BIOS 6612 Midterm Review

- This review is taken from previous midterm exams given in this class.
- You do not need to turn this in (no extra credit).
- This practice exam is about twice as long as your exam will be, but the questions will be similar.
- Your exam will include questions drawn from the material we have covered in lectures 1-12.
- SAS output is given here but your exam will contain simple tables of estimates, standard errors, etc. (i.e., will not favor R or SAS). There will be no SAS or R code on your exam.
- We will discuss during the in-class review any questions that students have had trouble with, but we will not go through the entire practice exam, **so come prepared with attempts at solutions on your own to get the most out of the review session.**

$t_{196,0.975} = 1.9723$	$F_{2,198,0.95} = 3.042$
$t_{197,0.975} = 1.9721$	$F_{2,199,0.95} = 3.041$
$t_{198,0.975} = 1.9720$	
$t_{199,0.975} = 1.9719$	

$z_{0.975} = 1.96$	$\chi^2_{3,\,0.95} = 7.815$	$t_{4,0.975} = 2.776$
$\chi^2_{1,\,0.95} = 3.841$	$\chi^2_{4,\,0.95} = 9.488$	$t_{8,0.975} = 2.306$
$\chi^2_{2,\,0.95} = 5.991$	$\chi^2_{5,\,0.95} = 11.070$	$t_{10,0.975} = 2.228$

Question 1. A study was performed to examine the effect of diet on depression in 600 graduate students at one university.

Graduate students were randomized to diet groups such that 300 graduate students were assigned a standard American diet and 300 graduate students were assigned a plant based diet.

All food and beverages were provided by the study center for two months and no students dropped out of the study.

Information was also collected on the students' self-reported exercise for an average week.

After two months, the graduate students took an exam in order to determine if they were clinically depressed or not.

The following table provides the variable coding.

Variable Coding

diet	0 = standard American diet 1 = plant based diet
exercise	0 = exercise less than an average of 3 hours a week 1 = exercise greater than or equal to an average of 3 hours a week
depressed	0 = Not depressed 1 = Depressed

The following SAS programs were run and partial output is included on the next few pages.

Note: SAS gave the following output for all models:

Convergence criterion (GCONV=1E-8) satisfied.

Please note that this is not a real study. This example is just being used to test and illustrate concepts from the class.

Model 1.A

```
PROC LOGISTIC;
 MODEL depressed (EVENT = '1') = diet / COVB;
 FREQ n;
RUN;
```

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	830.547	772.283
SC	834.944	781.077
-2 Log L	828.547	768.283

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	60.2640	1	<.0001
Score	59.2252	1	<.0001
Wald	57.1256	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.4895	0.1189	16.9390	<.0001
diet	1	-1.3053	0.1727	57.1256	<.0001

Model 1.B

```
PROC LOGISTIC;
 MODEL depressed (EVENT = '1') = exercise / COVB;
 FREQ n;
RUN;
```

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	830.547	798.156
SC	834.944	806.950
-2 Log L	828.547	794.156

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.3913	1	<.0001
Score	34.0805	1	<.0001
Wald	33.4018	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.3483	0.1192	8.5341	0.0035
exercise	1	-0.9744	0.1686	33.4018	<.0001

Model 1.C

```
PROC LOGISTIC;
 MODEL depressed (EVENT = '1') = diet exercise / COVB;
 FREQ n;
RUN;
```

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	830.547	736.602
SC	834.944	749.793
-2 Log L	828.547	730.602

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	97.9446	2	<.0001
Score	92.7105	2	<.0001
Wald	81.9487	2	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.0821	0.1619	44.6895	<.0001
diet	1	-1.3913	0.1809	59.1761	<.0001
exercise	1	-1.0832	0.1807	35.9537	<.0001

Estimated Covariance Matrix			
Parameter	Intercept	diet	exercise
Intercept	0.0262	-0.01832	-0.01901
diet	-0.01832	0.032712	0.005487
exercise	-0.01901	0.005487	0.032637

Model 1.D

```
PROC LOGISTIC;
 MODEL depressed (EVENT = '1') = diet exercise diet* exercise / COVB;
 FREQ n;
RUN;
```

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	830.547	730.313
SC	834.944	747.900
-2 Log L	828.547	722.313

