Statistics Qualifying Examination

Answer all questions and show all work.
This exam is closed-note/book. You need to use a calculator.

1. Let X and Y be independent identically distributed random variables, each with exponential distribution with mean $\lambda > 0$. Define $U = X + Y$ and $V = X - Y$.

   (a) Find the probability density function of V.

   (b) Find the correlation coefficient between U and V. (*Hint*: Use the formula for covariance directly.)

   (c) Are U and V independent? Justify your answer.

2. Answer the following questions:

   (a) Let Z be a random variable having the standard normal distribution. Find the moment generating function, $M_Z(t)$, of Z. Show work.

   (b) Let X have a normal distribution with mean μ and standard deviation σ. Find the moment generating function, $M_X(t)$, of X using the answer in (a).

   (c) Let $Y = e^X$. Find $E(Y)$ and $V(Y)$.

3. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with the probability density function (pdf) as

$$f(x;\theta) = \begin{cases} -\dfrac{1}{\theta}, & \theta < x < 0 \text{ and } -\infty < \theta < 0; \\ 0, & \text{elsewhere.} \end{cases}$$

   Let $Y_1 < Y_2 < \ldots < Y_n$ be the order statistics.

   (a) Derive the cumulative distribution function of the sample minimum Y_1.

   (b) Show that Y_1 converges in probability to θ, i.e., Y_1 is a consistent estimator for θ.

4. Let $X_1, X_2, \ldots, X_n$ be identically and independently distributed random variables with the pdf

$$f(x;\theta) = \theta x^{\theta-1},$$

where $0 < x < 1$ and $\theta > 0$. It is known that the maximum likelihood estimator (MLE) is $-\frac{n}{\sum_{i=1}^{n} \log X_i}$.

(a) Based on the asymptotic normality of the MLE, find the asymptotic $100(1-\alpha)\%$ confidence interval for θ.

(b) Find the exact $100(1-\alpha)\%$ confidence interval for θ.

(c) Find the exact likelihood ratio test (LRT) for the hypotheses $H_0 : \theta = 1$ vs $H_1 : \theta \neq 1$. Clearly specify the test statistic, its sampling distribution under H_0, and the decision rule with the rejection region for a size α-test.

5. Suppose that we conduct the simple regression analysis with $n = 4$ observations: $\{(y_1, x_1), (y_2, x_2), \cdots, (y_n, x_n)\}$ with the simple linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \;\; \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

The R code below is showing: 1) the vector y (response) and x (predictor); 2) the calculated hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$; 3) the linear model output from the `lm` function in R.

```
> y=c(4,5.5,6.2,7.8)
> x=c(2,3,4,5)
>
> X=cbind((rep(1,4)),(x))
>
> H=X%*%solve(t(X)%*%X)%*%t(X)
> H
     [,1] [,2] [,3] [,4]
[1,]  0.7  0.4  0.1 -0.2
[2,]  0.4  0.3  0.2  0.1
[3,]  0.1  0.2  0.3  0.4
[4,] -0.2  0.1  0.4  0.7
```

```
> SLR=lm(y~x)
> summary(SLR)

Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   ??????     0.4455   3.681  0.06651 .
x             1.2100     0.1212   9.980  0.00989 **
---
Signif. codes:  0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2711 on 2 degrees of freedom
Multiple R-squared:  0.9803,    Adjusted R-squared:  0.9705
F-statistic:  99.6 on 1 and 2 DF,  p-value: 0.009892
```

(a) Write down the fitted model *and* interpret $\hat{\beta}_1$, the estimated regression coefficient of x_1.

(b) Based on this output, find the residuals $\mathbf{e} = (e_1, e_2, e_3, e_4)^T$ from the above regression. (*Hint:* Recall the matrix/vector expression of the residual vector $\mathbf{e}$ in terms of the hat matrix and data vector.)

(c) Given the above output, find the *estimated* variance of all 4 residuals, $Var(e_i)$ for $i = 1, 2, 3, 4$. Are these estimated variances equal? (*Hint:* $Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$. You also need to estimate σ^2.).

