Financial Time Series I and Methods of Statistical Prediction

Homework 4: Contingency table, regression, ANOVA

Due Date: December 12th, 2002

1. For the 23 space shuttle flights that occurred before the Challenger mission Disaster in 1986, the following table shows the temperature (^oF) at the time of the flight and whether at least one primary O-ring suffered thermal distress.

Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD
1	66	0	9	57	1	17	70	0
2	70	1	10	63	1	18	81	0
3	69	0	11	70	1	19	76	0
4	68	0	12	78	0	20	79	0
5	67	0	13	67	0	21	75	1
6	72	0	14	53	1	22	76	0
7	73	0	15	67	0	23	58	1
8	70	0	16	75	0			

Here Ft= flight no., Temp= temperature, TD = thermal distress ($1 = yes$, $0 = no$).

(a) Use logistic regression to model the effect of temperature on the probability of thermal distress. Interpret the model fit.

(b) Calculate the predicted probability of thermal distress at 31^o, the temperature at the time of the Challenger flight. At what temperature does the predicted probability equal 0.5?

(c) Interpret the effect of temperature on the odds of thermal distress. Test the hypothesis that temperature has no effect, using the likelihood ratio test.

2. The following data is reported at 1991 General Society Survey at USA. The variables are gender and party identification. Subjects indicated whether they identified more strongly with the Democratic or Republic party or as Independents.

	Party Identification			
Gender	Democrat	Independent	Republic	Total
Female	279	73	225	577
Male	164	47	191	403
Total	444	120	416	980

Use Pearson Chi-Square test to test the null hypothesis of statistical independence of gender and party identification and report your conclusion.

3. Let $Y_1, \ldots, Y_n$ be independent random variables with distribution

$$Y_i \sim N(\mu x_i, \sigma^2), \quad i = 1, \ldots, n,$$

where $x_1, \ldots, x_n$ are constants. Find the maximum likelihood estimators for μ and σ^2, and find the distribution of $\hat{\mu}$.

4. The following table gives the survival times (y) in hours after getting a poison for 12 sheep with different weights (x) in pounds. Analyze the data using a linear

regression model. Include a scatter plot of y versus x, a normal probability plot of residuals and provide the estimates for the intercept, slope, and the variance of Y with their standard errors. Make a report based on your analysis, respecting the following rules:

- Presentation of the problem and the data.
- Statistical analysis.
- Conclusions.

Refer to the last two pages for reference on a sample report.

x	46	55	61	75	64	75	71	59	66	67	60	63
y	44	27	24	24	36	36	44	44	36	29	36	36

5. A nationwide real estate brokerage house wants to study the relationship between rent per square foot and the size of the property. The data collected are summarized in the following table.

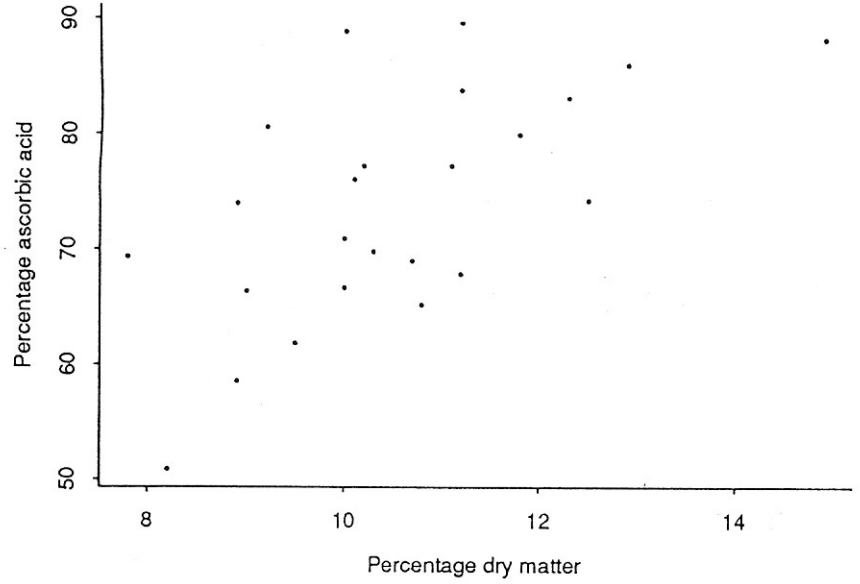
	Size of the Property (in square feet)		
Location	Less than $1,000$	$1,000$ to $2,000$	$2,000$ or more
Bad	3	2	3
	4	5	6
So-so	5	5	7
	5	6	7
Good	5	5	7
	4	6	6

(a) Using these data, can we reject the null hypothesis that the average rent per square foot are equal?

