FINAL EXAM Blue

BMED2400
Brani Vidakovic
Tuesday, 12/10/2013.

Name _____

Group #/ GT # _____ / _____

• For a full credit support your answers by uploading an annotated MATLAB-published output from calculations to T-square.

• On the last page of this exam booklet you will find a peer evaluation form. Filling this form is worth 5 points.

• By taking this exam, you pledge that this is your work and you have neither given nor received inappropriate help during the taking of this exam in compliance with the Academic Honor Code of Georgia Tech. You CAN use all internet resources available, however NO communication with a 3rd party is allowed. Academic misconduct will not be tolerated. You are to uphold the honor and integrity bestowed upon you by the Georgia Institute of Technology. Do NOT sign nor take this exam if you do not agree with the honor code.

1. Prostate	2. Leukoplakia	3. Burns	4. CBF/CMRO2	5. **T/F**	6. Peer	Total
/22	/17	/18	/18	/20	/5	/100

1. Prostate Cancer Data. This data set comes from the study by Stamey et al (1989) that examined the relationship between the level of serum prostate specific antigen (Yang polyclonal radioimmunoassay) and a number of histological and morphometric measures in 97 patients who were about to receive a radical prostatectomy. The data are organized as data structure `prost` with first 8 fields (`prost.lcavol - prost.pgg45`) as predictors, and 9th filed (`prost.lpsa`) as the response.

x_1	`prost.lcavol`	logarithm of cancer volume
x_2	`prost.lweight`	logarithm of prostate weight
x_3	`prost.age`	patient's age
x_4	`prost.lbph`	logarithm of benign prostatic hyperplasia amount
x_5	`prost.svi`	seminal vesicle invasion, 0 – no, 1 – yes.
x_6	`prost.lcp`	logarithm of capsular penetration
x_7	`prost.gleason`	Gleason score
x_8	`prost.pgg45`	percentage Gleason scores 4 or 5
y	`prost.lpsa`	logarithm of prostate specific antigen

Table 1: Fields in structure file `prost`. First 8 fields are predictors, and the last is the response to be modeled.

(a) Load the data into MATLAB and run procedure `stepwise`. Write down the regression equation suggested by the `stepwise`.

(b) Mr Smith (a new patient) has response $y = 2.3$ and covariates:

$x_1 = 1.4$, $x_2 = 3.7$, $x_3 = 65$, $x_4 = 0.1$, $x_5 = 0$, $x_6 = -0.16$, $x_7 = 7$, and $x_8 = 30$.

How close to the measured response $y = 2.3$ the regression from (a) predicts y for Mr Smith? Denote this prediction by $\hat{y}_p$. Calculate the residual $r = \hat{y}_p - y$. *Hint:* In calculating $\hat{y}_p$ you should use only covariates x_i suggested by `stepwise` procedure.

(c) The best (in R^2 sense) single predictor for y is x_1 – the logarithm of cancer volume. Fit the univariate regression using x_1 as the predictor. What is $\hat{y}_p$ for Mr Smith based on this univariate regression? Find 95% prediction interval for y_p. Is $y = 2.3$ in the interval?

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. Radical prostatectomy treated patients. *Journal of Urology*, **141**, 5, 1076–1083.

2. Leukoplakia. Data from Hamerle and Tutz (1980) come from a study on leukoplakia, which is a clinical term used to describe patches of keratosis visible as adherent white patches on the membranes of the oral cavity. Although most leukoplakia patches are benign and considered not dangerous, sometimes they are coexistent with oral cancer. Often cancers on the floor of the mouth, beneath the tongue, occur next to areas of leukoplakia.

The objective is to explore the association between the disease and smoking. The data on this association are stratified by the alcohol consumption level, also considered to be a risk factor.

		Leukoplakia	
		Yes	No
Alcohol	Smoker		
No	Yes	26	10
	No	8	8
$(0g, 40g]$	Yes	38	8
	No	43	24
$(40g, 80g]$	Yes	4	1
	No	14	17
$> 80g$	Yes	1	0
	No	3	7

Table 2: Contingency table for oral leukoplakia.