(d) Given the above output, find the *estimated* correlation between residual e_1 and e_4.

6. A multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

with $\epsilon \sim N(0, \sigma^2\mathbf{I})$ is fitted to a dataset ($\mathbf{I}$ is an identity matrix).

Refer to the software output below and answer the following questions.

Model: MODEL1
Dependent Variable: Y

Number of Observations Read	81
Number of Observations Used	81

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	******	138.32691	34.58173	26.76	<.0001
Error	******	98.23059	1.29251		
Corrected Total	******	236.55750			

Root MSE	1.13689	R-Square	0.5847
Dependent Mean	15.13889	Adj R-Sq	0.5629
Coeff Var	7.50970		

```
## Coefficients:
##               Estimate    Std. Error  t value  Pr(>|t|)
## (Intercept) 12.20059      0.57796     21.11    <.0001
## X1          -0.14203      NA          NA       <.0001
## X2           0.28202      0.06317      4.46    <.0001
## X3           0.61934      1.08681      0.57    0.5704
## X4           0.00000792  0.00000138    5.72    <.0001
```

We also have got two types of sums of squares.

```
Variable      Type I SS
X1            14.70852
X2            (1)
X3            8.38142
X4            (2)

Variable      Type III SS
X1            57.15802
X2            (3)
X3            0.41975
X4            42.32496
```

To answer the questions, you may need to use (some of) the following values.

(df_1, df_2)	$(2, 74)$	$(2, 75)$	$(2, 76)$	$(2, 77)$	$(2, 78)$
$F_{0.025;df_1,df_2}$	3.8790	3.8764	3.8739	3.8714	3.8690
$F_{0.05;df_1,df_2}$	3.1203	3.1186	3.1170	3.1154	3.1138
(df_1, df_2)	$(3, 74)$	$(3, 75)$	$(3, 76)$	$(3, 77)$	$(3, 78)$
$F_{0.025;df_1,df_2}$	3.2982	3.2958	3.2932	3.2908	3.2885
$F_{0.05;df_1,df_2}$	2.7283	2.7266	2.7249	2.7233	2.7218

(a) Fill in the blanks denoted by $\star\star\star\star\star\star$ in the ANOVA table.

(b) Fill in the numbered blanks **(1)**, **(2)**, and **(3)**.

(c) Fill in the blanks denoted by `NA`.

(d) *[For part d, even if you did not get the blanks (1), (2) and (3), please use (1)=72.802, (2)= 42.325, and (3)=25.759 in case you need to use these quantities.]* For the full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$, test the following hypotheses using $\alpha = 0.05$. Clearly specify the test statistic, its sampling distribution under the null hypothesis (including degrees of freedom), p-value (or range of p-value), and your conclusion.

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \text{ vs. } H_a : \text{ not all equals } 0$$

7. A clay tile company is interested in studying the effects of cooling temperature on strength. Since the company has five ovens which produce the tiles, four tiles were baked in each oven and then randomly assigned to one of the four cooling temperatures (° C). The data are shown below.

Cooling	Oven					
Temp	1	2	3	4	5	Mean
5	3	10	7	4	3	5.40
10	3	8	12	2	4	5.80
15	9	13	15	3	10	10.00
20	7	12	9	8	13	9.80
Mean	5.50	10.75	10.75	4.25	7.50	7.75

To answer the questions, you may need to use (some of) the following values.

(df_1, df_2)	$(3, 11)$	$(3, 12)$	$(4, 11)$	$(4, 12)$
$F_{0.025;df_1,df_2}$	4.6300	4.4742	4.2751	4.1212
$F_{0.05;df_1,df_2}$	3.5874	3.4903	3.3567	3.2592

(df_1, df_2)	$(3, 11)$	$(3, 12)$	$(4, 11)$	$(4, 12)$
$q_{0.025;df_1,df_2}$	4.3913	4.3243	4.8427	4.7614
$q_{0.05;df_1,df_2}$	3.8196	3.7729	4.2561	4.1987

df	3	4	11	12
$t_{0.01,df}$	4.5407	3.7469	2.7181	2.6810
$t_{0.02,df}$	3.4819	2.9985	2.3281	2.3027
$t_{0.03,df}$	2.9505	2.6008	2.0961	2.0764

(a) Give an appropriate model to analyze this data, and state the assumptions.

