEARS	1	-0.57184	0.18794	-3.04	0.0044
WEIGHT	1	1.35408	0.26722	5.07	<.0001

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. This is easy in SAS because these are the WEIGHT and YEARS of an individual in the dataset, specifically individual 39. It is easy to ask SAS to provide predictions and prediction intervals for datapoints in the dataset, and this is what is done in the analyses below. (To predict for a non-existent set of x -variables, we need to add a datapoint or datapoints to the dataset, which have the x -variables we wish to predict, but with a missing y -variable. Then the same commands as shown in the analysis will omit the new datapoint(s), as their y -values are missing, but will give predictions and prediction intervals for all the datapoints.)

In the command route, we include the commands for the lower (l95m=lowerpm) and upper (u95m=upperpm) bounds for predicting the mean SYSTOL for WEIGHT = 87, YEARS = 40, as well as the lower (l95=lowerp) and upper (u95=upperp) bounds for predicting a single datapoint's SYSTOL. The former take no account of scatter round the line, and represent only the uncertainty in prediction resulting from uncertainty in the parameter estimates. The latter in addition take into account the variability around the regression line, and so are inevitably further from the predicted value itself, making a wider interval. The output dataset will be found in the work folder, and called data1. As usual, the entire dataset is displayed with the prediction interval variables at the end.

In the menu route, the Predictions button is tempting, but it gives only the predictions for the mean, and not for an individual datapoint, and doesn't make this clear. So for our purposes, the Save Data button is the useful one, with which we can request the same information as we did via the command route, and obtain the same output. To see the values, find the most recent 'Diagnostics Table' in the left hand pane of the Analyst window. Double-click, and the dataset will be displayed, with the confidence interval variables at the end.

Remember in both cases, that it is individual 39 who has the x -variables we wish to predict for, so the final row will hold the relevant output.

SAS COMMANDS FOR BOX 11.2

A prediction interval using a model with two explanatory variables

```
Commands  proc reg data=gandh.Chapter11;
           model SYSTOL = YEARS WEIGHT;
           output p=pred l95m=lowerpm u95m=upperpm
                l95=lowerp u95=upperp;
           run;
           quit;
```

Menu route Statistics > Regression > Linear

 SYSTOL → Dependent

 YEARS WEIGHT → Explanatory

Create and save diagnostics data

 PREDICTED, L95, U95, L95M, U95M → Add

SAS OUTPUT FOR BOX 11.2 An excerpt from the output dataset with the prediction intervals from the model with two explanatory variables

YEARS	WEIGHT	SYSTOL	pred	lowerpm	upperpm	lowerp	upperp
1	71	170		145.8868	137.0548	154.7189	123.2981	168.4755
6	56.5	120		123.3935	118.5314	128.2556	102.042	144.7449
5	56	125		123.2883	118.1831	128.3935	101.8802	144.6964
1	61	148		132.346	126.5154	138.1766	110.7535	153.9386
1	65	140		137.7624	131.0877	144.437	115.9267	159.598
19	62	106		123.4069	119.5223	127.2915	102.2566	144.5572
5	53	120		119.226	113.2668	125.1852	97.59839	140.8537
25	53	108		107.7891	98.9919	116.5864	85.21405	130.3642
6	65	124		134.9031	129.7664	140.0398	113.4875	156.3187
13	57	134		120.0676	115.5364	124.5988	98.78908	141.3461
13	66.5	116		132.9313	128.9362	136.9265	111.7605	154.1022
10	59.1	114		124.6267	120.707	128.5464	103.47	145.7834
15	64	130		128.4024	125.0472	131.7577	107.343	149.4619
18	69.5	118		134.1343	129.6256	138.6431	112.8606	155.4081
2	64	138		135.8364	129.7524	141.9205	114.1741	157.4988
12	56.5	134		119.9624	115.3182	124.6066	98.65953	141.2653
15	57	120		118.9239	114.1754	123.6724	97.59806	140.2497
16	55	120		115.6439	109.9078	121.38	94.07666	137.2111
17	57	114		117.7802	112.708	122.8524	96.37997	139.1804
10	58	124		123.1372	118.9773	127.2971	101.9347	144.3397
18	59.5	114		120.5936	116.2433	124.9438	99.35285	141.8343
11	61	136		126.6276	123.0385	130.2167	105.5296	147.7256
11	57	126		121.2113	116.7754	125.6472	99.95286	142.4697
21	57.5	124		116.1699	110.4067	121.933	94.59542	137.7443
24	74	128		136.7966	130.6239	142.9694	115.1092	158.4841
14	72	134		139.8069	133.8548	145.7591	118.1812	161.4326
25	62.5	112		120.6529	115.3768	125.929	99.20339	142.1024
32	68	128		124.0974	117.4173	130.7775	102.2601	145.9347
5	63.4	134		133.3084	128.2726	138.3443	111.9168	154.7001
12	68	128		135.5343	130.8759	140.1927	114.2283	156.8402
25	69	140		129.4544	124.5068	134.4019	108.0833	150.8254
26	73	138		134.2989	128.3058	140.2919	112.6619	155.9358
10	64	118		131.2617	127.3434	135.18	110.1052	152.4181
19	65	110		127.4691	123.8424	131.0959	106.3647	148.5736
18	71	142		136.1655	131.1052	141.2257	114.768	157.5629
10	60.2	134		126.1162	122.3567	129.8757	104.9885	147.2438
1	55	116		124.2216	118.2393	130.2038	102.5876	145.8556
43	70	132		120.5153	110.3959	130.6346	97.39288	143.6376
40	87	152		145.2501	132.9616	157.5386	121.0995	169.4007

