DEPARTMENT OF STATISTICS, ACTUARIAL AND DATA SCIENCE
PH.D. QUALIFYING EXAMINATION – APPLIED STATISTICS
Time: 8am-11am (STA 590), 1pm-4pm (STA682), August 25, 2023

General Instructions

- There are two parts in this exam: STA 590 and STA 682. You are to answer all questions. The score for each part will be converted to its percentage.
- Write on one side only. Clearly label the problem number and subpart. You must **show** all your work and **justifications** correctly and completely to receive full credits. Partial credits may be given for partially correct solutions.
- For each problem/subproblem, hand in only the answer that you want to be graded. If necessary, please make clear, e.g., by crossing out the other answer(s), which answer should be graded. Crossed-out work will be ignored. Failure to follow this *instruction* for a *problem will result* in a *zero score* for that *problem*.
- If a theorem is applied, you must clearly state the theorem, identify its assumption(s) and conclusion(s), and justify why it is applicable. New notations must be defined before use.
- When finished, please collate all pages according to the problem numbers and then number the pages accordingly. Hand in also the exam paper.

By signing below, I hereby acknowledge that I have completely read and fully understand the instructions.

Signature

Printed Name

PART A: STA 590

This part consists of five problems, each with subparts. It has a possible total of 150 points.

Problem 1 (33 points)**:** A simulated dataset (n=30) has been generated by the following model:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + \mu_t$$

μ_t are independent $N(0, \sigma^2)$.

The **first four** columns at the following table listed the response variable Y, the explanatory variable X, autocorrelated error term ε, and the normal random variable μ. One way to deal with correlated data is using transformed data, $Y_t' = Y_t - \rho Y_{t-1}, X_t' = X_t - \rho X_{t-1}$. The first five observations of the dataset are listed.

t	X_t	μ_t	ε_t	Y_t	$Y_t{}'$	$X_t{}'$
0	20.00		2.00	52.00		
1	19.70	0.18	-1.12	48.28	$Y_1{}'$=?	$X_1{}'$=?
2	18.86	0.90	1.63	49.35	$Y_2{}'$=80.73	$X_2{}'$=31.67
3	19.78	-0.07	-1.13	48.42	$Y_3{}'$=80.50	$X_3{}'$=32.04
4	19.93	4.13	4.86	54.72	$Y_4{}'$=86.19	$X_4{}'$=32.78

The Cochrane-Orcutt procedure has estimated the ρ to be r=-0.65. The transformed data based on r=-0.65 are in the **fifth** and **sixth** columns.

The results from the simple linear regression based on response variable $Y_t{}'$ and independent variable $X_t{}'$ are:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3796.30517	3796.30517	931.10	<.0001
Error	27	110.08557	4.07724		
Corrected Total	28	3906.39074			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.56258	2.53604	6.93	<.0001
Xtrans	1	1.97597	0.06476	30.51	<.0001

Durbin-Watson D	1.815
Pr < DW	0.2385
Pr > DW	0.7615
Number of Observations	29
1st Order Autocorrelation	0.089

Please answer the following questions.

a) (2 points) Y_1'=
b) (2 points) X_1'=
c) (2 points) Estimate $\sigma^2\{\varepsilon_4\} =$
d) (3 points) Estimate $\sigma\{\varepsilon_3, \varepsilon_5\} =$
e) (6 points) Estimate $\boldsymbol{\sigma^2}\{\boldsymbol{\varepsilon}\}_{3x3}$ =? For $\boldsymbol{\varepsilon} = [\varepsilon_4, \varepsilon_5, \varepsilon_6]'$.
f) (6 points) Test whether the negative autocorrelation remains after transformation using α=0.05.

H_0 : H_1 :

Test Statistics:
p-value:
Conclusion: Reject H_0 or Fail to reject H_0

g) (6 points) Restate the estimated regression function in terms of the original variables. Also obtain $s\{b_0\}$ and $s\{b_1\}$.
h) (6 points) Test whether Y_t is positively linearly associated with X_t.

H_0 : H_1 :

Test Statistics:
i) p-value:
Conclusion: Reject H_0 or Fail to reject H_0

Problem 2 (25 points)**:** In an enzyme kinetic study the velocity of a reaction (Y) is expected to be related to the concentration (X) as follows:

$$Y_i = \frac{\gamma_0 X_i}{\gamma_1 + X_i} + \varepsilon_i$$

a) (5 points) Intrinsically linear models are nonlinear, but by using a correct transformation they can be transformed into linear regression models. Is this function,

$$Y_i = \frac{\gamma_0 X_i}{\gamma_1 + X_i} + \varepsilon_i$$

an **intrinsically linear** response function or **nonlinear** response function?