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	106.2344	3	<.0001
Score	101.0480	3	<.0001
Wald	86.6121	3	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.3949	0.2074	45.2163	<.0001
diet	1	-1.9354	0.2700	51.3933	<.0001
exercise	1	-1.6034	0.2632	37.1051	<.0001
diet*exercise	1	1.0454	0.3651	8.1959	0.0042

Estimated Covariance Matrix				
Parameter	Intercept	diet	exercise	dietexercise
Intercept	0.043029	-0.04303	-0.04303	0.043029
diet	-0.04303	0.072886	0.043029	-0.07289
exercise	-0.04303	0.043029	0.069287	-0.06929
dietexercise	0.043029	-0.07289	-0.06929	0.133332

Question 1.1 [5 points] For **Model 1.A**, is there a significant association between the odds of depression and diet using a Wald test? (Provide the odds ratio and corresponding 95% Wald CI.)

Question 1.2 [5 points] For **Model 1.A**, is there a significant association between the odds of depression and diet using a Likelihood Ratio test? (Provide the test statistic and corresponding p-value.)

Question 1.3 [5 points] For **Model 1.A**, despite a fairly large sample size of n=600 and a non-rare outcome in the sample with no sparsity in the cells of the contingency table, the chi-square test statistic for the association between the odds of depression and diet is smallest and the corresponding p-value is largest for the Wald test compared to the Score test and Likelihood Ratio test. Why is this the case in general?

Question 1.4 [10 points] For **Model 1D**, is there a significant association between the odds of depression and average weekly exercise among graduate students assigned to the standard American diet (diet=0)? (Provide an OR and 95% Wald CI.)

Question 1.5 [10 points] For **Model 1D**, is there a significant association between the odds of depression and average weekly exercise among graduate students assigned to the plant based diet (diet=1)? (Provide an OR and 95% Wald CI)

Question 1.6 [10 points] In **Model 1.D**, using a Likelihood Ratio Test, is the interaction between diet and exercise significantly associated with the odds of depression? (Provide a Likelihood Ratio Test Statistics to support your answer.)

Question 1.7 [5 points] For **Models 1.A, 1.B, 1.C**, and **1.D**, which is the best model based on AIC?

Question 1.8 [5 points] For **Models 1.A, 1.B, 1.C**, and **1.D**, which is the best model based on BIC?

Question 1.9 [5 points] In general, why do the models selected by AIC and BIC differ in Question 1.7 and 1.8?

Question 1.10 [5 points] Which model (1.A, 1.B, 1.C, & 1.D) is the saturated model? Or is the saturated model not given?

Question 1.11 [5 points] Use the Deviance Chi-Square Test Statistic to compare model 1.D and 1.C. If this question cannot be answered with the output provided, please state what output you would need to answer this question.

Question 1.12 [10 points] In the study for question 1 (Models 1.A, 1.B, 1.C, 1.D), are exactly 300 graduate students (i.e. 50% of the 600 graduate students) depressed after the 2 month diet? Justify your answer for full credit.

Question 2. An investigator ran the following code for a small study and was very confused. The study had a binary outcome (disease=1 for the disease and 0 otherwise), a binary exposure variable (exposure=1 for the exposure and 0 otherwise), and one binary covariate (covariate = 0 or 1). The SAS code and partial output is given below.

Model 2.

```
DATA test;
INPUT covariate exposure disease n;
DATALINES;
0 1 1 6
0 1 0 3
0 0 1 1
0 0 0 6
;
RUN;

PROC LOGISTIC;
 MODEL disease (EVENT = '1') = exposure covariate / COVB;
 FREQ n;
RUN;
```

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	4.7312	1	0.0296
Score	4.3900	1	0.0361
Wald	3.7049	1	0.0543

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

covariate =	0

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-1.7918	1.0801	2.7518	0.0971
exposure	1	2.4849	1.2910	3.7049	0.0543
covariate	0	0	.	.	.

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
exposure	12.000	0.956	150.688

Question 2.1 [10 points] For Question 2,

$$logit(\Pr(\text{disease}_i = 1)) = \beta_0 + \beta_E exposure_i + \beta_C covariate_i$$

In matrix form,

$$logit(\Pr(\mathbf{Y} = \mathbf{1})) = \boldsymbol{X\beta}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \; and \; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_E \\ \beta_C \end{bmatrix}$$

Give the dimensions of the matrix **X** and write ***X*** in matrix form with ALL the values inputted from Model 2 (i.e. elements of **X** should be 0s and 1s. Not $exposure_i$ or $covariate_i$). Use **Y** given above as a guide. Do NOT use dots (i.e. …) or arrows (i.e. ->). Give ALL of the elements of **X**.

