DEPARTMENT OF STATISTICS, ACTUARIAL AND DATA SCIENCE
PH.D. QUALIFYING EXAMINATION – APPLIED STATISTICS
Time: 8am-11am (STA 590), 1pm-4pm (STA682), August 21, 2025

General Instructions

- The exam consists of two parts: STA 590 and STA 682. You are required to answer all questions. The raw score for each part will be converted to its percentage.
- Write on one side of the paper only. Begin each subpart on a new sheet with the problem number clearly labeled. You must show all your work and justifications completely and correctly to receive full credit. Partial credit may be given for partially correct solutions.
- For each problem or subproblem, submit only the answer you want to be graded. Any crossed-out work will be ignored. Failure to follow this instruction for a problem will result in a zero score for the problem.
- If you apply a theorem, you must state the theorem, identify its assumptions and conclusions, and justify why it is applicable. New notations must be defined before use.
- When finished, please collate all pages based on problem labels and then number the pages accordingly. Hand in 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: Time series data on US income (*X*) and consumption expenditure (*Y*) during 1950–1993 (n=44) were collected. Part of the data is listed below.

Time	Income X (in 1987 USD)	expenditure Y (in 1987 USD)
1950	6284	5820
1951	6390	5843
1952	6476	5917
…	…	…
1991	14003	12899
1992	14279	13110
1993	14341	13391

Preliminary analysis indicated there is an autocorrelation on collected data. Therefore, an AR(1) model is proposed.

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

μ_t are independent $N(0, \sigma^2)$.

The Cochrane-Orcutt procedure has estimated the ρ to be r=0.792. A simple linear regression based on transformed response variable $Y_t{}'$ and independent variable $X_t{}'$ is used to model the data and the results are listed below.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12291182	12291182	1210.32	<.0001
Error	41	416369	10155		
Corrected Total	42	12707550			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-35.05153	62.71206	-0.56	0.5793
Xtrans	1	0.92626	0.02662	34.79	<.0001

The Durbin-Watson statistics based on transformed are:

Durbin-Watson D	1.973
Pr < DW	0.4044
Pr > DW	0.5956
Number of Observations	43
1st Order Autocorrelation	-0.055

a) (4 points) Estimate $\sigma^2\{\varepsilon_2\} =$
b) (4 points) Estimate $\sigma\{\varepsilon_2, \varepsilon_5\} =$

c) (6 points) Estimate $\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\}_{3x3}$ =? For $\boldsymbol{\varepsilon} = [\varepsilon_2, \varepsilon_3, \varepsilon_4]'$.

d) (6 points) Restate the estimated regression function in terms of the original variables. Also obtain $s\{b_0\}$ and $s\{b_1\}$.

e) (6 points) Test whether consumption expenditure (*Y*) is positively linearly associated with US income (*X*) after transformation.

H_0 : H_1 :

Test Statistics:

p-value:

Conclusion: Reject H_0 or Fail to reject H_0

f) (6 points) The US income (in 1987 USD) for 1994 is 14485. Predict the expenditure for 1994; employ a 95% prediction interval. Given $s\{pred\}$=110.03.

Problem 2: Consider the following model:

$$Y_i = exp(\gamma_0 + \gamma_1 \log X_i) + \varepsilon_i$$

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,

$$Y_i = exp(\gamma_0 + \gamma_1 \log X_i) + \varepsilon_i$$

an **intrinsically linear** response function or **nonlinear** response function? Please justify your answer.

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 through transformation.

c) (10 points) Obtain the two normal equations for γ_0 and γ_1 with least square estimates g_0 and g_1.

Surgical dataset: A hospital unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 54 patients was included for the study. The information for the following variables is collected for each patient:

Survival: in days after the operation.
Enzyme: Enzyme function test score
Age: in years
Gender: Male vs Female
Alcohol: None, Moderate, Heavy

First five rows of the data:

ID	Survival	Enzyme	Age	Gender	Alcohol
1	695	81	50	Male	Moderate
2	403	66	39	Male	None
3	710	83	55	Male	None
4	349	41	48	Male	None
5	2343	115	45	Male	Heavy

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

Problem 3 (35 points) addressed the first research question, " Are both gender (1=Female, 2=Male) and alcohol intake (1=Heavy, 2=Moderate, 3=None), and their interaction associated with survival days?" An ANOVA model for two-factor is proposed and the results is listed below.

