Department of Statistics

Course STATS 330

Model Answer for Assignment 3, 2005

Note: Unlike assignments 1 and 2, your answer for this assignment was expected to be in the form of a report. A sample report follows. Note that there are many possible models that could be fitted. You will get a good mark if you produce a good report, describing how you were led to choose a good model. I make no claims that my model is "best".

Predictions of runoff for six watersheds

Report by Alan Lee, XYZ Consultants

Executive Summary

In this report, we present predictions of water runoff for six different watersheds under a variety of climatic conditions. The predictions are based on a statistical model which has been developed from past runoff data.

Introduction

Water runoff in watersheds during severe storms is determined by a variety of factors, some depending on the nature of the watershed, some depending on the stare of the ground before the storm, and some on the intensity of the storm. In this report we use data on past storms to develop a statistical model that enables us to predict run-off peak flow under a variety of climatic conditions for the six watersheds of interest.

Data

We were given data on 30 past storms. The data set contained the following variables:

`watershed`:	The watershed being recorded, one of 1, 2, 3, 4, 5, 6.
`imperv:`	The area in the watershed impervious to water (sq km). This changes with time, the figure given is the value at the time of the storm.
`absorb:`	Surface absorbency index (0=no absorbency, 100=total absorbency) This depends on the amount of water in the ground prior to the storm.
`storage:`	Estimated soil storage capacity (mm of water) This also depends on the amount of water in the ground prior to the storm.
`rainfall:`	Rainfall in mm during the storm.

intensity: Time period (hours) for which the rainfall exceeded 5mm/hour.
peak.flow Peak flow of water (cubic meters/second).

We were also given data on each watershed. However, since the predictions were specific to the six watersheds under discussion, we initially ignored this information and instead fitted a "watershed effect". In case at a later date, the model might be applied to other watersheds, we also provide another prediction formula using the characteristics of the watersheds as predictors.

Analysis

As a first step in the analysis, we conducted a graphical exploration of the data using a series of pairs plots. In addition, a model was fitted to the full data set in order to identify any outliers that might adversely affect the model selection process. The pairs plots were inconclusive but the full-model fit identified points 28 and 30 as possible outliers. These points could have been removed but as the data set was small and the outliers not too severe, we left them in for the time being. Of more concern is the curvature in the residual plot. After some experimentation it was found that a log transformation improved matters.

We used the "all possible regressions" to choose a subset. Since the aim is ultimately prediction, we are particularly interested in the Cp criterion (which is based on prediction error), its close relative AIC, and CV. Note that using stepwise regression can't be any better than APR under these circumstances. We are also interested in getting a cheap predictor with a small number of variables, so the BIC criterion, which tends to select simpler models, will also be of interest.

From the APR output (see the technical appendix) the best model seems to be the one with all 5 dummy variables corresponding to the 6 watersheds, plus rainfall and intensity. A final fit of this model indicates point 28 is still extreme, so that this was deleted and the model refitted. The resulting R2 is 98%, an extremely good fit.

```
Call:
lm(formula = log(peak.flow) ~ watershed + rainfall + intensity,
    data = all.df, subset = (1:30)[-28])

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.851919   0.138863  27.739  < 2e-16 ***
watershed2   1.306713   0.134934   9.684 3.39e-09 ***
watershed3   3.027719   0.135711  22.310 4.17e-16 ***
watershed4   3.976886   0.170055  23.386  < 2e-16 ***
watershed5   3.701926   0.165347  22.389 3.88e-16 ***
watershed6   3.975177   0.145507  27.320  < 2e-16 ***
rainfall     0.021177   0.003071   6.896 8.16e-07 ***
intensity   -0.444517   0.062893  -7.068 5.65e-07 ***
---
```

```
Residual standard error: 0.2336 on 21 degrees of freedom
Multiple R-Squared: 0.9839,     Adjusted R-squared: 0.9786
F-statistic: 183.7 on 7 and 21 DF,  p-value: < 2.2e-16
```

Predictions

Our predictions are based on the model

```
log(peak flow) = 3.851919 + watershed effect + 0.021177 x rainfall
                -0.444517  x intensity
```

where the watershed effect is

1.306713 for watershed 2,
3.027719 for watershed 3,
3.976886 for watershed 4
3.701926 for watershed 5,
3.975177 for watershed 6.

