MATH3424 Regression Analysis

Assignment 2

1. Using the following summary statistics:

$n = 20,\quad \sum_{i=1}^{20} x_{i1} = 58,\quad \sum_{i=1}^{20} x_{i2} = 87,\quad \sum_{i=1}^{20} y_i = 219,$

$\sum_{i=1}^{20} x_{i1}^2 = 194,\quad \sum_{i=1}^{20} x_{i1}x_{i2} = 265,\quad \sum_{i=1}^{20} x_{i2}^2 = 1003,\quad \sum_{i=1}^{20} x_{i1}y_i = 696,$

$\sum_{i=1}^{20} x_{i2}y_i = 1559,\quad \sum_{i=1}^{20} y_i^2 = 3091,$

$S_{x_1x_1} = 25.8000,\quad S_{x_1x_2} = 12.7000,\quad S_{x_2x_2} = 624.5500,\quad S_{x_1y} = 60.9000,$

$S_{x_2y} = 606.3500,\quad S_{yy} = 692.9500.$

and

$$\begin{pmatrix} 25.8000 & 12.7000 \\ 12.7000 & 624.5500 \end{pmatrix}^{-1} = \begin{pmatrix} 0.0391516 & -0.000796133 \\ -0.000796133 & 0.00161734 \end{pmatrix},$$

to fit the following model

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i} \quad , \quad \boldsymbol{e_i \sim_{iid} N(0, \sigma^2)}$$

Assume that β_0 β_1 and β_2 is unknown.

(a) Re-write the model to a centered model. Find the least squares estimates of the unknown parameters β_0, β_1 and β_2. Then, write down the fitted line.

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

(c) Construct an ANOVA table and then test H_0: $\beta_1 = \beta_2 = 0$ at significance level of $\alpha = 0.05$. Write down your conclusion clearly.

(d) Test the null hypothesis that $H_0 : \beta_1 - \beta_2 = 0$ against the alternative hypothesis that $H_1 : \beta_1 - \beta_2 \neq 0$ at the significant level of $\alpha = 0.05$. Construct the test statistic using

i. t-test. Write down the test statistic, the critical value and your conclusion clearly.

ii. F test in terms of "Increase in Regression Sum of Squares". Write down the test statistic, the critical value and your conclusion clearly.

iii. F test for testing $H_0 : \underset{\sim}{C}\underset{\sim}{\beta} = \underset{\sim}{d}$ Write down the test statistic, the critical value and your conclusion clearly.

2. Ten men were studied during a maximal exercise treadmill test. The dependent and independent variables are: y = VO$_{2max}$, x_1 = weight, x_2 = HR$_{max}$, x_3 = SV$_{max}$. The table of parameter estimates, standard error and covariance matrix is given below:

			Covariance Matrix			
Variable	$\hat{\beta}_i$	St. Error	Intercept	x_1	x_2	x_3
Intercept	-1.4545	22.2144	493.4780	-2.1663	-1.5222	-0.4450
x_1	-0.6985	0.1281	-2.1663	0.01641	0.004525	0.0001291
x_2	0.2895	0.07810	-1.5222	0.004525	0.006099	0.0008443
x_3	0.4481	0.05110	-0.4450	0.0001291	0.0008443	0.002611

(a) Find the t-value for testing the statistical significance of $\beta_3 = 0$. Do we reject $\beta_3 = 0$ at the 5% significance level?

(b) Construct a 95% confidence interval for β_1.

(c) Test whether the ratio of the regression coefficient of x_2 to that of x_3 is equal to 0.5 at the 5% significance level. Write down your test statistic, critical value and your conclusions clearly.

(d) Fill in the missing values in the analysis of variance table below. Is the regression significant at the 5% significance level?

Source	Sum of Squares	D.F.	Mean Squares	F value
Regression				
Residual	55.9687			—
Total	1305.0760		—	—

3. Consider the following model (Model A)

$$y_i = \alpha + \beta(x_i - \bar{x}) + \gamma(z_i - \bar{z}) + e_i, \quad 1 \leq i \leq n$$

where now α, β and γ are unknown scalar parameters, where $e_i \underset{\sim}{iid} N(0, \sigma^2)$ with σ^2 known, and where

$$\bar{x} = n^{-1} \sum x_i, \ \bar{z} = n^{-1} \sum z_i$$

(a) Find $\hat{\beta}$ and $var(\hat{\beta})$.

(b) Let $\tilde{\beta}$ be the least squares estimate of β under the model (Model B)

$$y_i = \alpha + \beta(x_i - \bar{x}) + e_i, \quad 1 \leq i \leq n$$

Find $\tilde{\beta}$ and $var(\tilde{\beta})$, and show that $Var(\tilde{\beta}) \leq Var(\hat{\beta})$. When does this equality hold?

4. Consider the studentized residuals

$$\frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}} \qquad \text{where } s = \sqrt{\sum_{i=1}^{n} \hat{e}_i^2/(n-2)}, \quad \hat{e}_i = y_i - \hat{y}_i$$

The denominator is found by merely constructing the variance of $y_i - \hat{y}_i$, namely

$$Var(y_i - \hat{y}_i) = \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right]$$

and then standardizing $y_i - \hat{y}_i$.

(a) Show that

$$\sum_{i=1}^{n} \frac{Var(y_i - \hat{y}_i)}{\sigma^2} = n - 2$$

(b) Under the conditions that the e_i are i.i.d. N(0, σ^2), does the studentized residual have a t-distribution with $n - 2$ degrees of freedom? If not, why not?

5. (Bonus) Suppose that one assumes the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \qquad e_i \sim N(0, \sigma^2)$$

but in fact $\beta_2 = 0$ and thus the above model is an overfitted model. Prove that the residual mean squares for the overfitted model is still an unbiased estimator for σ^2 when $\beta_2 = 0$.

Hint: Use the face that:

Let $\underset{\sim}{Y}$ be a n random vector and let E($\underset{\sim}{Y}$) $= \underset{\sim}{\mu}$, Cov($\underset{\sim}{Y}$) $= \underset{\sim}{\Sigma}$. Then E[$\underset{\sim}{Y}^T A \underset{\sim}{Y}$] $=$ trace($\underset{\sim}{A}\underset{\sim}{\Sigma}$) $+ \underset{\sim}{\mu}^T \underset{\sim}{A}\underset{\sim}{\mu}$