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1898478.119	379695.624	2.82	0.0261
Error	48	6471042.418	134813.384		
Corrected Total	53	8369520.537			

R-Square	Coeff Var	Root MSE	Survival Mean
0.226832	52.29644	367.1694	702.0926

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Gender	1	251808.570	251808.570	1.87	0.1781
Alcohol	2	1467235.830	733617.915	5.44	0.0074
Gender*Alcohol	2	179433.719	89716.859	0.67	0.5187

Level of Gender	N	Survival	
		Mean	Std Dev
Female	25	775.640000	392.455611
Male	29	638.689655	397.366249

Level of Alcohol	N	Survival	
		Mean	Std Dev
Heavy	10	1046.30000	638.134791
Moderate	29	636.31034	253.376640
None	15	599.80000	320.685471

Level of Gender	Level of Alcohol	N	Survival	
			Mean	Std Dev
Female	Heavy	4	1205.00000	601.890909
Female	Moderate	14	676.07143	238.079322
Female	None	7	729.42857	404.493452
Male	Heavy	6	940.50000	694.294102
Male	Moderate	15	599.20000	269.647707
Male	None	8	486.37500	184.201704

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 alcohol 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 alcohol intake for survival days were obtained using the Bonferroni procedure.

However, the comparison for heavy and none is missing. Please compute the interval for $\mu_{Heavy} - \mu_{None}$ 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			
Alcohol Comparison	**Difference Between Mean**	**Simultaneous 90% Confidence Limits**	
Heavy - Moderate	409.99	114.97	705.01
Heavy - None			
Moderate - None	36.51	-219.35	292.37

f) (6 points) Using the Scheffe procedure, obtain confidence interval for the following comparisons for survival days with 95% family confidence coefficient:

$$L_1 = \frac{\mu_{Heavy}+\mu_{Moderate}}{2} - \mu_{None}.$$

Problem 4 (20 points) addressed the second research question," how alcohol intake affects the survival days". 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
Alcohol					
Error					
Total					