The predictions for each watershed for rainfalls of 50, 75, and 100 mm, and intensities of 2 and 5 hours are given in Table 1, shown on the next page.

Conclusions

A model was successfully fitted to the 30 storms, and a prediction calculated for a variety of rainfall and intensity combinations. The model fitted involved a different prediction equation form each watershed, involving the variables rainfall and intensity only. The R^2 was very high (98%). The predictions may be found in Table 1.

Table 1. Predictions for 6 watersheds, for different rainfalls and intensities.

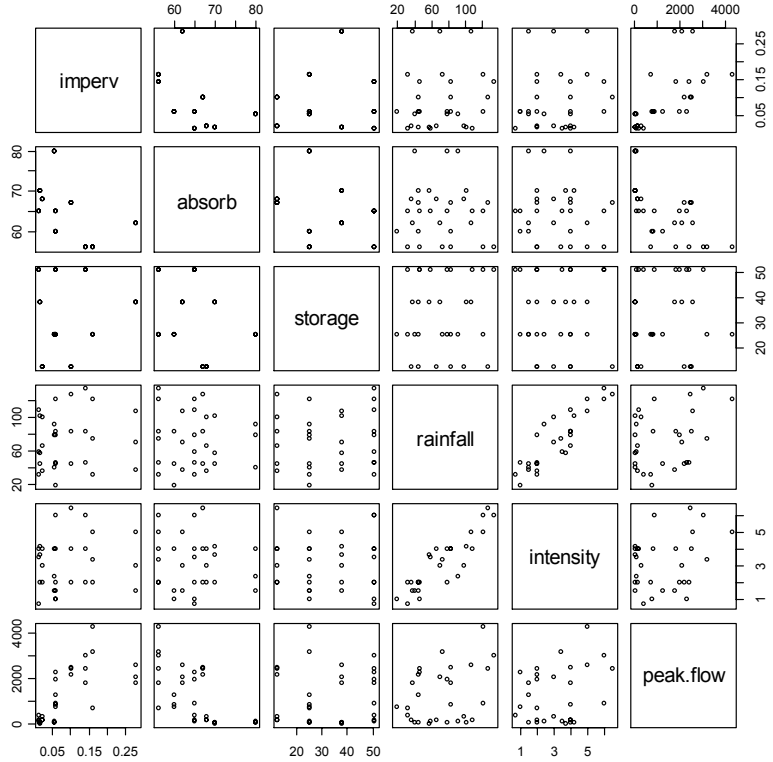
	watershed	rainfall	intensity	fit	lwr	upr
1	1	50	2	55.799	32.898	94.642
2	2	50	2	206.122	121.613	349.354
3	3	50	2	1152.251	680.317	1951.563
4	4	50	2	2976.906	1670.576	5304.737
5	5	50	2	2261.265	1284.264	3981.516
6	6	50	2	2971.825	1717.468	5142.305
7	1	75	2	94.745	54.718	164.051
8	2	75	2	349.989	202.074	606.174
9	3	75	2	1956.489	1138.070	3363.457
10	4	75	2	5054.700	2755.476	9272.443
11	5	75	2	3839.563	2134.965	6905.147
12	6	75	2	5046.074	2869.970	8872.168
13	1	100	2	160.874	87.232	296.685
14	2	100	2	594.271	321.874	1097.194
15	3	100	2	3322.063	1822.468	6055.581
16	4	100	2	8582.737	4376.975	16829.744
17	5	100	2	6519.468	3410.440	12462.752
18	6	100	2	8568.089	4598.756	15963.479
19	1	50	5	14.705	7.706	28.063
20	2	50	5	54.321	28.482	103.603
21	3	50	5	303.664	156.585	588.896
22	4	50	5	784.534	410.958	1497.706
23	5	50	5	595.934	304.875	1164.863
24	6	50	5	783.195	408.755	1500.639
25	1	75	5	24.969	14.061	44.338
26	2	75	5	92.236	51.929	163.829
27	3	75	5	515.613	286.709	927.272
28	4	75	5	1332.116	742.089	2391.269
29	5	75	5	1011.879	553.744	1849.048
30	6	75	5	1329.843	748.415	2362.970
31	1	100	5	42.397	24.680	72.831
32	2	100	5	156.614	91.053	269.382
33	3	100	5	875.497	505.725	1515.637
34	4	100	5	2261.895	1287.577	3973.486
35	5	100	5	1718.141	968.465	3048.133
36	6	100	5	2258.035	1318.973	3865.676

