Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: tisdagen den 11 mars, 2014 kl 14.00-18.00
Examinator och jour: Serik Sagitov, tel. 772-5351, mob. 0736 907 613, rum H3026 i MV-huset.
Hjälpmedel: Chalmersgodkänd räknare, **egen** formelsamling (fyra A4 sidor).
CTH: för "3" fordras 12 poäng, för "4" - 18 poäng, för "5" - 24 poäng.
GU: för "G" fordras 12 poäng, för "VG" - 20 poäng.
Inclusive eventuella bonuspoäng.

Partial answers and solutions are also welcome. Good luck!

1. (5 points) The article "Effects of gamma radiation on juvenile and mature cuttings of quaking aspen" (Forest science, 1967) reports the following data on exposure time to radiation (x, in kr/16 hr) and dry weight of roots (y, in mg$\times 10^{-1}$):

x	0	2	4	6	8
y	110	123	119	86	62

The estimated quadratic regression function is $y = 111.9 + 8.1x - 1.8x^2$.

a. What is the underlying multiple regression model? Write down the corresponding design matrix.

b. Compute the predicted responses. Find an unbiased estimate s^2 of the noise variance σ^2.

c. Compute the coefficient of multiple determination.

2. (5 points) The accompanying data resulted from an experiment carried out to investigate whether yield from a certain chemical process depended either on the formulation of a particular input or on mixer speed.

		Speed			
		60	70	80	Means
Formulation	1	189.7	185.1	189.0	
	1	188.6	179.4	193.0	187.03
	1	190.1	177.3	191.1	
	2	165.1	161.7	163.3	
	2	165.9	159.8	166.6	164.66
	2	167.6	161.6	170.3	
	Means	177.83	170.82	178.88	175.84

A statistical computer package gave

$$SSForm = 2253.44, \quad SSSpeed = 230.81, \quad SSForm * Speed = 18.58, \quad SSE = 71.87.$$

a. Calculate estimates of the main effects.

b. Does there appear to be interaction between the factors? In which various ways interaction between such two factors could manifest itself? Illustrate with graphs.

c. Does yield appear to depend either on formulation or speed.

d. Why is it important to inspect the scatter plot of residuals?

3. (5 points) A study of the relationship between facility conditions at gasoline stations and aggressiveness in the pricing of gasoline is based on $n = 441$ stations.

	Pricing policy			
	Aggressive	Neutral	Nonaggressive	Total
Substandard condition	24	15	17	56
Standard condition	52	73	80	205
Modern condition	58	86	36	180
Total	134	174	133	441

a. Suggest a parametric model for the data and write down the corresponding likelihood function.

b. What is a relevant null hypothesis for the data?

c. Properly analyze the data and draw your conclusions.

4. (5 points) Mice were injected with a bacterial solution: some of the mice were also given penicillin. The results were

	Without penicillin	With penicillin
Survived	8	12
Died	48	62

a. Find a 95% confidence interval for the difference between two probabilities of survival.

b. Assume that both groups have the probability of survival p. How would you compute an exact credibility interval for the population proportion p, if you could use a computer? Compute an approximate 95% credibility interval using a normal approximation.

5. (5 points) In a controlled clinical trial which began in 1982 and ended in 1987, more than 22000 physicians participated. The participants were randomly assigned in two groups: Aspirin and Placebo. The aspirin group have been taking 325 mg aspirin every second day. At the end of trial, the number of participants who suffered from myocardial infarctions was assessed.

	MyoInf	No MyoInf	Total
Aspirin	104	10933	11037
Placebo	189	10845	11034

The popular measure in assessing the results in clinical trials is Risk Ratio

$$RR = R_A/R_P = \frac{104/11037}{189/11034} = 0.55.$$

a. How would you interpret the obtained value of the risk ratio? What ratio of conditional probabilities is estimated by RR?

b. Is the observed value of RR significantly different from 1?

6. (5 points) Given a sample $(X_1, \ldots, X_n)$ of independent and identically distributed observations, we are interested in testing H_0: $M = M_0$ against the two-sided alternative H_1: $M \neq M_0$ concerning the population median M. No parametric model is assumed. As a test statistic we take $Y = \sum_{i=1}^{n} 1_{\{X_i \leq M_0\}}$, the number of observations below the null hypothesis value.

a. Find the exact null distribution of Y. What are your assumptions?

b. Suppose $n = 25$. Suggest an approximate confidence interval formula for M.

