STAT 3008 Applied Regression Analysis
Midterm
9:00-11:00. Tuesday, 5 March 2013

Name: ______________________ **Major:** ______________________

1. (40 marks) An insurance company is studying the relationship between the size of medical insurance claims y and the current insured value x, both in \$. Data from a sample of recent claims are given below.

x	50,000	100,000	150,000	200,000	250,000	300,000	400,000
y	3,120	4,800	6,500	9,100	16,100	23,900	39,000

a (6 marks) What diagram is the best to represent the dataset? Plot it and comment on the suitability of a simple linear regression of y on x for modeling these data.

b (8 marks) An actuary suggests that the data might reasonably be fitted by a model of the form

$$y = Aexp(Bx + e)\,,$$

where A and B are constants and e is a zero-mean, constant-variance error term. Find a function f such that the model can be expressed in the form

$$f(y) = a + bx + \epsilon\,.$$

Identify a, b and ϵ in terms of A, B and e. Plot the values of $f(y)$ and x on a diagram and comment on the actuary's suggestion.

c (12 marks) You are given that

$$\sum x_i = 1.45{\times}10^6 \ \sum f(y_i) = 64.76, \sum f(y_i)^2 = 603.99, \ \sum x_i^2 = 3.875{\times}10^{11}, \sum x_i f(y_i) = 1.406{\times}10^7 .$$

Calculate the least squares simple linear regression line of $f(y)$ on x and draw this line on the scatter diagram in part (b). Use this regression to calculate a point estimate of the average claim size for a insurance plan of current insured value \$ 250,000.

d (10 marks) Draw an ANOVA table to test the dependence (at 5% significance level) between the (transformed) size of insurance claim and current insured value. (State the range of p-value).

e (4 marks) You are given that the simple linear regression of y on x is

$$y = -6723 + 0.1032x\,.$$

Plot this regression on the scatter diagram in part (a) and use it to find a point estimate of the average claim size for a plan of current insured value \$ 250 000, compare your result with that of part (c) and say with reasons which estimate you consider to be the more reliable.

2. (40 marks) Let $Y = (2, 3, 4, 5, 8, 2)'$, $X1 = (2, 1, 3, 0, 8, 5)'$, $X2 = (3, 2, 5, 6, 3, 0)'$. It is found that

$$(X'X)^{-1} = \begin{pmatrix} 1.313 & -0.1417 & -0.2204 \\ ? & 0.0284 & 0.0164 \\ ? & ? & 0.0532 \end{pmatrix}$$

a (10 marks) Fit a regression model $Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + e$. Give the fitted coefficients and estimate of error variance.

b (5 marks) Test the significance of $X1$ at 10% significance level.

c (10 marks) Test the joint significance of $X1$ and $X2$ at 5% significance level.

d (5 marks) Construct a 95% prediction interval for a new observation with $X1 = 6, X2 = 4$.

e (5 marks) Find $\widehat{\text{var}}(\hat{\beta}_1 - \hat{\beta}_2)$.

f (5 marks) Use (e), test at 5% significance level
$H_o : \beta_1 - \beta_2 = 0$, against
$H_A : \beta_1 - \beta_2 \neq 0$.

3. (5 marks) For the multiple regression model $Y = X\beta + e$, $e \sim N(0, \sigma^2)$, Let $\hat{\beta} = (X'X + kI)^{-1}X'Y$. What is $E(\hat{\beta})$?

4. (5 marks) Consider the mean function: $E[Y|X_i = x_i, i = 1, 2, 3] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Suppose that you are going to test if X_1 is significant after adjusting for X_3 but ignoring X_2. Write down the null and alternative hypotheses. If the number of observations $n = 20$, write down the degrees of freedom of 1) the t statistic and 2) the F statistics for this hypothesis test.

5. (5 marks) Y, X_1 and X_2 are observations from five subjects. Consider the regression models

$$\begin{array}{ll} 1) & Y = a_1 + b_1 X_2 + e\,. \\ 2) & X_1 = a_2 + b_2 X_2 + e\,. \\ 3) & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e\,. \end{array}$$

The first four residuals from the fitting of model 1) are -0.415, 0.108, -1.13, 0.994.
The last four residuals from the fitting of model 2) are -0.373, -0.0868, -1.21, 1.29.
Find the estimate of β_1 for the regression model 3).

6. (5 marks) Consider the regression model $Y = X\beta + e$, $e \sim N(0, \sigma^2 I)$, where Y and e are $n \times 1$, X is $n \times p$, I is the identity matrix of size n. Let $H = X(X'X)^{-1}X'$. Recall that the residual sum of square is $RSS = Y'(I - H)Y$.

a) (2 marks) Show that $RSS = e'(I - H)e$.

b) (3 marks) Suppose that $(I - H)$ can be written as QDQ^T, where Q satisfies $Q^T Q = I$ (orthonormal matrix), and D is a diagonal matrix with $n - p$ ones and p zeros in the diagonal entries. Show that the Residual sum of square is $\sigma^2 \chi^2_{n-p}$ distributed.