Generalized linear models

Normal data – Gamma data

1. Larsen and Marx (1986, An Introduction to Mathematical Statistics and its Applications, Prentice Hall) write: in folk belief, the full moon represents something threatening, a kind of evil force that has the power of controlling our behavior. Over the centuries, many outstanding authors and philosophers have shared this belief. A typical example is Othello who, after the murder of Dysdemona lament that "everything is a fault of the moon. It comes closer to the earth than the earth desires. And this makes men crazy". The data in the file fullMoon.txt gives the rate of admission to the intensive care units of a psychiatric clinic in Virginia USA, before, during and after the 12 full moons from August 1971 to July 1972. Consider each month as a block and the rate of admission as a new variable that follows the normal distribution.

a. Fit a model that considers only the effect of full moon ignoring the effect of months. What are your conclusions?

b. Add the effects of months in the model. Are there any differences in the estimation of moon's effect between models of a. and b.?

c. What is the estimate of the data variance for each of the two models? Which of them is better?

d. Use a type of residuals for checking whether the model fits the data well.

2. A fan of squash noticed that ball's behavior seems to be different at the beginning and at the end of the game and that this behavior also depends on ball's age. So, he decided to conduct a factorial experiment, with the purpose of examining ball's behavior under different experimental conditions.

Factors of the experiment:

Ball's type: a ball with a yellow point on it (very slow) and another with a double x on it (too slow) were used.

Temperature: During the game, ball tends to become warmer. It was used two different temperatures for the ball: Room temperature and game temperature, that was achieved by putting the ball on boiling water for 45 seconds.

Age: It was used old and new balls.

Measurement process: Each ball was launched by a machine (special for the training of those involved with squash). In this way, the dispersion on the launch speed was minimized. Each ball hit the wall bounced back and hit the ground. The distance of the point where the ball hit the ground from the wall was recorded each time.

The results are presented on the table

Ball: Yellow point/Double x +/-

Temperature: Room/Game +/-

Age: New/Old +/-

Series	Ball	Temp.	Age	Dist.(cm)

7	+	+	+	540
13	+	+	+	567
11	+	+	−	553
8	+	+	−	465
14	+	−	−	637
12	+	−	−	562
4	+	−	+	613
2	+	−	+	685
16	−	+	−	467
6	−	+	−	412
1	−	+	+	497
5	−	+	+	525
3	−	−	−	647
10	−	−	−	619
15	−	−	+	719
9	−	−	+	673

a. Start from the largest possible model that considers all possible interactions between the three factors. In each step, remove the (interaction) variable which is the most statistically insignificant until finding an optimal model which includes statistically significant main or interaction effects. Describe what can be concluded for the data from this model.

b. Apply multiple comparisons for explaining the statistically significant interactions of the model.

3. In autumn, the small maple nuts, that are called samara, spin as they fall. A scientist from the local forest institute studied the association between fall velocity and their size (an index that based on the size and the weight of the nut). Data can be found for three trees in the file samara.txt. The purpose of this analysis is to examine whether the nuts fall with the same velocity from the three trees. It is desirable to take into consideration the size of the nuts, to compare nuts of the same size.

4. The following data (data from DASL library) comes from a survey that aimed to measure the mass and other physical characteristics of 22 men aged 16-30. The subjects were randomly selected from healthy volunteers. Except for mass, all the other measurements are in cm. The type of measurements and the data are given below.

Variable	Description
Mass	Weight in kg
Fore	Maximum circumference of forearm
Bicep	Maximum circumference of bicep
Chest	Distance around chest directly under the armpits
Neck	Distance around neck, approximately halfway up
Waist	Distance around waist, approximately trouser line
Thigh	Circumference of thigh, measured halfway between the knee and the top of the leg
Calf	Maximum circumference of calf
Height	Height from top to toe
Shoulders	Distance around shoulders, measured around the peak of the shoulder blades

Mass	Fore	Bicep	Chest	Neck	Shoulder	Waist	Height	Calf	Thigh	Head
77	28.5	33.5	100	38.5	114	85	178	37.5	53	58
85.5	29.5	36.5	107	39	119	90.5	187	40	52	59
63	25	31	94	36.5	102	80.5	175	33	49	57
80.5	28.5	34	104	39	114	91.5	183	38	50	60
79.5	28.5	36.5	107	39	114	92	174	40	53	59
94	30.5	38	112	39	121	101	180	39.5	57.5	59
66	26.5	29	93	35	105	76	177.5	38.5	50	58.5
69	27	31	95	37	108	84	182.5	36	49	60
65	26.5	29	93	35	112	74	178.5	34	47	55.5
58	26.5	31	96	35	103	76	168.5	35	46	58
69.5	28.5	37	109.5	39	118	80	170	38	50	58.5
73	27.5	33	102	38.5	113	86	180	36	49	59
74	29.5	36	101	38.5	115.5	82	186.5	38	49	60
68	25	30	98.5	37	108	82	188	37	49.5	57
80	29.5	36	103	40	117	95.5	173	37	52.5	58
66	26.5	32.5	89	35	104.5	81	171	38	48	56.5
54.5	24	30	92.5	35.5	102	76	169	32	42	57
64	25.5	28.5	87.5	35	109	84	181	35.5	42	58
84	30	34.5	99	40.5	119	88	188	39	50.5	56
73	28	34.5	97	37	104	82	173	38	49	58
89	29	35.5	106	39	118	96	179	39.5	51	58.5
94	31	33.5	106	39	120	99.5	184	42	55	57

a. Find the correlations between mass and each of the other explanatory variables.

