Name: Solution

uniq name: _____

BIOSTAT 651
Applied Statistics II: Extensions of Linear Regression

Test #2
Monday, March 21, 2016
1:10-2:30 p.m.

Question	Points Possible	Points Received
1	14	_____
2	14	_____
3	22	_____
Total	50	

1. (14 points, total) Azidothymidine (AZT) is an antiretroviral medication used to prevent and treat HIV/AIDS. The table below displays data from a study on the effect of AZT in slowing the development of AIDS symptoms. In the study, 338 veterans whose immune systems were beginning to falter after infection with the HIV virus were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. The table cross-classified the veteran's race, whether they received AZT immediately (AZT=use), and whether they developed AIDS symptoms during the 3-year study.

Race	AZT Use	Symptoms=YES	Symptoms=NO
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

Researchers considered the following logistic regression model:

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} * X_{2i}$$

where X_{1i} is a dummy variable for race ($X_{1i} = 1$ for white, $X_{1i} = 0$ for black) and X_{2i} is a dummy variable for AZT use ($X_{2i} = 1$ for immediate AZT use and $X_{1i} = 0$ otherwise). Note that $\pi_i = P(Y_i = 1|X)$ with $Y_i = 1$ if the patient showed AIDS symptoms.

(a) (5 points) Estimate β_0, β_2 and β_3.

$\widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3 = log(14/93) = -1.89$

$\widehat{\beta}_0 + \widehat{\beta}_1 = log(32/81) = -0.93$

$\widehat{\beta}_0 + \widehat{\beta}_2 = log(11/52) = -1.55$

$\widehat{\beta}_0 = log(12/43) = -1.27$

Therefore,

$\widehat{\beta}_0 = -1.28$

$\widehat{\beta}_2 = -0.28$

$\widehat{\beta}_3 = -0.69$

(b) (4 points) Provide interpretations for $\exp(\widehat{\beta_2})$ and $\exp(\widehat{\beta_3})$

$\exp(\widehat{\beta_2}) = 0.76$

$\exp(\widehat{\beta_3}) = 0.50$

$\exp(\widehat{\beta_2})$: Odds ratio of developing AIDS symptoms between immediate AZT uses vs. no immediate AZT use when the race is black.

$\exp(\widehat{\beta_3})$: Odds ratio of developing AIDS symptoms between immediate AZT uses vs. no immediate AZT use for white patients is estimated 50% lower that that for black patients.

(c) (5 points) Researchers are interested in estimating a relative risk (RR) of developing AIDS symptoms between $X_{2i} = 1$ vs $X_{2i} = 0$ among white patients. Is it possible to estimate the RR (Yes/No)? If Yes, please estimate the RR. If No, please justify your answer.

Yes, because it is a randomized trial.

$P(Y_i = 1|X_{2i} = 1, X_{1i} = 1) = 14/(14 + 93) = 0.13$

$P(Y_i = 1|X_{2i} = 0, X_{1i} = 1) = 32/(32 + 81) = 0.28$

$RR = 0.13/0.28 = 0.46$

2. (14 points, total) In the data set of interest, pest insects were exposed to various levels of gaseous carbon disulphide (CS_2, recorded on the $\log_{10}$ scale) for 5 hours. The observed data are given by:

j	$log_{10}CS_2$ (X_j)	n_j	$Death$ $(Y=1)$
1	1.5	59	6
2	1.6	60	13
3	1.7	62	20
4	1.8	56	38
5	1.9	63	52
6	2.0	59	53
7	2.1	62	61

We use the following regression model with the probit link function to model the relationship between mortality and CS_2:

$$\Phi^{-1}(\pi_j) = \beta_0 + \beta_1 X_j, \tag{1}$$

where $\pi_j = P(Y_{ij} = 1|X_j)$ and $\Phi(\cdot)$ is a normal cumulative distribution function. SAS output is provided at the back of the exam.