Statistical tables :

95% percentiles of the F_{n_1,n_2} distribution

	$n_1 = 1$	$n_1 = 2$	$n_1 = 3$	$n_1 = 10$	$n_1 = 12$	$n_1 = 15$	$n_1 = 20$
$n_2 = 1$	161.4	199.5	215.7	241.9	243.9	245.9	248.0
$n_2 = 2$	18.51	19.00	19.16	19.40	19.41	19.43	19.45
$n_2 = 3$	10.13	9.55	9.28	8.79	8.74	8.70	8.66
$n_2 = 10$	4.96	4.10	3.71	2.98	2.91	2.85	2.77
$n_2 = 12$	4.75	3.89	3.49	2.67	2.60	2.53	2.46
$n_2 = 15$	4.54	3.68	3.29	2.54	2.48	2.40	2.33
$n_2 = 20$	4.35	3.49	3.10	2.35	2.28	2.20	2.12

NUMERICAL ANSWERS

1a. Multiple regression model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, where the random variables ϵ_i, $i = 1, \ldots, 5$ are independent and have the same normal distribution $\mathrm{N}(0, \sigma^2)$. The corresponding design matrix has the form

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 8 & 64 \end{pmatrix}$$

1b. Using the formula $\hat{y}_i = 111.9 + 8.1 x_i - 1.8 x_i^2$ we get

x_i	0	2	4	6	8
y_i	110	123	119	86	62
$\hat{y}_i$	111.9	120.9	115.5	95.7	61.5

and then $s^2 = \frac{\|\mathbf{y}-\hat{\mathbf{y}}\|^2}{n-p} = \frac{114.6}{2} = 57.3$.

1c. Coefficient of multiple determination

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{114.6}{2630} = 0.956.$$

2a. In terms of the two-way ANOVA model $Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk}$ (grand mean + main effects + interaction+noise), we estimate the main effects as

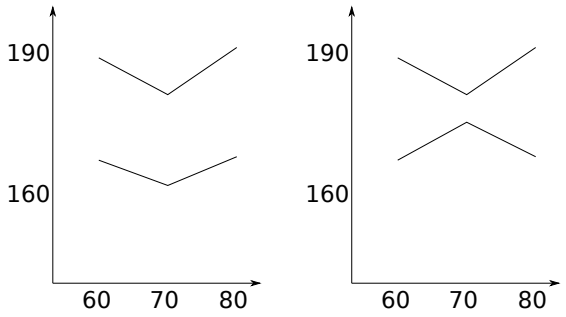
$$\hat{\alpha}_1 = 11.9, \hat{\alpha}_2 = -11.8, \qquad \hat{\beta}_1 = 1.99, \hat{\beta}_2 = -5.02, \hat{\beta}_3 = 3.04.$$

(Notice the effect of rounding errors.)

2b. Compute the cell means

	Speed		
	60	70	80
1	189.7	185.1	189.0
1	188.6	179.4	193.0
1	190.1	177.3	191.1
Cell means	189.5	180.6	191.0
2	165.1	161.7	163.3
2	165.9	159.8	166.6
2	167.6	161.6	170.3
Cell means	166.2	161.0	166.7

and draw two lines for the speed depending on two different formulations, see the left panel on the figure below. These two lines are almost parallel indicating to the absence of interaction between two main factors. This is confirmed by the ANOVA table below showing that the interaction is not significant.



One possible interaction effect could have the form on the right panel. In this case the formulation 2 interacts with the speed factor in such a way that the yield becomes largest at the speed 70.

2c. Anova-2 table

Source	df	SS	MS	F	Critical values	Significance
Formulation	1	2253.44	2253.44	376.2	$F_{1,12} = 4.75$	Highly significant
Speed	2	230.81	115.41	19.3	$F_{2,12} = 3.89$	Highly significant
Interaction	2	18.58	9.29	1.55	$F_{2,12} = 3.89$	Not significant
Error	12	71.87	5.99			
Total	17					

3a. This is a single sample of size $n = 441$. Each of n observations falls in of 9 groups. The multinomial distribution model

$$(n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}, n_{31}, n_{32}, n_{33}) \sim \text{Mn}(n, p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33})$$

gives the likelihood function

$$\begin{aligned} &L(p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33}) \\ &= P(n_{11} = 24, n_{12} = 15, n_{13} = 17, n_{21} = 52, n_{22} = 73, n_{23} = 80, n_{31} = 58, n_{32} = 86, n_{33} = 36) \\ &= \frac{441!}{24!15!17!52!73!80!58!86!36!} p_{11}^{24} \cdot p_{12}^{15} \cdot p_{13}^{17} \cdot p_{21}^{52} \cdot p_{22}^{73} \cdot p_{23}^{80} \cdot p_{31}^{58} \cdot p_{32}^{86} \cdot p_{33}^{36}. \end{aligned}$$

3b. The null hypotheis of independence $H_0 : p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$ meaning that there is no relationship between facility conditions at gasoline stations and aggressiveness in the pricing of gasoline.