We will use the normal equation to obtain the least square estimates. To obtain the normal equations for

$$Y_i = f(\boldsymbol{X}_i, \boldsymbol{\gamma}) + \varepsilon_i$$

we need to minimize $Q = \sum_{i=1}^{n}[Y_i - f(\boldsymbol{X}_i, \boldsymbol{\gamma})]^2$ with respect to γ_0 and γ_1.

The partial derivative of Q with respect to γ_k is:

$$\frac{dQ}{d\gamma_k} = \sum_{i=1}^{n} -2[Y_i - f(\boldsymbol{X}_i, \boldsymbol{\gamma})]\left[\frac{df(\boldsymbol{X}_i, \boldsymbol{\gamma})}{d\gamma_k}\right].$$

When the *p* partial derivatives are each set equal to 0

b) (10 points) Describe how to obtain the initial value for γ_0 and γ_1.
c) (10 points) Obtain the two normal equations for γ_0 and γ_1 with estimates g_0 and g_1.

SENIC dataset: The primary objective of the study on the efficacy of nosocomial infection control (SENIC) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial infection in United States hospitals. This data set contains of a random sample of 113 hospital selected from the original 338 hospitals surveyed. The variables we are interested include:

Length of Stay (LOS): Average length of stay of all patients in hospital (in days)
Age (Age): Average age of patients (in years)
Infection risk (Risk): Average estimated probability of acquiring infection in hospital (in percent)
Medical school affiliation (School): 1=Yes, 2=No.
Region (Region): Geographics region, where 1=NE, 2=NC, 3=S, 4=W.
First five rows of the data:

ID	LOS	Age	Risk	School	Region
1	7.13	55.7	4.1	2	4
2	8.82	58.2	1.6	2	2
3	8.34	56.9	2.7	2	3
4	8.95	53.7	5.6	2	4
5	11.2	56.5	5.7	2	1

This dataset is for **Problem 3**, **Problem 4** and **problem 5**.

Problem 3 (35 points) addressed the first research question," how medical school affiliation and region affect the infection risk". An ANOVA model for two-factor is proposed and the results is listed below.

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, i = 1.2, j = 1,2,3,4,$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	30.1153614	4.3021945	2.64	0.0150
Error	105	171.2644616	1.6310901		
Corrected Total	112	201.3798230			

R-Square	Coeff Var	Root MSE	Risk Mean
0.149545	29.32676	1.277141	4.354867

Source	DF	Anova SS	Mean Square	F Value	Pr > F
School	1	10.93551541	10.93551541	6.70	0.0110
Region	3	13.99693932	4.66564644	2.86	0.0404
School*Region	3	5.18290665	1.72763555	1.06	0.3698

Level of School	N	Risk	
		Mean	Std Dev
1	17	5.09411765	1.11213229
2	96	4.22395833	1.34028628

Level of Region	N	Risk	
		Mean	Std Dev
1	28	4.86071429	1.27114393
2	32	4.39375000	1.33921920
3	37	3.92702703	1.45900435
4	16	4.38125000	0.87652248

Level of School	Level of Region	N	Risk	
			Mean	Std Dev
1	1	5	5.60000000	1.28062485
1	2	7	4.62857143	1.08122505
1	3	3	5.86666667	0.30550505
1	4	2	4.30000000	0.42426407
2	1	23	4.70000000	1.23840073
2	2	25	4.32800000	1.41554465
2	3	34	3.75588235	1.39440248
2	4	14	4.39285714	0.93353281

a) (6 points) Please state the assumptions for the model proposed.
b) (6 points) Estimate α_2, β_2 and $(\alpha\beta)_{22}$.
c) (6 points) Test whether or not the two factors interact; using α=0.05.

H_0 : H_1 :

Test Statistics:
p-value:
Conclusion: Reject H_0 or Fail to reject H_0

d) (6 points) Test whether or not the effect for region is present; using α=0.05.

H_0 : H_1 :

Test Statistics:
p-value:
Conclusion: Reject H_0 or Fail to reject H_0

e) (6 points) The 90% family confidence coefficient intervals for all pairwise comparison of the means for region were obtained using the Bonferroni procedure. However, the comparison for NE (1) and S (3) are missing. Please compute the interval for $\mu_1 - \mu_3$ by hand to complete the table. State your findings and prepare a graphical summary by lining nonsignificant comparisons.

