MATH3424 Regression Analysis

Assignment 4

1. Using the following summary statistics from a data set with three columns: y, x_1 and x_2.

$n = 20$, $\quad \sum_{i=1}^{20} x_{i1} = 132$, $\quad \sum_{i=1}^{20} x_{i2} = 119$, $\quad \sum_{i=1}^{20} y_i = 375$,

$\sum_{i=1}^{20} x_{i1}^2 = 1100$, $\quad \sum_{i=1}^{20} x_{i1}x_{i2} = 969$, $\quad \sum_{i=1}^{20} x_{i2}^2 = 955$, $\quad \sum_{i=1}^{20} x_{i1}y_i = 3104$,

$\sum_{i=1}^{20} x_{i2}y_i = 2926$, $\quad \sum_{i=1}^{20} y_i^2 = 9187$,

$S_{x_1x_1} = 228.8000$, $\quad S_{x_1x_2} = 183.6000$, $\quad S_{x_2x_2} = 246.9500$, $\quad S_{x_1y} = 629.0000$,

$S_{x_2y} = 694.7500$, $\quad S_{yy} = 2155.7500$.

$\boldsymbol{x_{2c} = 0}$
$n = 10$, $\quad \sum x_{i1} = 52$, $\quad \sum y_i = 112$, $\quad \sum x_{i1}^2 = 366$, $\quad \sum x_{i1}y_i = 782$,

$\sum y_i^2 = 1742$, $\quad S_{x_1x_1}^{(0)} = 95.6000$, $\quad S_{x_1y}^{(0)} = 199.6000$, $\quad S_{yy}^{(0)} = 487.6000$.

$\boldsymbol{x_{2c} = 1}$
$n = 10$, $\quad \sum x_{i1} = 80$, $\quad \sum y_i = 263$, $\quad \sum x_{i1}^2 = 734$, $\quad \sum x_{i1}y_i = 2322$

$\sum y_i^2 = 7445$, $\quad S_{x_1x_1}^{(1)} = 94.0000$, $\quad S_{x_1y}^{(1)} = 218.0000$, $\quad S_{yy}^{(1)} = 528.1000$.

$\boldsymbol{x_{1c} = 0, x_{2c} = 0}$
$n = 5$, $\quad \sum y_i = 26$, $\quad \sum y_i^2 = 228$, $\quad S_{yy}^{(00)} = 92.8000$.

$\boldsymbol{x_{1c} = 0, x_{2c} = 1}$
$n = 5$, $\quad \sum y_i = 96$, $\quad \sum y_i^2 = 1846$, $\quad S_{yy}^{(01)} = 2.8000$.

$\boldsymbol{x_{1c} = 1, x_{2c} = 0}$
$n = 5$, $\quad \sum y_i = 86$, $\quad \sum y_i^2 = 1514$, $\quad S_{yy}^{(10)} = 34.8000$.

$\boldsymbol{x_{1c} = 1, x_{2c} = 1}$
$n = 5$, $\quad \sum y_i = 167$, $\quad \sum y_i^2 = 5599$, $\quad S_{yy}^{(11)} = 21.2000$.

and

$$\begin{pmatrix} 228.8000 & 183.6000 \\ 183.6000 & 246.9500 \end{pmatrix}^{-1} = \begin{pmatrix} 0.010834 & -0.008055 \\ -0.008055 & 0.010038 \end{pmatrix},$$

Part I **Consider a situation in which the data set is divided into two parts according to the values of $\boldsymbol{x_2}$. The categorical variable, $\boldsymbol{x_{2c}}$, is equal to 0 if $\boldsymbol{x_2}$ is less than its sample median and equal to 1 otherwise. For the categorical variable $\boldsymbol{x_{2c}}$, choose "0" as reference group**.

Fit a model of $\boldsymbol{y}$ on $\boldsymbol{x_1}$ and $\boldsymbol{x_{2c}}$, i.e.,

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2c} + \beta_{12} x_{i1} x_{i2c} + e_i \quad , \quad e_i \sim_{iid} N(0, \sigma^2)}$$

(a) Write down the fitted lines of y on x_1 for different values of x_{2c}.

(b) Hence or otherwise, find the fitted line for the model of y on x_1, x_{2c} and their interaction term.

(c) Find the unbiased estimate of the unknown parameter σ^2. No need to show that it is unbiased.

(d) Test $\beta_{12} = 0$ by t-test at the significance level $\alpha = 0.05$. State the test statistic, the critical value and your decision rule clearly.

