Midterm II Statistics 500 (Winter 2014)

Name: Key

Instructions: Answer the following questions carefully. You must **show all work** where needed to receive credit. Partial credit will be given where work is shown. Space is provided for your answers. You may use scrap paper to arrive at solutions, but **show all work** on the test. Most calculators use LN for the natural logarithm.
Total points: 65

1) The model

$$y_i \mid p_i \sim \text{Binomial}(n_i, p_i) \qquad y_i \mid p_i \text{ independent} \qquad i = 1, \ldots, N$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i$$

was fit with the following results.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3325     0.2061  -6.467    1e-10 ***
x             0.4747     0.2226   2.133    0.033 *
---
    Null deviance: 17.566  on 9  degrees of freedom
Residual deviance: 12.854  on 8  degrees of freedom
```

a) Test the hypothesis $H_0 : \beta = 0$ using the Wald test. Must state the value of the test statistic, its approximate distribution when H_0 is true, and conclusion. (3 points)

$Z_0 = \frac{\hat{\beta}}{se(\hat{\beta})} \sim N(0,1)$ H_0 true $Z_0 = 2.133 > 1.96$ Reject H_0

or $Z_0^2 \sim \chi^2_1$ $Z_0^2 = (2.133)^2 = 4.55 > 3.84$ Reject H_0

b) Test the hypothesis $H_0 : \beta = 0$ using the likelihood ratio test. Must state the the value of the test statistic, its approximate distribution when H_0 is true, and conclusion. (4 points)

$G^2 =$ Null deviance $-$ Residual deviance $= 4.71$ $G^2 \sim \chi^2_1$ H_0 true

$G^2 = 4.71 > 3.84$ Reject H_0

c) Calculate a 95% confidence interval for 2β. (4 points)

$2\hat{\beta} \pm 1.96\, se(2\hat{\beta})$

$2(0.4747) \pm 1.96\,(2)(0.2226) = (0.077, 1.82)$

$.9494 \pm .8726$

2) Suppose response variables y_i are independent and belong to the exponential family with $E[y_i \mid \mu_i] = \mu_i$. Consider the generalized linear model with link function h

$$h(\mu_i) = \alpha + \beta x_i \qquad i = 1, \ldots, N$$

where α is the intercept and β is the slope parameter. The model was fit by the method of maximum likelihood with results shown below.

```
Coefficients:
             Estimate Std Error  z-value  p-value
(Intercept)    -2.00      1.00    -2.00    0.046 *
    x           4.00      2.00     2.00    0.046 *

Covariance matrix:
             (Intercept)     x
(Intercept)        1.00     1.00
    x              1.00     4.00
```

Let $g(\hat{\alpha}, \hat{\beta}) = \dfrac{\hat{\alpha}\hat{\beta}}{2}$ and estimate a large sample standard error for $g(\hat{\alpha}, \hat{\beta})$. (6 points)

By delta method $Var(g(\hat{\alpha},\hat{\beta})) \approx \hat{B}\hat{\Sigma}\hat{B}'$

$g(\alpha,\beta) = \frac{\alpha\beta}{2}$ $B = \left[\frac{\partial g}{\partial \alpha}, \frac{\partial g}{\partial \beta}\right] = \left[\frac{\beta}{2}, \frac{\alpha}{2}\right]$

$\frac{\partial g}{\partial \alpha} = \frac{\beta}{2}$ $\frac{\partial g}{\partial \beta} = \frac{\alpha}{2}$ $Var\left(\frac{\hat{\alpha}\hat{\beta}}{2}\right) \approx \begin{bmatrix} 2 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}\begin{bmatrix} 2 \\ -1 \end{bmatrix}$

$= \begin{bmatrix} 1 & -2 \end{bmatrix}\begin{bmatrix} 2 \\ -1 \end{bmatrix} = 4$

$se\left(\frac{\hat{\alpha}\hat{\beta}}{2}\right) \approx 2$

3) A data analyst fit the standard Poisson model for rates

$$y_i \mid \theta_i \sim \text{Poisson}(t_i\theta_i) \qquad y_i \mid \theta_i \text{ independent} \qquad i = 1, \ldots, N$$

$$\log \theta_i = \alpha + \beta x_i$$

with the following results.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.7302     0.1148 -67.363  < 2e-16 ***
 x            0.5639     0.1620   3.481 0.000499 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 28.699  on 9  degrees of freedom
Residual deviance: 15.485  on 8  degrees of freedom
```