(a) (4 points) LD80 is the dose required to kill 80% of insects in the tested population during a pre-specified period. Estimate the LD80. You need to use the following standard normal table.

$\Phi(z)$	0.1	0.25	0.5	0.8	0.9
z	-1.28	-0.67	0	0.84	1.28

LD80 is the $log_{10}CS_2$ level to kill 80% of insects. By the definition of LD80, $\Phi^{-1}(0.8) = \beta_0 + \beta_1 LD80$. From this equation,

$$LD80 = \frac{\Phi^{-1}(0.8) - \widehat{\beta}_0}{\widehat{\beta}_1} = \frac{0.84 + 9.75}{5.59} = 1.89$$

(b) (5 points) Calculate a 95 % confidence interval for the LD80.

Suppose $g(\beta) = \frac{0.84-\beta_0}{\beta_1}$. And then

$$g(\widehat{\beta}) = LD80 \tag{2}$$

$$\frac{\partial g(\widehat{\beta})}{\partial \beta} = (\frac{-1}{\widehat{\beta}_1}, \frac{\widehat{\beta}_0 - 0.84}{\widehat{\beta}_1^2})^T = (-0.179, -0.339)^T \tag{3}$$

$$\tag{4}$$

By the Delta method,

$$\widehat{Var}(LD80) = \frac{\partial g(\widehat{\beta})}{\partial \beta}^T \widehat{Var}(\widehat{\beta}) \frac{\partial g(\widehat{\beta})}{\partial \beta} = 0.00022$$

95 % CI

$$LD80 \pm 1.96\sqrt{0.00022} = (1.865, 1.923)$$

(c) (5 points) Now we change the endpoint as alive, $I(Y_{ij} = 0)$, and use the following model

$$\Phi^{-1}(\pi_j) = \beta_0^* + \beta_1^* X_j,$$

where $\pi_j^* = P(Y_{ij} = 0|X_j)$. Is it possible to estimate β_0^* and β_1^* using the existing SAS output for the model (1)? If Yes, please estimate β_0^* and β_1^*. If No, please justify your answer.

Yes, because the probit link function is symmetric.

$\widehat{\beta}_0^* = -\widehat{\beta}_0 = 9.75$

$\widehat{\beta}_1^* = -\widehat{\beta}_1 = -5.59$

3. (22 points, total) Researchers at the University of Michigan carried out a case-control study to evaluate the effect of a genetic mutation in the TCF7L2 gene (SNP rs7903146) to the type 2 diabetes (T2D). Covariate information was assembled on 2000 cases and 2000 controls, including:

   AGE = age of the study subject at the enrollment
   $Gender = 1$ for female (0 otherwise)
   $SNP = 1$ for the presence of the mutation (0 otherwise)

   The response variable was coded as $Y = 1$ for cases (T2D patients) and $Y = 0$ for controls. The following model was fitted:

$$\begin{aligned} \log\left\{\frac{\pi_i}{1-\pi_i}\right\} &= \beta_0 + \beta_1 AGE_i + \beta_2 Gender_i + \beta_3 SNP_i \\ &+ \beta_4 AGE_i \times SNP_i \end{aligned} \tag{5}$$

   where $\pi_i = P(Y_i = 1|X_i)$. SAS output is provided at the back of the exam.

(a) (4 points) Researchers want to know whether the logistic regression model (5) well fits the data. Using the SAS output, perform the Hosmer−Lemeshow test. You need to write down the hypothesis, test statistic, reference null distribution and conclusion. (NOTE: the test statistic is given in the SAS output)

H0: the logistic regression model fits the data well

H1: the logistic regression model does not fit the data well

Test statistic

$$H = 8.75$$

which follows χ^2_8 under H0.

Since $H < 15.51 = \chi^2_{8,0.95}$, we can not reject H0. We can conclude that the logistic regression model well fits the data.

