

**Bios 312:  
Modern Regression Analysis**

**Midterm Examination  
March 3, 2011**

Name: \_\_\_\_\_

**Instructions:** Please provide concise answers to all questions. Questions are of varying levels of difficulty, so you may find it advantageous to skip questions you find especially difficult, and return to these questions at the end of the exam.

You are allowed three (3) pages of your own notes to assist you when taking the exam.

You may use a calculator to assist with arithmetic. When making intermediate calculations, always use at least four significant digits; report at least three significant digits.

If you come to a problem that you believe cannot be answered without making additional assumptions, clearly state the reasonable assumption that you make, and proceed.

Please adhere to the following pledge. If you are unable to truthfully sign the pledge for any reason, turn in your paper unsigned and discuss the circumstances with the instructor.

**PLEDGE:** On my honor, I have neither given nor received unauthorized aid on this examination

\_\_\_\_\_

This exam consists of

11 total pages

175 total points

4 total questions, each with multiple parts

Question 1: 65 pts, parts (a) – (l)

Question 2: 30 pts, parts (a) – (d)

Question 3: 30 pts, parts (a) - (c)

Question 4: 50 pts, parts (a) - (g)

**Question 1 (65 points):** Suppose that we are interested in the association between albumin, height, and current smoking history. The following is a summary of a linear regression analysis using robust standard errors of the following variables:

- *alb*: serum albumin measured in g/l
- *lnalb*: Natural logarithm of *alb*
- *height*: height measured in cm
- *anysmoke*: any smoking history (0=no; 1=yes)

Summary for variables: alb height  
by categories of: anysmoke

anysmoke	N	mean	sd	min	p25	p50	p75	max
0	319	3.972414	.2733282	3.2	3.8	4	4.1	5
	321	163.7467	9.366219	141	157	162	170.5	189.5
1	414	3.961353	.3017369	3	3.8	3.9	4.2	5.1
	414	167.3519	9.69052	139	160	168	174.2	190.5
Total	733	3.966166	.2895751	3	3.8	3.9	4.2	5.1
	735	165.7774	9.710078	139	158	165.9	173.5	190.5

Linear regression

Number of obs = 733  
F( 1, 731) = 0.27  
Prob > F = 0.6039  
R-squared = 0.0004  
Root MSE = .28972

	alb	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
anysmoke		-.0110611	.0213093	-0.52	0.604	-.0528958	.0307735
_cons		3.972414	.0153003	259.63	0.000	3.942376	4.002452

. regress alb height, robust

Linear regression

Number of obs = 733  
F( 1, 731) = 6.24  
Prob > F = 0.0127  
R-squared = 0.0105  
Root MSE = .28825

	alb	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
height		.0030542	.0012231	2.50	0.013	.0006529	.0054555
_cons		3.459824	.2027865	17.06	0.000	3.061711	3.857937

. regress alb anysmoke height, robust

Linear regression

Number of obs = 733  
 F( 2, 730) = 3.44  
 Prob > F = 0.0324  
 R-squared = 0.0120  
 Root MSE = .28823

alb	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
anysmoke	-.0228393	.0218508	-1.05	0.296	-.0657373	.0200587
height	.0032687	.0012641	2.59	0.010	.0007869	.0057504
_cons	3.437173	.2070915	16.60	0.000	3.030607	3.843739

```
. gen lnalb = log(alb)
. regress lnalb anysmoke height, robust
```

Linear regression

Number of obs = 733  
 F( 2, 730) = 3.57  
 Prob > F = 0.0285  
 R-squared = 0.0118  
 Root MSE = .07256

lnalb	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
anysmoke	-.0062513	.0055066	-1.14	0.257	-.0170621	.0045594
height	.0008111	.0003091	2.62	0.009	.0002043	.0014178
_cons	1.244214	.0506479	24.57	0.000	1.144781	1.343647

### End of Output.

1.a) Based on the regression model including only height, what is the best estimate for the mean Albumin in subjects with a height of 170 cm and are never-smokers? What is the best estimate for subjects with a height of 170 cm and are smokers? (5 points)

1.b) Based on the regression model including both height and smoking history, what is the best estimate for the mean Albumin in subjects with a height of 150 cm and have never smoked? (5 points)

**Question 1 (cont.)**

1.c) Based on the regression model including both height and smoking history, what is the best estimate for the mean Albumin in subjects with a height of 190 cm and have ever smoked? (5 points)

1.d) Based on the regression model including both height and smoking history, what is the best estimate for the change in mean Albumin when comparing a never-smoking subject with a height of 175 cm to a never-smoking subject with a height of 176 cm? (5 points)