3c. The chi-square test statistic $X^2 = 22.5$ should be compared with the critical values of χ^2_4-distribution. Even though the corresponding table is not given we may guess that the result must be significant as the square root of 22.5 is quite large. We reject the null hypothesis of independence and conclude that that there is a relationship between facility conditions at gasoline stations and aggressiveness in the pricing of gasoline.

	Pricing policy			
	Aggressive	Neutral	Nonaggressive	Total
Substandard condition	24 (17)	15 (22)	17 (17)	56
Standard condition	52 (62.3)	73 (80.9)	80 (61.8)	205
Modern condition	58 (54.7)	86 (71)	36 (54.3)	180
Total	134	174	133	441

It looks like the standard conditions are coupled with the least aggressive pricing strategy.

4a. Two independent dichotomous samples with $n = 56$, $\hat{p}_1 = \frac{8}{56} = 0.143$ and $m = 74$, $\hat{p}_2 = \frac{12}{74} = 0.162$. An asymptotic 95% confidence interval for the population difference is given by

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1}{n-1} + \frac{\hat{p}_2\hat{q}_2}{m-1}} = -0.019 \pm 0.125 = [-0.144, 0.106].$$

4b. For a credibility interval we can use the non-informative uniform prior $p \in Beta(1,1)$. Adding the pseudo-counts $(1,1)$ to the total counts $(8+12, 48+62)$ we get $p \in Beta(21, 111)$ as the posterior distribution. Using Matlab one can find the exact 95% credibility interval $[a, b]$ for p by finding the 2.5% and 97.5% quantiles of the posterior distribution.

Posterior mean $\mu = \frac{21}{21+111} = 0.16$ and standard deviation $\sigma = \sqrt{\frac{0.16(1-0.16)}{132}} = 0.03$ leads to the normal approximation of the posterior distribution with mean 0.16 and standard deviation 0.03. This yield an approximate 95% credibility interval : $0.16 \pm 1.96 \cdot 0.03 = [0.10, 0.22]$.

5a. The risk ratio compares the chances to suffer from myocardial infarction under the aspirin treatment vs the chances to suffer from myocardial infarction under the placebo treatment:

$$RR = \frac{P(MyoInf|Aspirin)}{P(MyoInf|Placebo)}.$$

5b. The null hypothesis of $RR = 1$ is equivalent to the hypothesis of homogeneity.

	MyoInf	No MyoInf	Total
Aspirin	104 (146.5)	10933 (10887.5)	11037
Placebo	189 (146.5)	10845 (10887.5)	11034
Total	293	21778	22071

The corresponding chi-square test statistic is

$$X^2 = \frac{42.5^2}{146.5} + \frac{42.5^2}{146.5} + \frac{42.5^2}{10887.5} + \frac{42.5^2}{10887.5} = 25.$$

Since df=1 we can use the normal distribution table. The square root of 25 is 5 making the result highly significant. Aspirin works!

6a. The null distribution of Y is $\text{Bin}(n, \frac{1}{2})$ as each observation is smaller than the true median (assuming that the distribution is continuous) with probability 0.5.

6b. A non-parametric CI for the midean M is given by $(X_{(k)}, X_{(n-k+1)})$ where k is such that

$$P_{H_0}(Y > n - k) \approx 0.025.$$

With $n = 25$ we find k using the normal approximation with continuity correction:

$$0.025 \approx P_{H_0}(Y > 25 - k) = P_{H_0}\Big(\frac{Y - 12.5}{2.5} > \frac{13-k}{2.5}\Big) \approx P\Big(Z > \frac{13-k}{2.5}\Big).$$

Thus $\frac{13-k}{2.5} \approx 1.96$ and we get $k = 8$. The approximate 95% CI for M is given by $(X_{(8)}, X_{(18)})$.