Comparisons significant at the 0.1 level			
Region Comparison	**Difference Between Means**	**Simultaneous 90% Confidence Limits**	
1 - 2	0.4670	-0.3371	1.2710
1 - 4	0.4795	-0.4943	1.4532
1 - 3			
2 - 4	0.0125	-0.9389	0.9639
2 - 3	0.4667	-0.2834	1.2168
3 - 4	-0.4542	-1.3839	0.4755

f) (5 points) Using the Scheffe procedure, obtain confidence interval for the following comparisons for weight gain with 95% family confidence coefficient:

$$L_1 = \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2}.$$

Problem 4 (20 points) addressed the second research question," how region affect the infection risk". An ANOVA model for one-factor is proposed and the results is listed below.

$$Y_{ij} = \mu_{..} + \alpha_i + \varepsilon_{ij}$$

a) (6 points) Please complete the analysis of variance table.

Source of Variation	df	SS	MS	F	p-value
Region					
Error					
Total					

b) (6 points) Test whether or not the effect for region is present; using α=0.05.

H_0 : H_1 :

Test Statistics:

P-value:

Conclusion: Reject H_0 or Fail to reject H_0

c) (8 points) The data is fitted by a multiple linear regression model using the following SAS code.

```
Proc GLM data=SENIC;
 class Region (ref=1);
 Model Risk=Region/solution;
run;
```

Please estimate all the parameters for this multiple linear regression model.

The hospital with infection risk greater than 5% is considered in the high-risk group. A binary variable, RiskHigh is defined as

$$RiskHigh = \begin{cases} 1 & if\ Risk > 5\% \\ 0 & if\ Risk < 5\% \end{cases}$$

Problem 5 (37 points) addressed the third research question, "How variables, such as age, length of stay and region associated with RiskHigh?" A set of four models (**A, B, C, D**) included some or all of the three predictor variables were considered. Three dummy variables, X_1, X_2, and X_3 were created for region (1=NE, 2=NC, 3=S, 4=W) variable.

$$X_1 = \begin{cases} 1 & if\ region = NE \\ 0 & Otherwise \end{cases}, X_2 = \begin{cases} 1 & if\ region = NC \\ 0 & Otherwise \end{cases}, X_3 = \begin{cases} 1 & if\ region = S \\ 0 & Otherwise \end{cases}$$

The four multiple logistic regression models considered were:

$$E\{RiskHigh = 1\} = \pi = \frac{exp(\boldsymbol{X}'\boldsymbol{\beta})}{1 + exp(\boldsymbol{X}'\boldsymbol{\beta})}$$

Model **A** (Region, Age, LOS): $\boldsymbol{X}'\boldsymbol{\beta}$=$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4\text{Age} + \beta_5\text{LOS}$
Model **B** (Age, LOS): $\boldsymbol{X}'\boldsymbol{\beta}$=$\beta_0 + \ \beta_4\text{Age} + \beta_5\text{LOS}$
Model **C** (Region, Age, LOS, Region*Age, Region*LOS): $\boldsymbol{X}'\boldsymbol{\beta}$=$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4\text{AGE} + \beta_5\text{LOS} + \beta_{14} X_1 * \text{Age} + \beta_{24} X_2 * \text{Age} + \beta_{34} X_3 * \text{Age} + \beta_{15} X_1 * \text{LOS} + \beta_{25} X_2 * \text{LOS} + \beta_{35} X_3 * \text{LOS}$
Model **D** (LOS): $\boldsymbol{X}'\boldsymbol{\beta}$=$\beta_0 + \beta_1\text{LOS}$
Analysis results were on page 9-13.

a) (4 points) Based on Model **A**, estimate the odds of been high-risk for a hospital from W(est) region with average patients' age=50 years old and length of stay=10 days.
b) (4 points) Based on Model **A**, what will be the maximum length of stay allowed to have the probability of been in high-risk less than 5% for a hospital in S(outh) and average patients' age=50?
c) (6 points) Conduct a Wald test to determine whether length of stay is related to the probability of been in high-risk group for Model **A**; using α=0.05.

   H_0 : H_1 :
   Test Statistics:
   p-value:
   Conclusion: Reject H_0 or Fail to reject H_0
d) (6 points) Conduct a likelihood ratio test to determine whether region is related to the probability of been in high-risk group for Model **A**; using α=0.05.

   H_0 : H_1 :
   Test Statistics:
   p-value:
   Conclusion: Reject H_0 or Fail to reject H_0
e) (6 points) Conduct a likelihood ratio test to determine whether the interaction terms, between age/length of stay and region, respectively, were related to the probability of been in high-risk group in Model **C**; using α=0.05.

   H_0 : H_1 :
   Test Statistics:
   p-value:
   Conclusion: Reject H_0 or Fail to reject H_0
f) (6 points) Conduct a goodness of fit test to detect whether Model **D** used logit link function is appropriate; using α=0.05.