(b) (4 points) Provide interpretations for $\widehat{\beta_3}$ and $\widehat{\beta_4}$

$\widehat{\beta_3} = 0.694$

$\widehat{\beta_4} = -0.0081$

$\widehat{\beta}_3$: log odds ratio of T2D comparing individuals with and without the mutation when Age=0, adjusting for all other covariates.

$\widehat{\beta}_4$: log odds ratio of T2D for a one year increase in age for individuals with the mutation is expected to 0.0081 lower than that for individuals without the mutation, adjusting for other covariates.

(c) (4 points) Does the mutation in the SNP has a significant effect on T2D? Please carry out a likelihood ratio test for $\beta_3 = \beta_4 = 0$. You need to write down full and reduced models, test statistic, reference null distribution and conclusion.

$H_0 : \beta_3 = \beta_4 = 0$

Full Model:

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_0 + \beta_1 AGE_i + \beta_2 Gender_i + \beta_3 SNP_i + \beta_4 AGE_i \times SNP_i$$

Reduced Model:

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_0 + \beta_1 AGE_i + \beta_2 Gender_i$$

Test statistic: $X_L = 2(l_{full} - l_{reduced}) = 5263 - 5243 = 20$ which follows χ^2_2 under H0.

Since $X_L > 5.991 = \chi^2_{2,0.95}$, we can reject H0. We can conclude that the SNP has a significant effect on T2D.

(d) (5 points) To improve the interpretation of parameters, researchers decided to use the following model:

$$\begin{aligned}\log\left\{\frac{\pi_i}{1-\pi_i}\right\} &= \beta_0^* + \beta_1^*(AGE_i - 40) + \beta_2^* Gender_i + \beta_3^* SNP_i \\ &+ \beta_4^*(AGE_i - 40) \times SNP_i\end{aligned}$$

Estimate β_1^*, β_2^*, β_3^* and β_4^*.

$$\begin{aligned}\log\left\{\frac{\pi_i}{1-\pi_i}\right\} &= \beta_0 + \beta_1 AGE_i + \beta_2 Gender_i + \beta_3 SNP_i \\ &+ \beta_4 AGE_i \times SNP_i \\ &= (\beta_0 + 40\beta_1) + \beta_1(AGE_i - 40) + \beta_2 Gender_i + (\beta_3 + 40\beta_4) SNP_i \\ &+ \beta_4 (AGE_i - 40) \times SNP_i \end{aligned} \tag{6}$$

$\widehat{\beta}_1^* = \widehat{\beta}_1 = 0.0526$

$\widehat{\beta}_2^* = \widehat{\beta}_2 = -0.7021$

$\widehat{\beta}_3^* = \widehat{\beta}_3 + 40\widehat{\beta}_4 = 0.6940 - 0.0081 * 40 = 0.37$

$\widehat{\beta}_4^* = \widehat{\beta}_4 = -0.0081$

(e) (5 points) It is known that the prevalence of T2D is 0.1 ($P(Y = 1) = 0.1$). Using this fact, estimate a risk of T2D for a 50 year old female without the mutation in rs7903146 (i.e. SNP = 0)

Since it is a case-control study, you need to calculate sampling factions to consistently estimate the intercept β_0

Sampling fraction:

$\tau_1 = P(S = 1|Y = 1) = P(Y = 1|S = 1)P(S = 1)/P(Y = 1)$

$\tau_0 = P(S = 1|Y = 0) = P(Y = 0|S = 1)P(S = 1)/P(Y = 0)$

$$\frac{\tau_1}{\tau_0} = \frac{P(Y = 1|S = 1)P(Y = 0)}{P(Y = 0|S = 1)P(Y = 1)} = 0.5 * 0.90.5 * 0.1 = 9$$

Bias adjusted β_0: $-2.495 - log(9) = -4.692$

Risk

$$P(Y_i = 1|Age = 50, Gender = 1, SNP = 0)$$
$$= exp(-4.692 + 50 * 0.0526 - 0.7021)/(1 + exp(-4.692 + 50 * 0.0526 - 0.7021)) = 0.059$$