The five right hand columns of the above table will appear in the diagnostics table (or work.data1) on the extreme right hand side, beyond all other data.

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, which SAS calls 'sle' for 'significance level for entry'. 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). We also tell SAS that we wish only to add variables, and not to remove them.

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

```
Commands      proc reg data=gandh.Chapter11;
                model LRGWHAL= TRIPID YEAR MONTH DAY NPASS
                CLOUD8AM RAIN8AM VIS8AM RAIN VIS DURNTOT /
                selection=forward sle=0.05;
                run;
```

```
Menu route    Statistics > Regression > Linear ...
                LRGWHAL → Dependent
                TRIPID YEAR MONTH DAY NPASS CLOUD8AM
                RAIN8AM VIS8AM RAIN VIS DURNTOT →
                Explanatory
                Model
                ☉ Forward Selection
```

SAS OUTPUT FOR BOX 11.6

Forwards stepwise regression of the whale watching dataset

The REG Procedure
 Model: MODEL1
 Dependent Variable: LRGWHAL

Forward Selection: Step 1

Variable VIS Entered: R-Square = 0.4567 and C(p) = 8.1979

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	61.16584	61.16584	193.35	<.0001
Error	230	72.75867	0.31634		
Corrected Total	231	133.92451			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-4.52464	0.06116	1731.42296	5473.26	<.0001
VIS	0.12522	0.00901	61.16584	193.35	<.0001

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable VIS8AM Entered: R-Square = 0.4665 and C(p) = 5.9565

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	62.47236	31.23618	100.11	<.0001
Error	229	71.45215	0.31202		
Corrected Total	231	133.92451			

The REG Procedure
 Model: MODEL1
 Dependent Variable: LRGWHAL LRGWHAL

Forward Selection: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-4.55501	0.06253	1655.88739	5307.02	<.0001
VIS8AM	0.02850	0.01393	1.30652	4.19	0.0419
VIS	0.10413	0.01365	18.16959	58.23	<.0001

Bounds on condition number: 2.3279, 9.3118

-----*(Contd.)*

SAS OUTPUT FOR BOX 11.6 (Contd.)

Forward Selection: Step 3

Variable RAIN Entered: R-Square = 0.4773 and C(p) = 3.2393

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	63.92544	21.30848	69.41	<.0001
Error	228	69.99907	0.30701		
Corrected Total	231	133.92451			

Variable	Parameter Estimate	Standard Error	Type III SS	F Value	Pr > F
Intercept	-4.64132	0.07362	1220.08460	3974.04	<.0001
VIS8AM	0.03665	0.01432	2.01270	6.56	0.0111
RAIN	0.14555	0.06690	1.45308	4.73	0.0306
VIS	0.10560	0.01355	18.63935	60.71	<.0001

Bounds on condition number: 2.4992, 18.172

No other variable met the 0.0500 significance level for entry into the model.

The REG Procedure
 Model: MODEL1
 Dependent Variable: LRGWHAL LRGWHAL

Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	VIS	VIS	1	0.4567	0.4567	8.1979	193.35	<.0001
2	VIS8AM	VIS8AM	2	0.0098	0.4665	5.9565	4.19	0.0419
3	RAIN	RAIN	3	0.0109	0.4773	3.2393	4.73	0.0306

The INCLUDE option to the MODEL statement allows the first n of the x -variables to be forced into and retained in the model. Thus adding INCLUDE=3 after SLE=0.05 would have resulted in TRIPID, YEAR and MONTH being kept in all models considered (simply because they are the first three explanatory variables in the model formula). If previous studies have indicated that specific variables are important, then these would be put first and the INCLUDE option used to include them (see main text). The menu route uses the Include tab of the Model subdialog box for the same purpose.

