STATISTICS

Comprehensive Examination Part I
Paper 2 – Applications

1. Consider the following data randomly collected from a clinic during a specific month, to examine the association between shoulder and back pains:

	Shoulder pain	
Back pain	Presence ($Y = 1$)	Absence ($Y = 0$)
Presence ($Z = 1$)	$n_{11}(= 220)$	$n_{12}(= 90)$
Absence ($Z = 0$)	$n_{21}(= 60)$	$n_{22}(= 30)$

   a What is (are) the necessary assumption(s) in order for this data to be reasonably analyzed with a generalized linear model?

   b Parameterize your distributional assumptions such that one parameter represents row effect, one for column effect and one for association.

   c Let $\{n_{rc}, \text{ for } r = 1, 2 \text{ and } c = 1, 2\}$ be the cell numbers and Let $n = \sum n_{rc}$. Justify that, conditional on n, $\{n_{rc}, \text{ for } r = 1, 2 \text{ and } c = 1, 2\}$ follows the multinomial distribution. What are the cell probabilities of this distribution?

   d Show that in general the multinomial likelihood is maximized at $\hat{\pi}_{ij} = n_{ij}/n$, and using the data from the Table, compute the likelihood estimates for the marginal probabilities

$$Pr[Y = 1], \text{ nd } Pr[Z = 1].$$

   e Provide the likelihood estimating equations for parameters in your model.

2. Suppose that n observations were classified into a three-way table with classifiers X , Y and Z. A loglinear model will be applied to analyze the data. Partial information about the three categorical variables is available. What are the corresponding models with the information provided?

   a X, Y and Z are mutually independent.

   b X is jointly independent of Y, Z.

   c X and Y are conditionally independent, given Z

   d There is no three-factor interaction.

   e No information of independence is available.

3. Suppose we are interested in the estimation of the size of a finite population by using a simple capture-recapture design, i.e., in a first stage, K subjects were sampled and tagged before being released into the population; after some time, in a second stage a new sample of size n was drawn from the same population. Out of those n captured at the second stage, y had a tag. Assume that the population size and the number of tagged units did not change between stages and the sampling was carried with replacement.

   a Show that while y/n is an unbiased estimator of K/N with minimum variance, nK/y doesn't share those properties.

   b Compute the expectation and variance of $\hat{N} = K(n+1)/(y+1)$ and comment on the properties of this estimator with respect to nK/y.

4. a) In designed experiments, we are interested to compare the different levels of one factor or combination of different factor levels (treatments). But in reality, there are other factors which may have influence on the response, but we are not interested to test the effect of those factors (nuisance factors). Discuss how these nuisance factors are controlled in designing the experiments, with suitable examples in the following three cases
   i) Controllable nuisance factors
   ii) Uncontrollable nuisance factors, but can be measured
   iii) Uncontrollable nuisance factors, but cannot be measured

   b) A taste panel will convene this afternoon to compare six different brands of ice cream. The panel is comprised of 10 persons who are expert tasters. The maximum number of different brands that an individual taster will taste is 3. Suggest a suitable design. What would you do if three of the expert tasters failed to come this afternoon because of illness, so that you could not run the design recommended earlier?

5. Design an eight-run fractional factorial design for an experimenter with the following five factors each at two levels: temperature, concentration, pH, agitation rate, and catalyst type. She tell you she is particularly concerned about the two factor interactions between temperature and concentration and between catalyst type and temperature. She would like a design, if possible to construct one, with main effects unconfounded with one another. Can you help her for constructing the desired design and if not possible give justification and propose alternate one? For the proposed design, show the alias structure of main effects and two factor interactions.

6. A consumer group conducted an experiment to compare the effectiveness of 3 commercially available weight-reducing diets 1, 2, and 3, Thirty volunteers were randomly assigned to the three diets (10 to each diet). Their weights (in pounds) were recorded both at the beginning and after 1 month on the respective diets. The group wanted to answer the following questions from the data collected.

   i. Are the three diets achieving similar weight reductions?

   ii. Does the initial weight affect the weight loss?

   iii. Is the effect of the initial weight (w_b) on the weight loss the same for the 3 diets?