$1.34(0.1148) = 0.1538$

$1.34(0.162) = 0.2171$

$\frac{-7.7302}{0.1538} = -50.26$

Noting some overdispersion, the analyst decided to adjust the standard errors and estimate the dispersion parameter as

$$\hat{\sigma}^2 = \frac{X^2}{N - p}$$

$\frac{0.5639}{0.2171} = 2.597$

where X^2 is the Pearson chi-squared statistic. Fill in the `Std. Error` and `z value` columns below. **Show calculations.** (4 points)

$\hat{\sigma}^2 = 1.7956$

$\hat{\sigma} = \sqrt{1.7956} = 1.34$

```
Coefficients:
             Estimate Std. Error z value
(Intercept)  -7.7302   0.1538   -50.26
 x            0.5639   0.2171   2.597
---
(Dispersion parameter for poisson family taken to be 1.7956)
```

4) Electronic stability control (ESC) is a technology designed to prevent motor vehicle loss of control. When ESC sensors detect impending loss of control, braking at individual wheels and engine torque are automatically adjusted in an attempt to keep the vehicle under control and on the road. The National Highway Traffic Safety Administration (NHTSA) sponsored a study to assess the effects of ESC on the likelihood of loss-of-control type crashes for passenger cars. Data were collected on the following 9 variables.

loc	0=vehicle struck and not at fault in crash with no loss of control, 1=vehicle in single-vehicle crash with loss of control (ran off road)
esc	1=second generation ESC, 2=first generation ESC, 3=no ESC device
scond	road surface condition, 0=dry, 1=wet
splim	speed limit (mph), 0=less than 55, 1=55 or greater
ralign	road alignment, 0=straight, 1=curved
light	light condition, 0=light, 1=dark
age	driver age in years
gender	driver gender, 0=male, 1=female
vtype	vehicle type, 0=luxury car, 1=non-luxury sedan or compact

A logistic regression model was fit with `loc` as the binary response and all other variables as predictors. The variable `esc` is a categorical predictor with three levels. `age` is a continuous variable. All other predictors are binary. Use $\alpha = 0.05$ as the level of significance and in constructing confidence intervals.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.493555   0.267927  -1.842   0.0655 .
esc2         0.337973   0.177493   1.904   0.0569 .
esc3         1.215587   0.175485   6.927 4.30e-12 ***
scond        1.515469   0.173410   8.739  < 2e-16 ***
splim        1.006600   0.136156   7.393 1.44e-13 ***
ralign       1.071445   0.144495   7.415 1.22e-13 ***
light        0.935101   0.136960   6.828 8.64e-12 ***
age         -0.055825   0.005524 -10.106  < 2e-16 ***
gender       0.047211   0.132629   0.356   0.7219
vtype        0.194378   0.136589   1.423   0.1547
---
    Null deviance: 1757.7  on 1281  degrees of freedom
Residual deviance: 1364.3  on 1272  degrees of freedom