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. If you wish to fit your variables in the same order as in the main text, then use the model button via the menu route, and the statistics button to get Type I and Type II sums of squares. As these are now established procedures we do not repeat them here.

SAS COMMANDS FOR BOX 11.10 (A) Cat flea analysis with minimal model	
Commands	<pre>proc glm data=gandh.Chapter11; class TRTMT; model LOGFLEAS = TRTMT; run;</pre>
Menu route	Statistics > Anova > Linear Models... LOGFLEAS → Dependent TRTMT → Class

SAS OUTPUT FOR BOX 11.10 (A): Cat flea analysis with minimal model					
The GLM Procedure					
Dependent Variable: LOGFLEAS					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.61179950	1.61179950	1.68	0.1987
Error	87	83.59146156	0.96082140		
Corrected Total	88	85.20326106			
	R-Square	Coeff Var	Root MSE	LOGFLEAS Mean	
	0.018917	24.36274	0.980215	4.023419	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRTMT	1	1.61179950	1.61179950	1.68	0.1987
Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRTMT	1	1.61179950	1.61179950	1.68	0.1987

SAS COMMANDS FOR BOX 11.10 (B): **Cat flea analysis with full model for FLEAS**

```

Commands   proc glm data=gandh.Chapter11;
            class TRTMT CARPET;
            model FLEAS = TRTMT HAIRL NCATS CARPET;
            run;

```

Menu route Statistics > Anova > Linear Models...

 FLEAS → Dependent

 TRTMT CARPET → Class

 HAIRL NCATS → Quantitative

SAS OUTPUT FOR BOX 11.10 (B) **Cat flea analysis with full model for FLEAS**

The GLM Procedure

Dependent Variable: FLEAS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	240319.6726	60079.9182	8.87	<.0001
Error	84	568947.2263	6773.1813		
Corrected Total	88	809266.8989			

	R-Square	Coeff Var	Root MSE	FLEAS Mean
	0.296960	92.50620	82.29934	88.96629

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRTMT	1	1412.1361	1412.1361	0.21	0.6491
HAIRL	1	17462.9857	17462.9857	2.58	0.1121
NCATS	1	188946.7985	188946.7985	27.90	<.0001
CARPET	1	32497.7524	32497.7524	4.80	0.0313

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRTMT	1	18437.9935	18437.9935	2.72	0.1027
HAIRL	1	3.7787	3.7787	0.00	0.9812
NCATS	1	188994.2135	188994.2135	27.90	<.0001
CARPET	1	32497.7524	32497.7524	4.80	0.0313

SAS COMMANDS FOR BOX 11.10 (C): **Cat flea analysis with full model for LOGFLEAS**

```

Commands  proc glm data=gandh.Chapter11;
           class TRTMT CARPET;
           model LOGFLEAS = TRTMT HAIRL NCATS CARPET;
           run;

```

Menu route Statistics > Anova > Linear Models...

 LOGFLEAS → Dependent

 TRTMT CARPET → Class

 HAIRL NCATS → Quantitative

SAS OUTPUT FOR BOX 11.10 (C) **Cat flea analysis with full model for LOGFLEAS**

The GLM Procedure

Dependent Variable: LOGFLEAS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	31.60884965	7.90221241	12.39	<.0001
Error	84	53.59441140	0.63802871		
Corrected Total	88	85.20326106			

	R-Square	Coeff Var	Root MSE	LOGFLEAS Mean
	0.370982	19.85294	0.798767	4.023419

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRTMT	1	1.61179950	1.61179950	2.53	0.1157
HAIRL	1	1.15225204	1.15225204	1.81	0.1826
NCATS	1	22.58432755	22.58432755	35.40	<.0001
CARPET	1	6.26047056	6.26047056	9.81	0.0024

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRTMT	1	5.75934967	5.75934967	9.03	0.0035
HAIRL	1	0.16422744	0.16422744	0.26	0.6132
NCATS	1	22.59152317	22.59152317	35.41	<.0001
CARPET	1	6.26047056	6.26047056	9.81	0.0024

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

Multiplicity of p-values

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