   Suppose that 3 variables x_1, x_2, and x_3 are defined as follows: For $j = 1$, 2, and 3,

$$x_j = 1, \quad \text{for people on diet } j,$$
$$x_j = 0, \quad \text{otherwise.}$$

   You, the analyst, then fit Model 1:

$$\text{Model 1:} \qquad w_l = b_1\, x_1 + b_2\, x_2 + b_3\, x_3 + b_4\, w_b + e,$$

   where w_l = weight loss and w_b = the initial weight. Some output from R is given below:

```
Coefficients:
     Estimate   Std. error  t value  Pr(> |t|)
x1  -18.38773    7.06714    -2.602    0.015104*
x2  -15.23466    7.33789    -2.076    0.047905*
x3  -19.28102    7.17408    -2.688    0.012384*
wb    0.13703    0.03176     4.314    0.000205***

Residual standard error: 7.967 on 26 degrees of freedom
Multiple R-squared: 0.7607, Adjusted R-squared: 0.7239
F-statistics: 20.66 on 4 and 26 DF, p-value: 9.199e-08
Residual Sum of Squares: 1650.12
---
```

   You also fit Model 2:

$$\text{Model 2:} \qquad w_l = b_0 + b_4\, w_b + e.$$

Partial output from R is given below:

```
Coefficients:
              Estimate   Std. error  t value  Pr(> |t|)
(intercept)  -18.16734    6.79884    -2.672    0.012422*
wb             0.13954    0.03132     4.455    0.000123***
Residual standard error: 7.883 on 28 degrees of freedom
Multiple R-squared: 0.4148, Adjusted R-squared: 0.3939
F-statistics: 19.84 on 1 and 28 DF, p-value: 0.0001229
Residual Sum of Squares: 1740.09
---
```

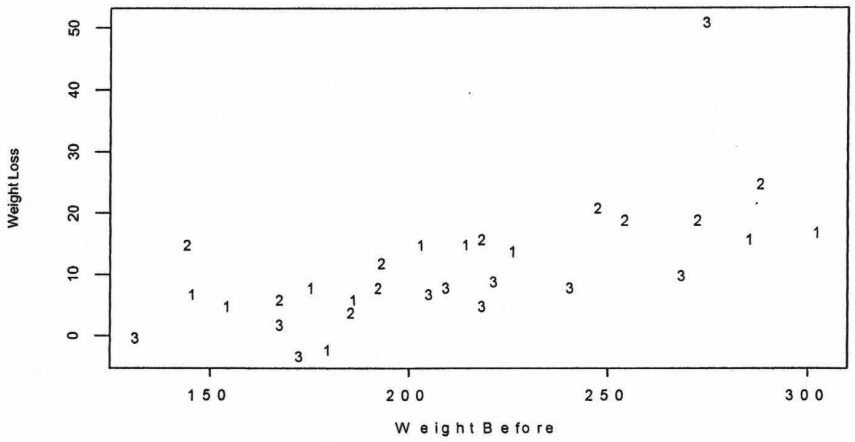
(a) Interpret the parameters b_4 and $b_2 - b_1$ in Model 1.

(b) Test the appropriate hypothesis to determine if the three diets are achieving similar weight reductions, clearly specifying the hypothesis being tested. State your conclusions.

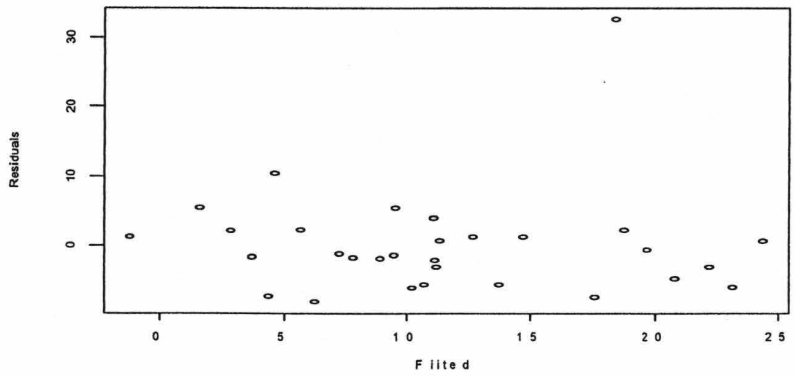
(c) The following 3 plots (see the next pages) and residual information are also available. **Briefly** comment on the plots and the residual information.