1.e) Based on the regression model including both height and smoking history, what is the best estimate for the change in mean Albumin when comparing a subject with a history of smoking and a height of 175 mg/l to a smoking subject with a height of 180 mg/l? Also, provide a 95% confidence interval for this estimate. (10 points)

1.f) Based on the regression model including both height and smoking history, provide an interpretation for the intercept. What scientific use would you make of this estimate? (5 points)

1.g) Based on the regression model including both height and smoking history, provide an interpretation for the height slope. What scientific use would you make of this estimate? (5 points)

**Question 1 (cont.)**

1.h) Based on the regression model including both height and smoking history, is there evidence that the slope for height is significantly different from 0? (5 points)

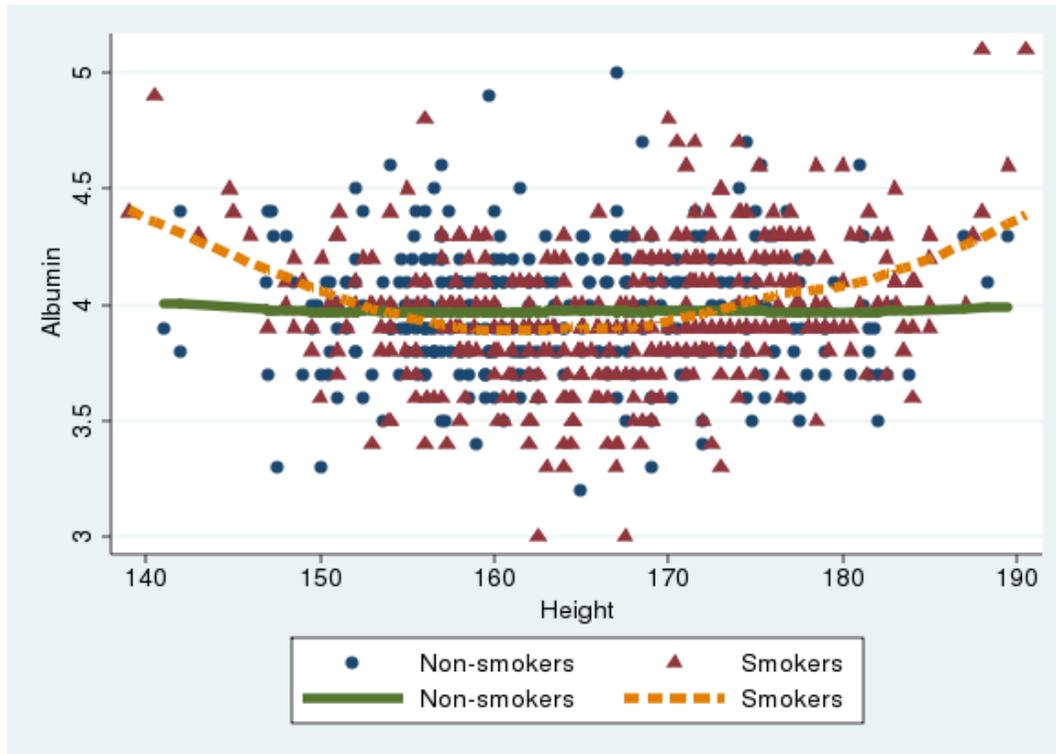
1.i) Based on the regression model including both height and smoking history, provide an interpretation for the *anysmoke* slope. What scientific use would you make of this estimate? (5 points)

1.j) Based on the regression model including both height and smoking history, is there evidence that the slope for *anysmoke* is significantly different from 0? (5 points)

1.k) Based on the regression model including both height and smoking history, what is the best estimate of the standard deviation of Albumin in groups that are homogeneous with respect to smoking history and height? (5 points)

1.l) Is there any evidence that smoking history confounds the relationship between height and Albumin? What would you have to consider for smoking to be a confounder? (5 points)

**Question 2 (20 points):** The following plot display Albumin (y-axis) by Height (x-axis) stratified by smoking history. In a post-hoc fashion, use the additional information in this plot to evaluate the analysis from question 1 by answering the following questions



2.a) From the plot, comment on the reliability of your answers to question 1, parts (b) – (c). (5 points)

2.b) From the plot, comment on the reliability of your answers to question 1, parts (d) and (e). (5 points)

**Question 2 (cont.)**

2.c) From the plot, comment on the reliability of your answers to question 1, parts (h) and (j). (5 points)

2.d) In problem 1, I also presented an analysis using log-transformed Albumin as the outcome with height and smoking history as predictors alone. Explain how this model differs in scientific interpretation from the model with (untransformed) Albumin as the outcome. Furthermore, provide full statistical inference (parameter estimates, CIs, p-values) for the *height* and *anysmoke* predictors in this regression model (15 points)