   H_0 : H_1 :

Test Statistics:
p-value:
Conclusion: Reject H_0 or Fail to reject H_0

g) (5 points) Based on Model D used probit link function, estimate the probability of been in high-risk for a hospital with length of stay =12 days.

Analysis results:

Problem 5 Model A: Multiple Logistic Regression analysis on Region, Age and Length of stay to RiskHigh

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	136.682	112.973
SC	139.409	129.337
-2 Log L	134.682	100.973

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	33.7092	5	<.0001
Score	28.2421	5	<.0001
Wald	19.5711	5	0.0015

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Region	3	5.8766	0.1178
Age	1	0.6164	0.4324
LOS	1	17.8976	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-13.0773	3.9586	10.9132	0.0010
Region	1	1	-1.0171	0.5311	3.6681	0.0555
Region	2	1	0.2107	0.4044	0.2714	0.6024
Region	3	1	-0.4237	0.4249	0.9945	0.3187
Age		1	0.0461	0.0587	0.6164	0.4324
LOS		1	0.9946	0.2351	17.8976	<.0001

Problem 5 Model B: Multiple Logistic Regression analysis on Age and Length of stay to RiskHigh

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	136.682	113.323
SC	139.409	121.505
-2 Log L	134.682	107.323

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	27.3590	2	<.0001
Score	24.3038	2	<.0001
Wald	16.8873	2	0.0002

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.0441	3.4833	8.3148	0.0039
Age	1	0.0309	0.0559	0.3054	0.5805
LOS	1	0.7572	0.1866	16.4622	<.0001

Problem 5 Model C: Multiple Logistic Regression analysis on Region, Age, Length of stay and interactions to RiskHigh

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	136.682	114.574
SC	139.409	147.303
-2 Log L	134.682	90.574

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	44.1074	11	<.0001
Score	36.0266	11	0.0002
Wald	20.9400	11	0.0340

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-18.7276	6.3670	8.6516	0.0033
Region	1	1	-22.4364	15.6072	2.0666	0.1506
Region	2	1	3.7300	8.7361	0.1823	0.6694
Region	3	1	12.3504	8.1363	2.3041	0.1290
Age		1	0.1531	0.0999	2.3487	0.1254
LOS		1	0.9343	0.2738	11.6426	0.0006
Age*Region	1	1	0.3782	0.2533	2.2289	0.1355
Age*Region	2	1	-0.0963	0.1257	0.5874	0.4434
Age*Region	3	1	-0.3122	0.1342	5.4115	0.0200
LOS*Region	1	1	0.1046	0.4712	0.0493	0.8243
LOS*Region	2	1	0.2170	0.4424	0.2407	0.6237
LOS*Region	3	1	0.4538	0.4576	0.9835	0.3213

Problem 5 Model D: Simple Logistic Regression analysis on length of stay to RiskHigh with link function=Logit

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	136.682	111.631
SC	139.409	117.085
-2 Log L	134.682	107.631

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	27.0512	1	<.0001
Score	24.1121	1	<.0001
Wald	16.9425	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.4550	1.8686	20.4729	<.0001
LOS	1	0.7627	0.1853	16.9425	<.0001

Partition for the Hosmer and Lemeshow Test					
		RiskHigh = 1		RiskHigh = 0	
Group	Total	Observed	Expected	Observed	Expected
1	11	1	0.58	10	10.42
2	11	1	0.89	10	10.11
3	11	2	1.20	9	9.80
4	11	1	1.61	10	9.39
5	11	1	2.02	10	8.98
6	11	4	2.67	7	8.33
7	11	4	3.24	7	7.76
8	11	1	4.17	10	6.83
9	12	7	6.11	5	5.89
10	13	10	9.51	3	3.49

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
7.2002	8	0.5152

Problem 5 Model D: Simple Logistic Regression analysis on length of stay to RiskHigh with link function=Probit

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	136.682	111.871
SC	139.409	117.325
-2 Log L	134.682	107.871

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	26.8111	1	<.0001
Score	24.1121	1	<.0001
Wald	18.6891	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.8984	1.0140	23.3368	<.0001
LOS	1	0.4413	0.1021	18.6891	<.0001

Partition for the Hosmer and Lemeshow Test					
		RiskHigh = 1		RiskHigh = 0	
Group	Total	Observed	Expected	Observed	Expected
1	11	1	0.52	10	10.48
2	11	1	0.87	10	10.13
3	11	2	1.22	9	9.78
4	11	1	1.68	10	9.32
5	11	1	2.11	10	8.89
6	11	4	2.79	7	8.21
7	11	4	3.35	7	7.65
8	11	1	4.24	10	6.76
9	12	7	6.07	5	5.93
10	13	10	9.41	3	3.59

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
7.4362	8	0.4904