b. Check in the bibliography what the partial plots are. Create partial plots to deduce which of the explanatory variables can be used for a linear model that associates mass with them.

c. Starting from the largest model that considers only linear effects of explanatory variables, find the best model using BIC.

d. Check goodness-of-fit of the final model using appropriate residual plots.

e. What are your final conclusions?

5. It is known that the concertation of cholesterol in the blood increases with the age, but it is not clear whether the levels of cholesterol are associated with the body weight. The data in file cholesterol.dat shows the levels of cholesterol (millimoles per liter), the age (years) and the BMI index (weight in kg divided by height in meters squared) for a number of subjects.

a. Make a graph to empirically deduce the relation between cholesterol-age and between cholesterol-BMI.

b. Fit an appropriate GLM for checking whether the cholesterol is related to BMI when age is already included in the model.

c. Interpret the estimated coefficient of the effect of BMI irrespective of its statistical significance.

d. Examine visually the adequacy of the model, using appropriate residual plots.

6. Use the data file data for GLM practice.pdf. In dataset 7 there is data from 21 consecutive days of operation of an ammonia to nitric acid oxidation process. The variables x1, x2 and x3 are the air flow, the temperature of the cold water and the concentration of nitric acid. The response (y) is (10 times) the percentage of lost ammonia because it cannot be absorbed as nitric acid.

a. Start with a GLM which assumes the effect of the three explanatory variables. See the literature on optimal model selection and apply a method (of your choice) to find the optimal model that adequately explains the data.

b. Check visually the assumptions of the model using residual plots.

c. What are your conclusions?

7. An engineer study the use of watermill in the production of electricity. He has measurements of the direct current (y) in Volt and of the wind speed (x). The data is:

Observation	Wind speed (mph)	DC exit (volt)	Observation	Wind speed (mph)	DC exit (volt)
1	5.00	1.582	13	4.60	1.562
2	6.00	1.822	14	5.80	1.737
3	3.40	1.057	15	7.40	2.088
4	2.70	0.500	16	3.60	1.137
5	10.00	2.236	17	7.85	2.179
6	9.70	2.386	18	8.80	2.112
7	9.55	2.294	19	7.00	1.800
8	3.05	0.558	20	5.45	1.501
9	8.15	2.166	21	9.10	2.303
10	6.20	1.866	22	10.20	2.310
11	2.90	0.653	23	4.10	1.194
12	6.35	1.930	24	3.95	1.144
			25	2.45	0.123

a. Fit a GLM with the linear effect of the wind speed for relating the DC exit with the wind speed. Use the identical link function.
b. Add the quadratic effect in the model. Does it provide a better fit? Use residuals or an F test to answer.
c. Taking into consideration that theory predicts the existence of one asymptotically maximum value for DC exit as wind speed increases, try to propose a transformation on dependent variable (for example 1/y) or on independent variable (1/x), in order to ensure an reasonable final proposed model.
d. It would be interesting to see if instead of making a transformation on the dependent variable, you changed the link function. Are these two models equivalent?

8. The data in the file nerve.dat shows the time between two successive impulses in a nerve fiber. Approximately, it could be assumed that the data follows a Gamma distribution. Fit an appropriate model and comment on the estimated model coefficients (including the dispersion parameter).

9. The data in the file carinsurance.dat shows the mean compensation for car damages in Great Britain in 1975. The mean values are given in pound sterling. The variables are: OwnerAge (age of the owner of the insurance, 8 levels), Model (car's model type according to engine capacity in liters, 4 levels), Car Age (car's age, 4 levels), Nclaims (number of claims for compensation), AveCost (mean cost for each compensation). Use only the main effects of the explanatory variables for modelling the mean cost for each compensation. Take into consideration the following: The mean cost comes from different number of compensations, there are extreme values and possible normal distribution is not adequate for modelling the data. Check the goodness-of-fit of the model and summarize your conclusions. Help can be found on Mc Cullagh and Nelder (1989) Generalized Linear Models, Chapman and Hall, Chapter 8.

10. The data in file cherry_trees.dat shows measurements from the diameter, the height and the volume of cherry trees with the purpose of helping us predicting the volume of the tree using the diameter of its body and its height. Don't use diam2 variable. Find the appropriate way to model the data, which are quite right skewed. What are your conclusions? Predict the volume of the trees when the diameter is 15.5 and the heights are 60,70,75,80,85,90.