(b) Redo the analysis without controlling the location factor, will it change the conclusion derived in (a)?

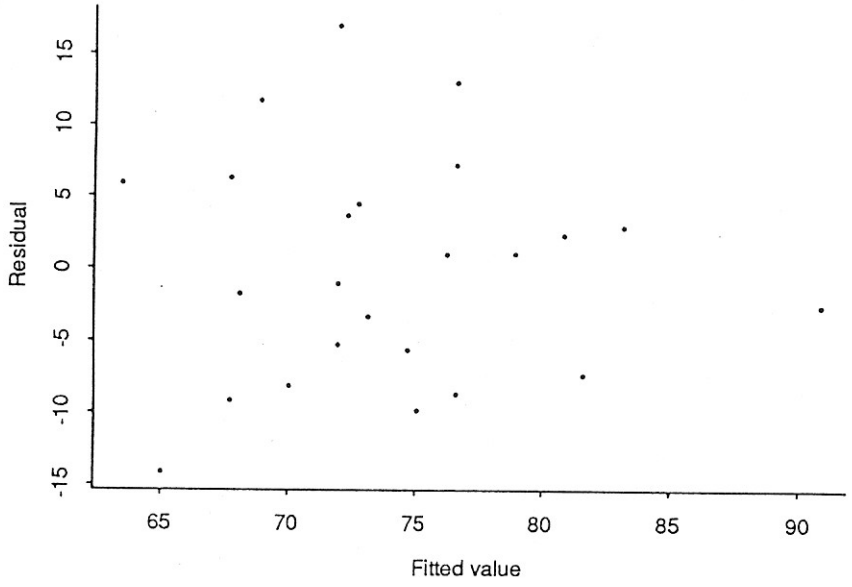
Example 1.1

Figure 1.2 shows the relationship between the percentage dry matter of fresh spinach (x) and the percentage preserved ascorbic acid after drying at 90°C (y). The data in Table 1.1, from Hald (1952, p. 548), are from a study concerning the preservation of ascorbic acid in vegetables during drying and storing. The choice of percentage ascorbic acid as response variable reflects the desire to predict this variable as a function of percentage dry matter, and is based on the (reasonable) assumption that x may have a causal effect on y in the present context. The scatterplot supports the linearity of the relationship, and the constancy of the variance of Y for x fixed. The independence of the observations may be assumed if the 24 experiments were performed separately, in some sense, although we have no detailed information about this point. We analyze this data set in detail in Section 1.7.



(i) Presentation of the problem.

As mentioned in Example 1.1, the data in Table 1.1 represent the rela tionship between the percentage dry matter of fresh spinach (x) and th percentage preserved ascorbic acid after drying at 90°C (y). The data ar from an investigation concerning the preservation of ascorbic acid in veg etables during drying and storing, so consequently percentage preserve ascorbic acid after drying is chosen as response variable (y). The question relevant to this investigation are if the relationship between x and y can b said to be linear in the x-interval under study, which ranges from 6 to 1 percent dry matter, and what the magnitude of the deviation from the lin ear relationship is. Furthermore, we may ask how precisely the parameter of the linear relationship have been estimated.



(ii) *Statistical analysis*

Figure 1.2 shows the scatterplot of y versus x, and as noted earlier, the plot does not suggest any substantial departures from the linear regression model for these data. The statistical model that we use is hence

$$Y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2), \quad i = 1, \ldots, 24,$$

where $Y_1, \ldots, Y_{24}$ are independent, the data y_i representing a realization of the random variable Y_i. A further check of the model is provided by the plot of residuals of the normal plot of residuals, shown in Figures 7.1 and 7.2. The first plot shows that the variance is constant, and the second shows a nice linear relationship, confirming the normality of the residuals. In any case, it is difficult to reject normality based on a sample of only 24 observations.