Technical Appendix

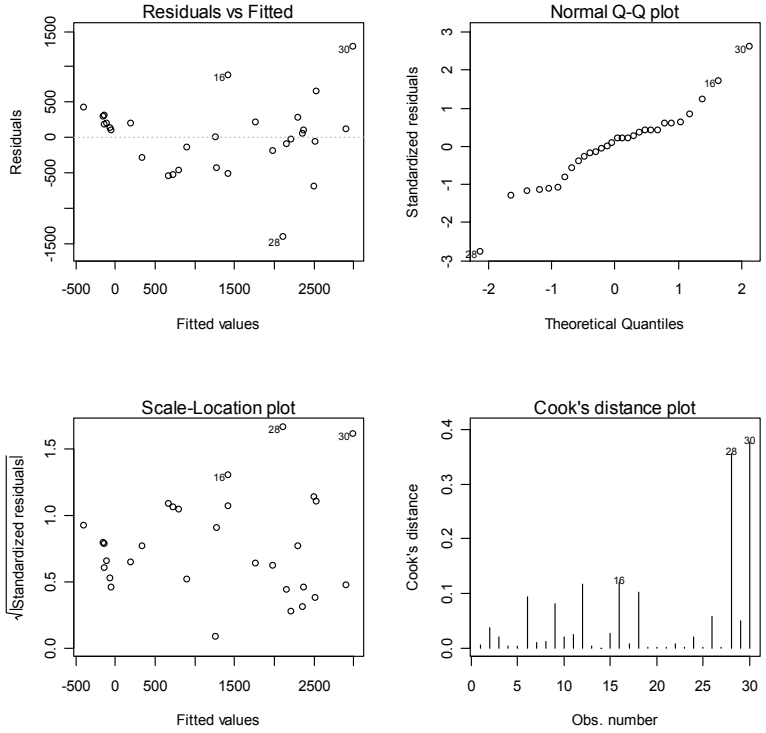
Data used (note that variable watershed is a factor)