(e) Based on the model in (b) (i.e., $\beta_{12} \neq 0$), test whether the difference of $E(y)$ between $x_{2c} = 1$ and $x_{2c} = 0$ at $x_1 = 4$, is significant at $\alpha = 0.05$. State the test statistic, the critical value and your decision rule clearly.

Part II **Consider a situation in which the data set is divided into four parts according to the values of x_1 and x_2. The categorical variable, x_{1c}, is equal to 0 if x_1 is less than its sample median and equal to 1 otherwise. The categorical variable, x_{2c}, is equal to 0 if x_2 is less than its sample median and equal to 1 otherwise. For both categorical variables, x_{1c} & x_{2c}, choose "0" as reference group**.

Define μ_{ij} to be the population mean of y when $x_{1c} = i, x_{2c} = j$ for $i, j = 0, 1$.

(a) Formulate a suitable model using the population means of these four parts, μ_{ij}. Then, estimate all unknown parameters in the model.

(b) Test the population means of these four parts are equal at $\alpha = 0.05$. Write down the test statistic, critical value and your conclusion clearly.

(c) Find orthogonal single-degree-of-freedom contrasts such that the sum of their contrast S.S. is equal to Reg.S.S. in (b). Hence or otherwise, find the significant contrast(s) at the level of $\alpha = 0.05$ to explain the inequality of population means of y for different combinations of x_{1c} & x_{2c}.

(d) Re-write the model in (a) as a model with the interaction term between x_{1c} and x_{2c} and then estimate the unknown parameters in the model.

(e) Write down the hypothesis that there is no interaction between x_{1c} and x_{2c} in terms of μ_{ij}. Hence or otherwise, perform the test at the significance level of 0.05. Write down the test statistic, critical value and your conclusion clearly.

(f) Assume that there is no interaction between x_{1c} and x_{2c}.

i. Find the unbiased estimate of the unknown parameter σ^2. No need to show that it is unbiased.

ii. Find the sum of squares for the main effect of "x_{1c}". Hence or otherwise, test whether the population means of y for $x_{1c} = 1$ and $x_{1c} = 0$ are different at $\alpha = 0.05$.

iii. Find the sum of squares for the main effect of "x_{2c}". Hence or otherwise, test whether the population means of y for $x_{2c} = 1$ and $x_{2c} = 0$ are different at $\alpha = 0.05$.

2. An experiment was conducted to study the effect of polymer and type of temperature on the amount of suspended solids in a coal cleaning system. Two types of polymers and two temperature levels were used in the experiment. The following results were recorded:

	Temperature	
Polymer	1	2
1	34, 32.7, 32, 33.2	29.8, 26.7, 28.7
2	30.1, 29.8, 29	28.1, 28.8, 27.3, 29.7, 28.3

(a) Estimate σ^2.

(b) Test the hypothesis that there is no interaction between polymer and temperature. Hence or otherwise, perform the test at the significance level of 0.05. Write down the test statistic, critical value and your conclusion clearly.

(c) Assume that there is no interaction between polymer and temperature.

i. Find the unbiased estimate of the unknown parameter σ^2. No need to show that it is unbiased.

ii. Find the sum of squares for the main effect of "polymer". Hence or otherwise, test whether the population mean amounts of suspended solids in a coal cleaning system are different at $\alpha = 0.05$.

iii. Find the sum of squares for the main effect of "temperature". Hence or otherwise, test whether the population mean amounts of suspended solids in a coal cleaning system are different at $\alpha = 0.05$.

(d) Assume that there is interaction between polymer and temperature. Test the mean difference of the amount of suspended solids between two levels of temperature for the 1st type of polymer at $\alpha = 0.05$.

3. In addition to ten men (details are given in Q2 of Assignment 2), eleven women were also studied during a maximal exercise treadmill test. Based on the observations from women, we obtain the following table of parameter estimates and standard error.

Variable	Parameter Estimate $\hat{\beta}_i$	Standard Error
Intercept	-51.9625	26.5128
x_1	-0.4168	0.2014
x_2	0.4415	0.09881
x_3	0.3629	0.1599

The $\hat{\sigma}^2$ for women is equal to 16.7327.

(e) Assuming that the population variances of y for men and women are equal, estimate the common variance.

(f) Test whether the regression coefficients of y on x_2 for men and women are equal at the 5% significance level. Write down your estimate, standard error, test statistic, critical value and your conclusion clearly.