(a) Using Haenszel-Mantel procedure test the hypothesis that smoking and the disease are associated. Use $\alpha = 0.05$.

(b) Aggregate over alcohol consumption levels into a single 2 x 2 table on smoking vs disease status. Does the test for the association agree with the decision from (a)?

Hamerle, A. K. P. and Tutz, G. (1980). Kategoriale Reaktionen in multifaktoriellen Versuchsplanen und mehrdimensionale Zusammenhangsanalysen. *Archiv für Psychologie*, 53–68.

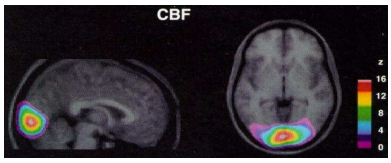
3. Third-degree Burns. [**Text 17.8 (p. 691)**] The data for this problem, discussed in Fan et al. (1995), refer to $n = 435$ adults who were treated for third-degree burns by the University of Southern California General Hospital Burn Center. The patients were grouped according to the area of third-degree burns on the body. For each midpoint of the groupings "log(area +1)," the number of patients in the corresponding group who survived, and the number who died from the burns are provided in table below.

log(area+1)	Survived	Died
1.35	13	0
1.60	19	0
1.75	67	2
1.85	45	5
1.95	71	8
2.05	50	20
2.15	35	31
2.25	7	49
2.35	1	12

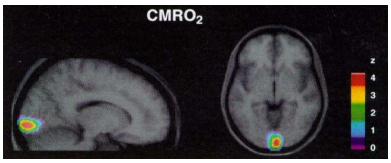
• Fit the logistic regression on the probability of death due to third-degree burns with `log(area+1)` as a covariate.

• Using your model, estimate the probability of survival for a person for which `log(area + 1)` equals 2.

4. Cerebral Blood Flow and Metabolic Rate of Oxygen. In the human brain, structures having high blood flow (e.g. grey matter) also tend to have increased metabolic rates, and structures containing low blood flow (e.g. white matter) have lower metabolic rates. This phenomenon is known as coupling, and it can be modeled by measuring cerebral blood flow (CBF) and comparing it with the variable cerebral metabolic rate of oxygen (CMR_{O_2}). This technique has been used to show how various brain regions differ in their metabolism characteristics and has been used as a fundamental basis for fMRI imaging. Fox and Raichle (1986) tested whether the measurements of CBF and CMR_{O_2} were dynamically coupled in the measurements they collected.



(a)



(b)

Three sets of stimulated-state measurements and one set of resting state measurements were acquired for each of the 9 subjects undergoing sensory stimulation. The 3 stimulation levels (S1, S2, S3) differed only in stimulus duration. The measured response to tactile stimulation for each subject was the regional blood flow ratio (contralateral/ipsilateral). Data are provided in the table below:

Subject	Resting	S1	S2	S3
1	0.82	1.14	1.05	1.12
2	1.05	1.45	1.38	1.42
3	1.10	1.45	1.49	1.58
4	0.90	1.05	1.22	1.14
5	1.02	1.27	1.27	1.39
6	1.03	1.25	1.23	1.20
7	1.04	1.30	1.33	1.29
8	0.98	1.30	1.42	1.34
9	1.04	1.29	1.23	1.24

Load the data set `cbfcmro2.xls` into MATLAB.

(a) Why should you not use a paired t-test to find significant differences between the population means for the four groups? (Resting vs S1, Resting vs S2, S1 vs S2, etc.) Give your arguments in a paragraph.

(b) Conduct the repeated measures analysis in a non-parametric fashion and test the hypothesis that population distributions of regional blood flow ratios are the same. This would imply that the population means for the four groups are the same,

$$H_0 : \mu_{Resting} = \mu_{S1} = \mu_{S2} = \mu_{S3}.$$

(c) If H_0 from (b) is rejected, which population means differ? Use $\alpha = 0.05$.

Hint. This is Friedman's test. You may use either built-in MATLAB's function `friedman.m` or "home-made" `friedmanGT.m` with `friedmanpairwise.m`. For the home-made make sure versions on your laptop are current (as in the course web page).