	watershed	imperv	absorb	storage	rainfall	intensity	peak.flow
1	1	0.016	70	38.1	44.450	2.0	46
2	1	0.016	70	38.1	57.150	3.7	28
3	1	0.016	70	38.1	101.600	4.2	54
4	1	0.054	80	25.4	40.640	1.5	70
5	1	0.054	80	25.4	78.740	4.0	47
6	1	0.054	80	25.4	91.440	2.4	112
7	2	0.013	65	50.8	31.750	0.7	398
8	2	0.013	65	50.8	58.420	3.5	98
9	2	0.013	65	50.8	107.950	4.0	191
10	2	0.021	68	12.7	36.830	2.0	171
11	2	0.021	68	12.7	66.040	4.0	150
12	2	0.021	68	12.7	99.060	3.0	331
13	3	0.060	60	25.4	19.050	1.0	772
14	3	0.060	60	25.4	44.450	1.5	1268
15	3	0.060	60	25.4	82.550	4.0	849
16	3	0.060	65	50.8	45.720	1.0	2294
17	3	0.060	65	50.8	78.740	2.0	1984
18	3	0.060	65	50.8	120.650	6.0	900
19	4	0.101	67	12.7	44.450	2.0	2181
20	4	0.101	67	12.7	82.550	4.0	2484
21	4	0.101	67	12.7	127.000	6.5	2450
22	5	0.282	62	38.1	38.100	1.5	1794
23	5	0.282	62	38.1	69.850	3.0	2067
24	5	0.282	62	38.1	106.680	5.0	2586
25	6	0.142	56	50.8	45.720	2.0	2410
26	6	0.142	56	50.8	82.550	4.0	1808
27	6	0.142	56	50.8	133.350	6.0	3024
28	6	0.163	56	25.4	31.750	2.0	710
29	6	0.163	56	25.4	73.660	3.4	3181
30	6	0.163	56	25.4	120.904	5.0	4279

Pairs plots



Diagnostic plots for full model



Curvature in residual plot suggests the need to transform. Box-Cox plot indicates a log transformation of the response. This produces a good fit with reasonable residual plots.

An APR indicates keeping watershed, rainfall and intensity:

	rssp	sigma2	adjRsq	Cp	AIC	BIC	CV	watershed2	watershed3	watershed4	watershed5
1	34.064	1.217	0.506	326.916	356.916	359.719	3.958	0	0	0	0
2	23.760	0.880	0.643	222.169	252.169	256.373	2.919	0	0	1	0
3	16.765	0.645	0.738	151.697	181.697	187.302	1.965	0	0	1	0
4	10.309	0.412	0.832	86.811	116.811	123.817	1.376	0	0	0	1
5	5.356	0.223	0.909	37.490	67.490	75.898	0.778	1	1	1	0
6	4.793	0.208	0.915	33.657	63.657	73.465	0.773	1	1	1	1
7	**1.953**	**0.089**	**0.964**	**6.229**	**36.229**	**47.439**	**0.412**	**1**	**1**	**1**	**1**
8	1.852	0.088	0.964	7.183	37.183	49.794	0.408	1	1	1	1
9	1.840	0.092	0.963	9.061	39.061	53.072	0.462	1	1	1	1
10	1.834	0.097	0.961	11.000	41.000	56.413	0.486	1	1	1	1

	watershed6	imperv	absorb	storage	rainfall	intensity
1	0	0	1	0	0	0
2	0	0	1	0	0	0
3	0	1	1	0	0	0
4	1	1	1	0	0	0
5	1	1	0	0	0	0
6	1	0	1	0	0	0
7	**1**	**0**	**0**	**0**	**1**	**1**
8	1	0	1	0	1	1
9	1	1	1	0	1	1
10	1	1	1	1	1	1

Summary for final model, after deleting point 28.

```
Call:
lm(formula = log(peak.flow) ~ watershed + rainfall + intensity,
    data = all.df, subset = (1:30)[-28])

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.851919   0.138863  27.739  < 2e-16 ***
watershed2   1.306713   0.134934   9.684 3.39e-09 ***
watershed3   3.027719   0.135711  22.310 4.17e-16 ***
watershed4   3.976886   0.170055  23.386  < 2e-16 ***
watershed5   3.701926   0.165347  22.389 3.88e-16 ***
watershed6   3.975177   0.145507  27.320  < 2e-16 ***
rainfall     0.021177   0.003071   6.896 8.16e-07 ***
intensity   -0.444517   0.062893  -7.068 5.65e-07 ***
---
Residual standard error: 0.2336 on 21 degrees of freedom
Multiple R-Squared: 0.9839,     Adjusted R-squared: 0.9786
F-statistic: 183.7 on 7 and 21 DF,  p-value: < 2.2e-16
```

Residual plots