(b) If $MSE = 6.275$, compute the F-statistic to determine if there is a difference among the four cooling temperatures (use $\alpha = 0.05$).

(c) We would like to perform pairwise comparisons for the four cooling temperatures. Which procedure should you use? Calculate the critical difference. (You don't need to report the comparison results.)

(d) Suppose the company believes there is a jump in the strength at 12.5° C but otherwise cooling temperature has no effect (i.e., step function $---\_\_\_$). To test this, we find the following contrasts:

$$\mathbf{C}_1 = (1, -1, 0, 0)$$
$$\mathbf{C}_2 = (0, 0, 1, -1)$$
$$\mathbf{C}_3 = (1, 1, -1, -1)$$

Test these contrasts *simultaneously* by using an appropriate procedure (using $\alpha = 6\%$).

8. The percentage of hardwood concentration (HC) in raw pulp and the vat pressure are being investigated for their effects on the strength of paper. Three levels of hardwood concentration and three levels of pressure are selected. A factorial experiment with three replicates is conducted, and the following data are obtained:

	Pressure		
HC Percentage	400	500	600
2	24.9, 26.7, 23.2	27.6, 29.3, 26.3	54.3, 52.5, 55.6
4	31.5, 28.8, 25.6	37.0, 41.4, 44.0	34.0, 35.4, 42.8
6	20.4, 25.1, 26.1	35.0, 38.0, 27.0	49.6, 43.6, 53.0

Some summary statistics are given above.

```
grand mean: 35.51

HC Percent       MEAN   |    Pressure      MEAN
1               35.60   |       1         25.81
2               35.61   |       2         33.96
3               35.31   |       3         46.76

HC Percent   Pressure         MEAN
1            1                24.93
1            2                27.73
1            3                54.13
2            1                28.63
2            2                40.80
2            3                37.40
3            1                23.87
3            2                33.33
3            3                48.73
```

Suppose the following statistical model is used to fit the data.

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}; k = 1, 2, 3$$

where τ_i ($i = 1, 2, 3$) and β_j ($j = 1, 2, 3$) are the main effects of HC percentage, the main effects of pressure, respectively, and $(\tau\beta)_{ij}$ are their interactions. For parameter estimation, we impose the following constraints: $\sum_i \tau_i = \sum_j \beta_j = \sum_i (\tau\beta)_{ij} = \sum_j (\tau\beta)_{ij} = 0$.

(a) Calculate the estimate of $(\tau\beta)_{13}$.

(b) Calculate the sum of squares due to the main effect pressure.

(c) Part of the ANOVA table from SAS is given on the next page. Test if the interaction between HC percentage and pressure is significant ($\alpha = 5\%$). To get full credit, give the hypotheses, the value of the test statistic, the sampling distribution under the null hypothesis (including degrees of freedom), the (range of) p-value, and your conclusion. (You don't need to complete the ANOVA table.)

(df_1, df_2)	$(2, 18)$	$(4, 18)$	$(8, 18)$	$(2, 26)$	(4,26)	(8,26)
$F_{0.025;df_1,df_2}$	4.5597	3.6083	3.0053	4.2655	3.3289	2.7293
$F_{0.05;df_1,df_2}$	3.5546	2.9277	2.5102	3.3690	2.7426	2.3205

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2739.531852	342.441481	26.66	<.0001
Error	18	231.206667	12.844815		
Cor.Total	26	2970.738519			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
percent	2	0.520741	0.260370	0.02	0.9800
pressure	*	*******	********	*****	******
percent*pressure	*	*******	********	*****	******