Partial Covariance Matrix:
            esc2    esc3   scond    splim   ralign   light
esc2     0.03150 0.01937 0.00030 -0.00010 -0.00007 0.00124
esc3     0.01937 0.03080 0.00242  0.00268  0.00174 0.00244
scond    0.00030 0.00242 0.03007  0.00149  0.00300 0.00281
splim   -0.00010 0.00268 0.00149  0.01854  0.00159 0.00210
ralign  -0.00007 0.00174 0.00300  0.00159  0.02088 0.00192
light    0.00124 0.00244 0.00281  0.00210  0.00192 0.01876
```

a) Can the **Residual deviance** be used as a Goodness-of-Fit statistic? Why or why not? (2 points)

No. The response is binary and $n_i = 1$.

Large sample theory does not apply.

b) Holding other variables in the model fixed, what is the estimated odds ratio of being in a single-vehicle crash with loss of control comparing a vehicle equipped with no ESC device to a vehicle with a second generation ESC device? Calculate a 95% confidence interval for the odds ratio. Is the result significant? Explain. (3 points)

$$\hat{OR} = e^{1.2156} = 3.37$$

$$1.2156 \pm 1.96(0.1755) = (0.8716, 1.5596)$$

$$e^{(0.8716, 1.5596)} = (2.39, 4.76)$$

significant, 1 is not in interval

c) Holding other variables in the model fixed, what is the estimated odds ratio of being in a single-vehicle crash with loss of control for a 10 year **decrease** in age? (Note the sign of the age coefficient. So, for example, the problem is asking for the odds ratio comparing a 30-year old driver to a 40 year-old driver). Calculate a 95% confidence interval for the odds ratio. (4 points)

$$-10\hat{\beta}_{age} = -10(-0.0558) = 0.558$$

$$\hat{OR} = e^{0.558} = 1.75$$

$$-10\hat{\beta} \pm 1.96\, se(10\hat{\beta})$$

$$0.558 \pm 1.96(10)\,0.0055 = 0.558 \pm 0.1078$$

$$= 0.4502, 0.6658$$

$$e^{0.4502, 0.6658} = (1.57, 1.95)$$

d) Using results from the fitted model, construct a 95% confidence interval for $\beta_{esc3} - \beta_{esc2}$. **Interpret the result.** (5 points)

$$\hat{\beta}_{esc3} - \hat{\beta}_{esc2} \pm 1.96\, se(\hat{\beta}_{esc3} - \hat{\beta}_{esc2})$$

$$Var(\hat{\beta}_{esc3} - \hat{\beta}_{esc2}) = Var(\hat{\beta}_{esc3}) + Var(\hat{\beta}_{esc2}) - 2Cov(\hat{\beta}_{esc3}, \hat{\beta}_{esc2})$$

$$= 0.0308 + 0.0315 - 2(0.01937)$$

$$= 0.02356 \qquad \sqrt{.02356} = 0.1535$$

$$1.2156 - 0.338 \pm 1.96(.1535)$$

$$0.8776 \pm 1.96(0.1535) = (0.577, 1.178)$$

The interval does not include 0. There is a significant difference in the odds of loss of control comparing ~~2nd gen~~ no ESC to 1st gen. No ESC has higher estimated odds.

e) Suppose you want to test the hypothesis $H_0 : \beta_{sex2} = \beta_{sex3}$ using a Wald test. From large sample theory of MLEs, $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1})$. Create the $\boldsymbol{A}$ matrix of constants so that $\boldsymbol{A}\boldsymbol{\beta}$ satisfies H_0 noting the order in which the parameters were fit in the model. Write down the equation for the Wald statistic when H_0 is true. What is the large sample distribution of the Wald statistic when H_0 is true? (5 points)

$$A_{1\times 10} = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A\hat{\beta} \sim N\left(A\beta,\ A(x'wx)^{-1}A'\right)$$

$$w = (A\hat{\beta})'\left[A(x'wx)^{-1}A'\right]^{-1}(A\hat{\beta}) \sim \chi^2_1$$

when Ho true $A\beta = 0$

$$w_{1\times 1} = \frac{(A\hat{\beta})^2}{A(x'wx)^{-1}A'}$$

5) The proportional odds model for an ordered categorical response, as given below, was fit. The variables x_1, x_2, and x_3 are continuous predictors.

$$y_{i1}, y_{i2}, y_{i3}, y_{i4} \sim \text{Multinomial}(n_i = 1; \pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4}) \qquad i = 1, \ldots, N \qquad j = 1, 2, 3$$

$$\log\left(\frac{P(Y \le j \mid \boldsymbol{x}_i)}{1 - P(Y \le j \mid \boldsymbol{x}_i)}\right) = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