Fox, P. T. and M. E. Raichle (1986). Focal physiological uncoupling of cerebral blood-flow and oxidative-metabolism during somatosensory stimulation in human subjects. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 4, 1140–1144.

5. Question [value: 1 point each]	TRUE	FALSE
1. The sample variance of {-2, 0, 0, 2} is 8/3		
2. An experiment results in N possible outcomes. $N(A)$ of those N are favorable for event A, $N(B)$ are favorable for event B, and $N(AB)$ are favorable for both A and B. The probability of either A or B is calculated as $(N(A) + N(B) + N(AB))/N$.		
3. Two events, A and B both have positive probabilities, but share no outcomes. They are exclusive, that is, AB is impossible event. Therefore A and B are independent events.		
4. Spearman's coefficient of correlation is, in fact, the Pearson coefficient of correlation applied on ranks of data.		
5. We use ANOVA table in ANOVA and regression procedures.		
6. VIF - Variance Inflation Factor for a variable in multiple or multivariable regression represents the factor by which variance of the error inflates if the variable is omitted.		
7. For the multivariable regression given as $y = X\beta + \epsilon$, the following is (standardly) assumed: components of ϵ are i.i.d. normal $\mathcal{N}(0, 1)$.		
8. Logistic regression assumes that responses are either Bernoulli or Binomial, depending whether there is a single or multiple response for each value of the covariate.		
9. Generalized Linear Models (GLM) have non-linear fits. For example, the logistic fit is a sigmoidal function. The epithet *linear* is used because the predictors enter the model as a linear combination.		
10. When $n_1 = n_2$, it is always possible to interpret the 2-sample t-test as one sample t-test on the differences between the observations.		
11. In testing H_0 versus H_1, the null hypothesis is not rejected. Therefore H_0 is confirmed.		
12. Testing ANOVA hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ at level $\alpha = 0.05$ is equivalent to testing three pairwise hypotheses $H_{01} : \mu_1 = \mu_2$, $H_{02} : \mu_2 = \mu_3$, and $H_{03} : \mu_3 = \mu_1$, each at 5% level. H_0 is rejected if at least one of H_{01}, H_{02}, or H_{03} is rejected.		
13. ANOVA is generalization of a two sample t-test with equal population variances and repeated measures design is generalization of a paired t-test.		
14. Contingency tables and 2-Way ANOVA are always applicable to the same data: tabular with two factors. The inferences made by the two are equivalent.		
15. In ANY testing situation you can devise a test for which the sensitivity exceeds 96.8%.		
16. Tech Trolley interarrival times at the station next to Klaus bldg are exponentially distributed with (rate) parameter $\lambda = 1/5$ (units 1/min) so the expected interarrival time is 5 min. If you arrive at the station at random time your expected waiting time for the next trolley is less than 5 minutes.		
17. Wilcoxon Signed rank test can be used to test both: the equality of two population locations using paired samples and testing the population location using a single sample.		
18. Random variable X is normal with mean $\mu = 1$ and variance $\sigma^2 = 9$. The probability $P(-5 \leq X \leq 7)$ is less than 90%.		
19. Observations $X_1 = 1$ and $X_2 = 2$ are coming from an exponential $\mathcal{E}xp(\lambda)$ distribution where λ is the rate parameter. The MLE of λ is 3/2.		
20. Power of the test can be interpreted as the probability of rejection region under H_1.		

6. Self/Peer Evaluation. The score should reflect the team-work efforts and contribution to Activities, Midterm 1, and Class Project.

In the table below please list your group members and assign the scores to yourself and group members. Use the scale from 0 to 100. For example, for Activities: Myself 30/100, Jim 30/100, Jane 40/100; Total 100/100. If you wish, you are welcome to provide justification for your evaluation, or comment on any other aspect of the group work.

Group member	Activities	Midterm 1	Project
Myself	/100	/100	/100
	/100	/100	/100
	/100	/100	/100
Total	100/100	100/100	100/100

Justification/Comments: