Exercise 14: Cross-validated R^2

The dataset `rock` taken from[11] contains measurements on four cross-sections of each of 12 oil-bearing rocks. The data-set is given in the end of the exercise. It is also available from the course home page.

The aim is to predict permeability (`perm`) (a property of fluid flow) from the other three measurements.

a) Calculate different summarising statistics of permeability. Make a histogram of the permeability. Comment on your results. Try different transformations for making the measurements more "equally" distributed. What is your prefered transformation? Use these transformed measurements in the following.

b) Make an analysis of the covariates, that is the three other measurements. In particular, calculate the correlations between the covariates. Will the correlations obtained have any effect on a regression analysis?

c) We now want to use the covariates to predict permeability. Plot permeability against the other measurements. Use the transformation for permeability you prefered from a). Perform a regression analysis with the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where x_1, x_2, x_3 are the three covariates `area`, `peri` and `shape` and y is either permeability or the tranformed permeability (where the transformation is the one you chose in a).

d) Order the covariates according to their T-values. Is this a proper way of ordering the covariates in this case? Use the ordering you got to perform cross-validation with 1, 2 and 3 covariates for calculating the cross-validated R-squared measure. What would your choise of model be?

(If you have problems performing cross-validation in your statistical package, use the adjusted R^2 instead.)

e) We will now extend our model to include both second order and interaction effects. Consider the full model

$$\begin{aligned} y = \beta_0 &+ \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &+ \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 \\ &+ \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_2 x_3 + \varepsilon. \end{aligned}$$

Use cross-validation for choosing an approperiate model in this case.

[11] Venables, W. N and Ripley, B. D. (1994). Modern Applied Statistics with S-Plus. Springer-Verlag.

f) There are many weaknesses in using the T-values from an ordinary regression analysis for ordering the covariates. By peforming a step-wise forward selection procedure, the following ordering was obtained: 2, 1, 7, 4, 5, 6, 8, 9, 3.

Use the cross-validation procedure for decision on a final model.

g) We will now analyse the model chosen in f). Use residual plots for investigating the usual asumptions made. Go through the different procedures from Lecture 5. Comment on your results.

rock	area	peri	shape	perm
1	4990	2791.8980	0.09032963	6.3
2	7002	3892.5984	0.14862240	6.3
3	7558	3930.6583	0.18331184	6.3
4	7352	3869.3190	0.11706328	6.3
5	7943	3948.5443	0.12241680	17.1
6	7979	4010.1545	0.16704479	17.1
7	9333	4345.7487	0.18965110	17.1
8	8209	4344.7459	0.16412710	17.1
9	8393	3682.0425	0.20365393	119.0
10	6425	3098.6518	0.16239442	119.0
11	9364	4480.0515	0.15094360	119.0
12	8624	3986.2422	0.14814132	119.0
13	10651	4036.5441	0.22859469	82.4
14	8868	3518.0357	0.23162315	82.4
15	9417	3999.3683	0.17256742	82.4
16	8874	3629.0733	0.15348108	82.4
17	10962	4608.6600	0.20431417	58.6
18	10743	4787.6205	0.26272664	58.6
19	11878	4864.2237	0.20007106	58.6
20	9867	4479.4077	0.14480992	58.6
21	7838	3428.7447	0.11385190	142.0
22	11876	4353.1388	0.29102946	142.0
23	12212	4697.6499	0.24007729	142.0
24	8233	3518.4405	0.16186492	142.0
25	6360	1977.3856	0.28088685	740.0
26	4193	1379.3490	0.17945461	740.0
27	7416	1916.2404	0.19180202	740.0
28	5246	1585.4187	0.13308318	740.0
29	6509	1851.2141	0.22521446	890.0
30	4895	1239.6551	0.34127298	890.0

rock	area	peri	shape	perm
31	6775	1728.1378	0.31164615	890.0
32	7894	1461.0583	0.27601553	890.0
33	5980	1426.7574	0.19765328	950.0
34	5318	990.3882	0.32663547	950.0
35	7392	1350.7630	0.15419249	950.0
36	7894	1461.0583	0.27601553	950.0
37	3469	1376.7014	0.17696851	100.0
38	1468	476.3220	0.43871230	100.0
39	3524	1189.4594	0.16358625	100.0
40	5267	1644.9558	0.25383179	100.0
41	5048	941.5429	0.32864058	1300.0
42	1016	308.6420	0.23008104	1300.0
43	5605	1145.6881	0.46412508	1300.0
44	8793	2280.4890	0.42047671	1300.0
45	3475	1174.1141	0.20074356	580.0
46	1651	597.8081	0.26265106	580.0
47	5514	1455.8751	0.18245258	580.0
48	9718	1485.5799	0.20044654	580.0