The proportional odds assumption was tested using a likelihood ratio statistic. The test statistic has a large sample chi-squared distribution on the appropriate number of degrees of freedom. What are the degrees of freedom? **Explain carefully. Must demonstrate method.** (4 points)

The full model has 12 parameters

$\alpha_j, \beta_{1j}, \beta_{2j}, \beta_{3j} \quad j = 1, 2, 3$

The reduced model has 6 parameters

$\alpha_j \quad \beta_1, \beta_2, \beta_3 \quad j = 1, 2, 3$

$12 - 6 = 6$

6) Consider the weighted linear model

$$y_i = \beta x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2/w_i) \qquad i = 1, \ldots, n$$

where β is the unknown regression parameter, x is the predictor variable, ϵ_i are independent, and w_i are known weights. Let

$$\epsilon' W \epsilon = \sum_{i=1}^{n} w_i (y_i - \beta x_i)^2$$

be the least squares criterion and find the least squares estimator for β. Be sure to show all work. (6 points)

$$\frac{\partial}{\partial \beta} \epsilon' w \epsilon = \frac{\partial}{\partial \beta} \sum w_i (y_i - \beta x_i)^2$$

$$= -2 \sum w_i (y_i - \beta x_i) x_i = 0$$

$$\Rightarrow \sum (w_i x_i y_i - \beta w_i x_i^2) = 0$$

$$\sum w_i x_i y_i = \beta \sum w_i x_i^2$$

$$\Rightarrow \hat{\beta} = \frac{\sum w_i x_i y_i}{\sum w_i x_i^2}$$

Problems 7 - 16 are multiple choice questions. Circle the letter corresponding to the **best** answer. (1.5 points each)

7) When running a Markov chain, the practice of discarding initial values is known as

a) burn-in
b) delay
c) thinning
d) chain stability
e) none of the above

8) When using a conjugate prior distribution, the posterior distribution

a) comes from the same family as the likelihood distribution
b) comes from the same family as the prior distribution
c) comes from a family intermediate between the likelihood and the prior distributions
d) comes from a noninformative distribution
e) none of the above

9) A general characteristic of a Bayesian estimator is that

a) it tends to the maximum likelihood estimator when the sample size is small
b) it is a weighted average of the data and the prior mean
c) it tends to the prior mean when the sample size is large
d) it is asymptotically normal for small to moderate samples
e) none of the above

10) In a bioassay experiment, the LD50 is

a) the dose level equal to the mean probability of death
b) the mean dose level
c) the dose level at which the probability of death is 1/2
d) the median dose level
e) none of the above

11) The bootstrap is a resampling plan, most commonly used to

a) estimate the mean of a sample statistic
b) estimate the median of a sample statistic
c) estimate the standard error of a sample statistic
d) approximate the sampling distribution of a sample statistic
e) c or d

12) When fitting a logistic regression model with an explanatory variable x, one way to check if a linear trend is appropriate is to make a plot of x versus

a) the sample proportions
b) the sample logits
c) the log of the sample proportions
d) the log of x
e) none of the above

13) The Residual Deviance is known as a

a) Likelihood Ratio statistic
b) Wald statistic
c) Pearson chi-squared statistic
d) Score statistic
e) none of the above

accepting a or e

14) The Central Limit Theorem applies to sums of independent random variables when

a) sample sizes are small to moderate
b) the population is not normal and the sample size is large
c) the population is normal
d) the population is normal and the sample size is large
e) none of the above

15) An *offset* is declared when fitting

a) a logit model for the analysis of counts
b) a logit model for the analysis of proportions
c) a loglinear model for the analysis of counts
d) a loglinear model for the analysis of rates
e) none of the above

16) The statistic G^2 has an approximate chi-squared distribution when

a) fitting the logistic regression model with a binary response
b) fitting the logistic regression model with a binomial response
c) fitting the log-linear model model with a Poisson response
d) the sample size is small
e) b or c