(d) You also fit Model 1 without observation 28 obtaining *Residual Sum of Squares = 378.83*. Similarly when Model 2 is fit without observation 28, you obtain *Residual Sum of Squares = 691.41*. Explain whether the removal of observation 28 alters any of your earlier conclusions regarding your answer to question (i).

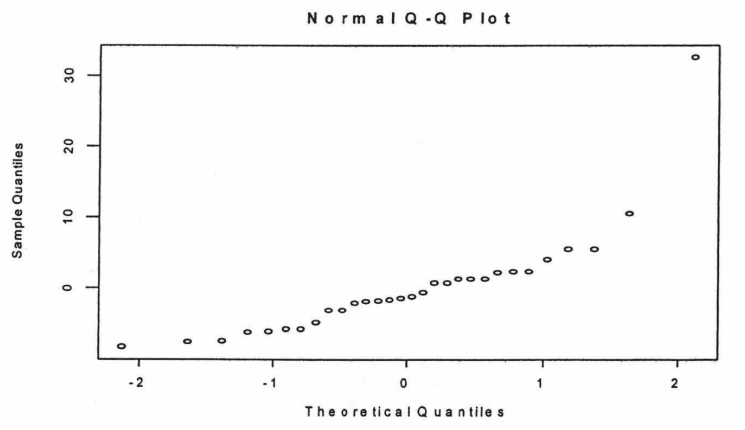
(e) How would you anser question (iii) (i.e., "Is the effect of the initial weight (w_b) on the weight loss the same for the 3 diets?"). What other model or models would you consider in order to answer question (iii) and what other hypotheses would you test?



wloss vs wb. Plot symbol=diet #



Residual vs Fitted plot



Normal Probability Plot

The leverages (diagonal elements of the 'hat matrix' $H = X(X'X)^{-1}X'$) are given below.

No	1	2	3	4	5	6	7	8	9	10
Leverage	0.1058	0.1969	0.1124	0.1162	0.1002	0.1445	0.2437	0.1609	0.1008	0.1069
No	11	12	13	14	15	16	17	18	19	20
Leverage	0.1229	0.1092	0.1153	0.1824	0.1001	0.1498	0.1824	0.1382	0.1084	0.1153
No	21	22	23	24	25	26	27	28	29	30
Leverage	0.1005	0.1018	0.1301	0.2004	0.1236	0.1000	0.1525	0.1641	0.1138	0.1009

The studentized residuals are given below

No	1	2	3	4	5	6	7	8	9	10
St Res	0.1798	-0.7103	-1.1647	0.3203	0.7595	0.3080	-0.9347	0.7789	0.5490	0.1734
No	11	12	13	14	15	16	17	18	19	20
St Res	-0.1001	-0.4511	-0.8812	1.5198	0.1713	-0.4563	0.0930	-0.2550	0.1084	0.3174
No	21	22	23	24	25	26	27	28	29	30
St Res	-0.2721	-0.2992	-0.2472	0.1770	-1.0513	-0.2088	-1.0914	4.7268	-0.8087	-0.8008

7. Consider the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of a response variable, $\boldsymbol{X}$ is an $n \times p$ full rank covariate matrix ($p < n$), $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors. The n errors have indpendent $N(0, \sigma^2)$ distributions.

(a) Write down the least squares estimate of $\boldsymbol{\beta}$ in terms of the vector $\boldsymbol{Y}$ and the matrix $\boldsymbol{X}$.

(b) Write down the vector of least squares residuals in terms of the vector $\boldsymbol{Y}$ and the matrix $\boldsymbol{X}$.

(c) Derive the distribution of the least squares estimator of $\boldsymbol{\beta}$. (You may use without proving standard results on linear transformations of normal random variables.)

(d) Provide an appropriate test statistic for testing $H_0 : \beta_j - \beta_k = 0$ against $H_1 : \beta_j - \beta_k \neq 0$, where $j, k = 1, \ldots, p$ and $j \neq k$ are two parameters in $\boldsymbol{\beta}$. Indicate how you would find the p–value.