**Question 3 (30 points):** The following output represents the results from a logistic regression analysis of the risk of pre-term birth after exposure to high concentrations of disinfection byproducts (DBP) in drinking water. A case-control study was conducted in which drinking water habits were retrospectively ascertained for 50 subjects with pre-term birth (cases) and 50 subjects without preterm birth (controls). Subjects were classified as being either exposed or unexposed to high concentrations of DBP for the purpose of this analysis

We will consider the following models for  $p$ , the probability of have a pre-term birth

- $\log(p/(1-p)) = \alpha_0 + \alpha_1 \cdot \text{exposed}$
- Cases were coded as 1, controls as 0
- Exposure was coded as an indicator variable (1=exposed, 0=not exposed)

```

Logistic regression                               Number of obs   =          100
                                                  LR chi2(1)      =           9.24
                                                  Prob > chi2     =          0.0024
Log likelihood = -64.694635                    Pseudo R2       =          0.0667

```

_____	_____	_____	_____	_____	_____	_____
_outcome	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
exposed	1.252763	.4225771	2.96	0.003	.424527	2.080999
_cons	-.5596158	.280306	-2.00	0.046	-1.109005	-.0102262

3.a) Based on the logistic regression model, what scientific conclusions can you make about the association between exposure to DBP and pre-term birth? If appropriate, include estimates of the effect size and statistical inference in your written description of the results. (10 points)

3.b) Provide an interpretation of the intercept. What scientific use can you make of this quantity? (5 points)

**Question 3 (cont).**

3.c) Use the the regression output and the given marginal totals to fill in the missing cells of the following 2x2 contingency table. (15 points)

	Case	Control	Total
Exposed			45
Unexposed			55
Total	50	50	100

Question 4 (50 points): A scientific colleague was examining how the relationship between creatinine and age differed across the sexes. Ideally, I would want output from a linear regression of creatinine including terms for age (measured in years), an indicator of male sex (variable  $\text{male}=0$  for females,  $\text{male}=1$  for males), and a variable  $\text{maleage} = \text{male} * \text{age}$ . He brought to me the following output from two linear regressions of CRP on age.

-----  
-> male = 0

Linear regression

Number of obs = 367  
F( 1, 365) = 0.12  
Prob > F = 0.7289  
R-squared = 0.0003  
Root MSE = 2610.7

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
crt100						
age	9.089228	26.20464	0.35	0.729	-42.4418	60.62025
_cons	8620.363	1946.583	4.43	0.000	4792.438	12448.29

-----  
-> male = 1

Linear regression

Number of obs = 366  
F( 1, 364) = 5.01  
Prob > F = 0.0258  
R-squared = 0.0284  
Root MSE = 2784.6

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
crt100						
age	84.21487	37.61646	2.24	0.026	10.242	158.1877
_cons	5693.233	2745.619	2.07	0.039	293.9654	11092.5

4.a) Suppose the researcher fit the desired model with *male*, *age*, and *maleage*. What would have been the estimated intercept in that model? (5 points)

4.b) Suppose the researcher fit the desired model with *male*, *age*, and *maleage*. What would have been the estimated slope for *age*? (5 points)

4.c) Suppose the researcher fit the desired model with *male*, *age*, and *maleage*. What would have been the estimated slope for *male*? (5 points)

4.d) Suppose the researcher fit the desired model with *male*, *age*, and *maleage*. What would have been the estimated slope for *maleage*? (5 points)

4.e) Is there a statistically significant difference between the *age* slope for females and the *age* slope for males? Calculate the Z-statistic for this test, and indicate if the p-value is greater or less than 0.05. You may assume the parameter estimates from both models are independent and follow an approximately Normal distribution. [Hint: Recall that for two independent random variables X and Y,  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$  and  $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$ .  $Z_{1-\alpha/2} = Z_{0.975} = 1.96$ ] (10 points)

4.f) Suppose that we really wanted to know the creatinine and age in the entire population, irrespective of gender. If we assume that 50% of the sample is male (and 50% female), create an estimate of the *age* slope ignoring gender. Would this *age* slope likely have been significantly different from 0? [Hint: Recall that if we multiply a statistic by a constant, then we multiply the variance by a factor of the constant square. For a constant  $k$  and random variable X,  $\text{Var}(k * X) = k^2 * \text{Var}(x)$ ] (10 points)

4.g) Is there any evidence that gender confounds the association between *age* and creatinine? If yes, what is that evidence. If not, what additional information would you need to evaluate symptoms of confounding in this data? (10 points)