Question 2.2 [5 points] Explain why the following note was given by SAS for the **Model 2** output.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

covariate =	0

Question 2.3 [5 points] The OR for the exposure is very large (i.e. OR=12) and the 95% Wald CI is very wide (0.956,150.688), but the corresponding p-value is relatively modest (i.e p-value=0.0543). Explain why the Wald test is not performing well in this scenario.

Study: A study was performed to examine whether dietary fiber intake has an effect on HbA1c levels. The HbA1c test (hemoglobin A1c test) is a laboratory test used to estimate average blood glucose levels. Normal HbA1c levels are 4%-6%, but are commonly higher in cigarette smokers. Dietary fiber intake was measured as a continuous variable (grams/day) and vitamin C usage was measured as a categorical variable from a food frequency questionnaire. The study sample consisted of 125 smokers and 75 non-smokers, for a total of 200 participants.

The following variables are available for the analysis:

hba1c:	hemoglobin A1c levels (%)
fiber :	dietary fiber intake (grams/day)
smoker:	current smoking status (0 = non-smokers; 1=smokers)
vitC:	supplement of vitamin C (2= large dose, 1= normal dose, 0=no dose)

Vitamin C Usage	*New indicator variables*		
	vitC_none	vitC_normal	vitC_large
vitC_none	1	0	0
vitC_normal	0	1	0
vitC_large	0	0	1

Model 1:

You perform a simple linear regression of HbA1c (*hba1c*) on dietary fiber intake (*fiber*). The following SAS output was obtained.

```
PROC REG;
 MODEL hba1c = fiber;
RUN;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	XXXXX	XXXXX	2.60748	XXXXX	0.2135
Error	XXXXX	XXXXX	XXXXX		
Corrected Total	XXXXX	XXXXX			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.62683	0.23910	27.72	<.0001
fiber	1	-0.01855	0.01487	-1.25	0.2135

Model 2:

You perform a t-test and a simple linear regression of HbA1c (*hba1c*) on smoking status (smoker). The following SAS output was obtained and then sections were blanked out.

```
proc ttest;
 var hba1c;
 class smoker;
 run;
```

smoker	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	125	5.5324	0.4719	0.0422	4.5300	7.5500
1	75	7.7157	1.0592	0.1223	5.5000	10.6600
Diff (1-2)		-2.1833	0.7475	0.1092		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	XXXXX	XXXXX	<.0001
Satterthwaite	Unequal	91.9	-16.87	<.0001

```
PROC REG;
 MODEL hba1c = smoker / covb;
RUN;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	Q2A	223.45052	XXXXX	Q2E	<.0001
Error	Q2B	Q2D	0.55881		
Corrected Total	Q2C	334.09564			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	Q2F	0.06686	82.74	<.0001
smoker	1	Q2G	XXXXX	Q2H	Q2I

Model 3:

You perform a linear regression of HbA1c (*hba1c*) on fiber, smoking status (smoker), and fiber*smoker (fiber_smoke). The following SAS output was obtained.

```
PROC REG;
MODEL hba1c = fiber smoker fiber_smoke / covb;
RUN;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	229.16117	76.38706	142.68	<.0001
Error	196	104.93447	0.53538		
Corrected Total	199	334.09564			

Root MSE	0.73170	R-Square	0.6859
Dependent Mean	6.35115	Adj R-Sq	0.6811
Coeff Var	11.52070		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.84243	0.16545	35.31	<.0001
fiber	1	-0.02118	0.01038	-2.04	0.0427
smoker	1	2.43152	0.28710	8.47	<.0001
fiber_smoke	1	-0.01549	0.01773	-0.87	0.3834

Covariance of Estimates

Variable	Intercept	fiber	smoker	fiber_smoke
Intercept	0.0273743493	-0.001577284	-0.027374349	0.0015772839
fiber	-0.001577284	0.0001077386	0.0015772839	-0.000107739
smoker	-0.027374349	0.0015772839	0.0824244199	-0.004724646
fiber_smoke	0.0015772839	-0.000107739	-0.004724646	0.0003144918

Model 4: You perform a linear regression of HbA1c (*hba1c*) on vitamin C usage where vitC_normal=1 for normal dosages of vitamin C & 0 otherwise, vitC_large=1 for large dosages of vitamin C & 0 otherwise, and vitC_none= 1 for no vitamin C dosage and 0 otherwise.

Model 4a:

```
PROC REG;
MODEL hba1c = vitC_none vitC_normal vitC_large  / noint covb;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	**Q4A**	**XXXXX**	**XXXXX**	**XXXXX**	<.0001
Error	**Q4B**	**XXXXX**	**Q4D**		
Uncorrected Total	**Q4C**	**XXXXX**			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
vitC_none	1	6.52479	0.11740	55.58	<.0001
vitC_normal	1	6.26842	0.20775	30.17	<.0001
vitC_large	1	5.94372	0.19530	30.43	<.0001

Covariance of Estimates			
Variable	**vitC_none**	**vitC_normal**	**vitC_large**
vitC_none	0.0137827786	0	0
vitC_normal	0	0.0431618594	0
vitC_large	0	0	0.0381430386

Model 4b:

```
PROC REG;
MODEL hba1c =vitC_normal vitC_large;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	10.98596	5.49298	3.35	0.0371
Error	197	323.10968	1.64015		
Corrected Total	199	334.09564			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	**Q4E**	**Q4G**	**XXXXX**	<.0001
vitC_normal	1	**Q4F**	**Q4H**	**XXXXX**	0.2840
vitC_large	1	-0.58107	0.22787	-2.55	0.0115

Model 5:

You perform a simple linear regression of HbA1c (*hba1c*) on dietary fiber intake (*fiber*) and smoking status (*smoker*). The following SAS output was obtained.

```
PROC REG;
 MODEL hba1c = fiber smoker;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	228.75255	114.37628	213.89	<.0001
Error	197	105.34308	0.53474		
Corrected Total	199	334.09564			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	5.92014	0.13943	42.46	<.0001
fiber	1	**XXXXX**	**XXXXX**	**Q5A**	0.0019
smoker	1	2.19877	0.10692	20.56	<.0001

Question 1. *10 points*. For **Model 1**, provide a brief interpretation of the association between dietary fiber intake and HbA1c levels, including the **95% CI**, point estimate, p-value, and decision.

Question 2A. *10 points*. Fill in the missing values for **Model 2** (parts **Q2A-Q2E**). **Justify your answers for full credit.**

Q2A.

Q2B.

Q2C.

Q2D.

Q2E.

Question 2B. *10 points.* Fill in the missing values for **Model 2** (parts **Q2F-Q2J**). **Justify your answers for full credit.**

Q2F.

Q2G.

Q2H.

Q2I.

Question 3A. *10 points*. Provide an interpretation of the relationship between HbA1c and fiber for non- smokers in **Model 3** (include a point estimate, test statistic and decision).

Question 3B. *10 points*. Provide an interpretation of the relationship between HbA1c and fiber for smokers in **Model 3** (include a point estimate, test statistic and decision).

Question 3C*. 10 points*. Does the relationship between HbA1c and fiber significantly depend on smoking status? Give a p-value to support this decision.

Question 4A. *10 points*. Fill in the missing values for **Model 4a** (parts **Q4A-Q4D**). **Justify your answers for full credit.**

Q4A.

Q4B.

Q4C.

Q4D.

Question 4B. *10 points*. Fill in the missing values for **Model 4** (parts **Q4E-Q4H**). **Justify your answers for full credit.**

Q4E.

Q4F.

Q4G.

Q4H.

Question 4C. *10 points.* Using **Model 4a**, test whether HbA1c is the same for those taking normal doses of vitamin C (vitC_normal) versus those taking large doses of vitamin C (vitC_large). **Provide only the null hypothesis and test statistic.**

Question 5. *10 points*. Give the absolute value of the t statistic for fiber in **Model 5** (part **Q5A**). **Show your work for full credit.** Hint: this question requires a partial F-test.

Question 6 Extra Credit*: 5 points.* Using the output for **Model 1** calculate the correlation between fiber and HbAc1. **Justify your answers for full credit.**