b) (6 points) Test whether or not the effect of alcohol 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=Surgical;
 class Alcohol (ref="Heavy");
 Model Survival= Alcohol /solution;
run;
```

Please specify the multiple linear regression model and estimate all the parameters for the model.

Survival days passed 730 days (2 years) is of interest. A binary variable, Survival2yr is defined as

$$Survival2yr = \begin{cases} 1 \ if \ Survival > 730 \\ 0 \ if \ Survival < 730 \end{cases}$$

Problem 5 (37 points) addressed the third research question, "How variables, such as age, enzyme and alcohol associated with Survival2yr?" 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, and X_2 were created for alcohol (Heavy, Moderate, None) variable.

$$X_1 = \begin{cases} 1 \ if \ Alcohol = Heavy \\ 0 \ Otherwise \end{cases}, \ X_2 = \begin{cases} 1 \ if \ Alcohol = Moderate \\ 0 \ Otherwise \end{cases},$$

The four multiple logistic regression models considered were:

$$E\{Survial2yr = 1\} = \pi = \frac{exp(\boldsymbol{X'\beta})}{1 + exp(\boldsymbol{X'\beta})}$$

Model **A** (Alcohol, Age, Enzyme): $\boldsymbol{X'\beta}$=$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \text{Age} + \beta_4 \text{Enzyme}$
Model **B** (Age, Enzyme): $\boldsymbol{X'\beta}$=$\beta_0 + \ \beta_4 \text{Age} + \beta_5 \text{Enzyme}$
Model **C** (Alcohol, Age, Enzyme, Alcohol*Age, Alcohol*Enzyme): $\boldsymbol{X'\beta}$=$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \text{AGE} + \beta_4 \text{Enzyme} + \beta_{13} X_1 * \text{Age} + \beta_{23} X_2 * \text{Age} + \beta_{14} X_1 * \text{Enzyme} + \beta_{24} X_2 * \text{Enzyme}$
Model **D** (Enzyme): $\boldsymbol{X'\beta}$=$\beta_0 + \beta_1 \text{Enzyme}$
Analysis results were on page 9-16.

a) (4 points) Based on Model **A**, estimate the odds of survived more than 2 years for a patient who is heavy alcohol user, 50 years old with enzyme score equal 80.
b) (4 points) Based on Model **A**, what will be the (maximum? minimum?) enzyme allowed to have the probability of survived more than 2 years to be (higher lower than?) 10% for a none alcohol user who is 50 years old?
c) (6 points) Conduct **a Wald test** to determine whether enzyme is related to the probability of survived more than 2 years 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 alcohol is related to the probability of survived more than 2 years 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/enzyme and alcohol, respectively, were related to the probability of survived more than 2 years 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 survived more than 2 years for a patient whose enzyme score equal 75.

Analysis results:

Problem 5 Model A: Multiple Logistic Regression analysis on Alcohol, Age and Enzyme to Survival2yr

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	67.631	56.801
SC	69.620	66.746
-2 Log L	65.631	46.801

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.8297	4	0.0008
Score	14.9547	4	0.0048
Wald	10.7917	4	0.0290

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Alcohol	2	3.2425	0.1977
Age	1	0.0038	0.9507
Enzyme	1	9.3768	0.0022

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.4945	2.8260	7.0328	0.0080
Alcohol	Heavy	1	0.9309	0.6346	2.1519	0.1424
Alcohol	Moderate	1	0.3217	0.5106	0.3971	0.5286
Age		1	0.00198	0.0320	0.0038	0.9507
Enzyme		1	0.0774	0.0253	9.3768	0.0022

Problem 5 Model B: Multiple Logistic Regression analysis on Age and Enzyme to Survival2yr

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	67.631	56.803
SC	69.620	62.770
-2 Log L	65.631	50.803

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.8277	2	0.0006
Score	11.8632	2	0.0027
Wald	9.0825	2	0.0107

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.9033	2.7526	6.2897	0.0121
Age	1	-0.00327	0.0313	0.0109	0.9168
Enzyme	1	0.0752	0.0250	9.0321	0.0027

Problem 5 Model C: Multiple Logistic Regression analysis on Alcohol, Age, Enzyme and interactions to Survival2yr

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	67.631	64.067
SC	69.620	81.968
-2 Log L	65.631	46.067

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	19.5635	8	0.0121
Score	16.2837	8	0.0385
Wald	10.6043	8	0.2251

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Alcohol	2	0.3517	0.8388
Age	1	0.0003	0.9851
Enzyme	1	6.3301	0.0119
Age*Alcohol	2	0.0032	0.9984
Enzyme*Alcohol	2	0.5675	0.7530

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-8.3370	3.4385	5.8786	0.0153
Alcohol	Heavy	1	-2.6140	5.7622	0.2058	0.6501
Alcohol	Moderate	1	2.3718	4.0899	0.3363	0.5620
Age		1	0.000962	0.0514	0.0003	0.9851
Enzyme		1	0.0879	0.0350	6.3301	0.0119
Age*Alcohol	Heavy	1	0.00394	0.0701	0.0032	0.9552
Age*Alcohol	Moderate	1	-0.00040	0.0561	0.0000	0.9944
Enzyme*Alcohol	Heavy	1	0.0388	0.0562	0.4761	0.4902
Enzyme*Alcohol	Moderate	1	-0.0240	0.0400	0.3589	0.5491

Problem 5 Model D: Simple Logistic Regression analysis on Enzyme to Survival2yr with link function=Logit

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	67.631	54.814
SC	69.620	58.792
-2 Log L	65.631	50.814

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.8168	1	0.0001
Score	11.8506	1	0.0006
Wald	9.0385	1	0.0026

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.0814	2.1777	10.5742	0.0011
Enzyme	1	0.0753	0.0250	9.0385	0.0026

Partition for the Hosmer and Lemeshow Test					
		Survival2yr = 1		Survival2yr = 0	
Group	Total	Observed	Expected	Observed	Expected
1	5	0	0.06	5	4.94
2	5	0	0.21	5	4.79
3	6	1	0.66	5	5.34
4	6	1	0.96	5	5.04
5	5	0	1.02	5	3.98
6	5	2	1.47	3	3.53
7	5	1	1.77	4	3.23
8	7	5	3.03	2	3.97
9	5	2	2.92	3	2.08
10	5	4	3.89	1	1.11

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.5283	8	0.6999

Problem 5 Model D: Simple Logistic Regression analysis on Enzyme to Survival2yr with link function=Probit

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	67.631	54.638
SC	69.620	58.616
-2 Log L	65.631	50.638

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.9931	1	0.0001
Score	11.8506	1	0.0006
Wald	10.1491	1	0.0014

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.1553	1.1879	12.2368	0.0005
Enzyme	1	0.0440	0.0138	10.1491	0.0014

Partition for the Hosmer and Lemeshow Test					
		Survival2yr = 1		Survival2yr = 0	
Group	Total	Observed	Expected	Observed	Expected
1	5	0	0.03	5	4.97
2	5	0	0.17	5	4.83
3	6	1	0.64	5	5.36
4	6	1	0.97	5	5.03
5	5	0	1.04	5	3.96
6	5	2	1.49	3	3.51
7	5	1	1.77	4	3.23
8	7	5	3.01	2	3.99
9	5	2	2.86	3	2.14
10	5	4	3.82	1	1.18

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.4663	8	0.7068