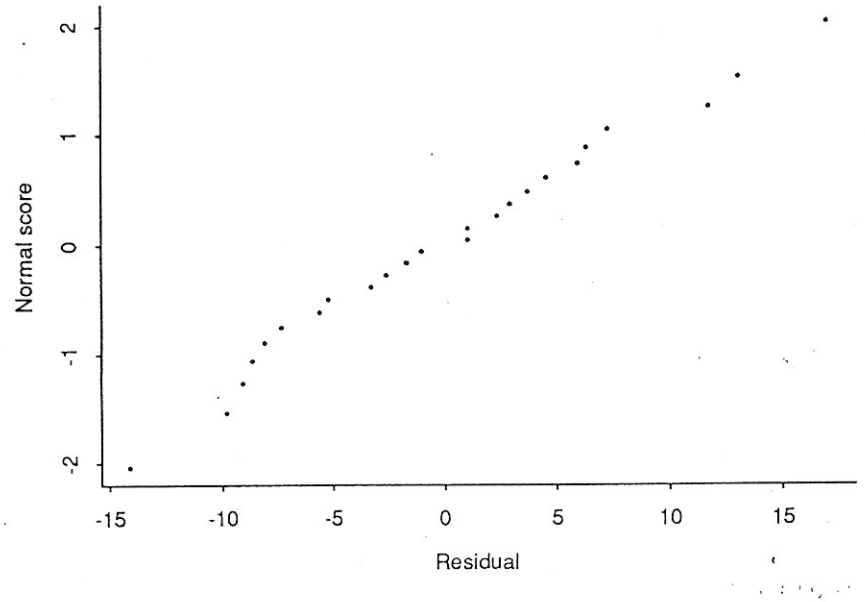


Figure 7.2 *Normal plot of residuals, spinach data.*

To complete the verification of the model, we note that the assumption of independence of the 24 observations requires that the 24 experiments were in some sense performed separately, both in space and time, although we have no specific information about this point here.

Table 7.1. Parameter estimates for linear regression model, spinach data

Parameter	Estimate	s.e.
β_1	33.48	11.10
β_2	3.85	1.04
	$\tilde{\sigma}^2 = 64.84$	d.f. = 22

The parameter estimates for the model and their standard errors are given in Table 7.1. Based on these values, the estimated linear relationship between $\mathrm{E}(Y)$ and x is given by

$$\mathrm{E}(Y) = 33.48 + 3.85x, \tag{7.1}$$

with a standard deviation estimated by $\tilde{\sigma} = 8.05$. A 95% confidence interval for β_2 is $[1.70, 6.00]$. The t-test for the hypothesis $\beta_2 = 0$ is

$$t(y) = \frac{3.85}{1.04} = 3.70,$$

with 22 degrees of freedom, which gives a p-value of less than 0.01. There is hence a strong indication that β_2 is not zero.

(iii) *Conclusions*

The statistical analysis shows that the data may reasonably be described by a linear regression model, the estimated relationship being given by (7.1). The estimates and their standard errors in Table 7.1 show that the parameters are not very precisely estimated, particularly so for the intercept β_1. The statistical test for the hypothesis that the slope is zero rejects the hypothesis. The percentage preserved ascorbic acid after drying hence depends on the percentage dry matter of fresh spinach, with a slope between about 1.7 and 6.0 (95% confidence interval). Hence, equation (7.1) may be useful for predicting y from x, but a vertical deviation ("prediction error") of about $1.96 \times 8.05 = 15.77$ should be expected. For example, for $x = 10$, a value of y between 56.21 and 87.75 is expected with probability 95%, with a median value of 71.98.

We have here used the normal distribution as a basis for the prediction interval. A more detailed discussion of prediction, which is given in Section 4.5, shows that the correct prediction intervals should be based on the t-distribution, although the above approach is approximately correct.