Name: Solution

uniq name: _____

BIOSTAT 651
APPLIED STATISTICS II: EXTENSIONS OF LINEAR REGRESSION

Test #1

Wednesday, February 17, 2016
1:10-2:30 p.m.

Statistical table and blank paper are provided.

Question	Points Possible	Points Received
1	13	_____
2	12	_____
3	25	_____
Total	50	

1. (13 points, total) The Pareto distribution with a known scaling parameter $\alpha > 0$ is given

$$f(Z = z; \beta) = \frac{\beta \alpha^{\beta}}{z^{(\beta+1)}}, \quad z > \alpha, \beta > 0.$$

(a) (4 points) Write out $P(Z = z; \beta)$ as an exponential family form (Note that $a(\phi) > 0$).

$$f(z=z;\beta) = \exp\{-(\beta+1)\log z + \beta \log \alpha + \log \beta\}$$

$$= \exp\{-\beta \log z + \beta \log \alpha + \log \beta - \log z\}$$

- $t(z) = \log z, \quad \theta = -\beta, \quad b(\theta) = -\beta \log \alpha - \log \beta.$

$a(\phi) = 1, \quad c(y, \phi) = -\log z.$

(b) (4 points) Log transformation of Y, $Y = \log(Z/\alpha)$, is commonly used as an outcome in regression analysis. The density function of Y is given as

$$f(Y = y; \beta) = \beta e^{-\beta y}, \quad y > 0, \beta > 0.$$

Suppose that we have n independent observations (X_i, Y_i). Write out the above density function as an exponential family form and determine the canonical link function $g(\mu_i) = X_i^T \beta$, where $\mu_i = E(Y_i)$. (Note that $a(\phi) > 0$)

$$f(Y=y;\beta) = \exp\{-\beta y + \log \beta\}.$$

- $t(y) = y, \quad \theta = -\beta, \quad b(\theta) = -\log \beta = -\log(-\theta)$

$a(\phi) = 1$

- $b'(\theta) = -\frac{1}{\theta} = \mu. \quad \Leftrightarrow \quad \theta = -\frac{1}{\mu}.$

- Canonical link function : $g(\mu) = -\frac{1}{\mu}.$

(c) (5 points) Determine the deviance. It should be expressed as a function of Y_i and $\widehat{\mu}_i$ $(i = 1, \ldots, n)$, where $\widehat{\mu}_i$ is the estimated μ_i under the model in (b).

- Saturated model

$$\tilde{\theta}_i = -\frac{1}{\tilde{\mu}_i} = -\frac{1}{Y_i}, \quad b(\tilde{\theta}_i) = -\log\left(\frac{1}{Y_i}\right)$$

- Fitted model

$$\hat{\theta}_i = -\frac{1}{\hat{\mu}_i}, \quad b(\hat{\theta}_i) = -\log\left(\frac{1}{\hat{\mu}_i}\right)$$

$$D = 2\sum_{i=1}^{n}\left[Y_i(\tilde{\theta}_i - \hat{\theta}_i) - \left(b(\tilde{\theta}_i) - b(\hat{\theta}_i)\right)\right] = 2\sum_{i=1}^{n} Y_i\left[-\frac{1}{Y_i} + \frac{1}{\hat{\mu}_i}\right] - \left\{-\log\frac{1}{Y_i} + \log\frac{1}{\hat{\mu}_i}\right\}$$

$$= 2\sum\left[\frac{Y_i}{\hat{\mu}_i} - \log\frac{Y_i}{\hat{\mu}_i} - 1\right]$$

2. (12 points, total) Suppose that the response Y_i follows a Poisson(λ_i) distribution

$$f(Y_i|\lambda_i) = \frac{\lambda_i^{Y_i} e^{-\lambda_i}}{Y_i!}.$$

Assume that the only covariate, x_i, acts additively on λ_i, i.e., $\lambda_i = x_i\beta$.

(a) (4 points) Derive score function, $U(\beta)$, and expected Fisher information, $I(\beta)$, as a function of Y_i, x_i and β.

$$L(L_i) = \frac{\lambda_i^{Y_i} e^{-\lambda_i}}{Y_i!}$$

$$\ell_i = Y_i \log\lambda_i - \lambda_i$$

$$U_i(\beta) = \frac{\partial \ell_i}{\partial \beta} = \frac{\partial \ell_i}{\partial \lambda_i} \cdot \frac{\partial \lambda_i}{\partial \beta} = \left(\frac{Y_i}{\lambda_i} - 1\right) \cdot x_i = \left(\frac{Y_i}{x_i\beta} - 1\right) \cdot x_i = \frac{Y_i}{\beta} - x_i$$

$$J_i(\beta) = -\frac{\partial U_i(\beta)}{\partial \beta} = \frac{Y_i}{\beta^2}, \qquad I_i(\beta) = E[J_i(\beta)] = \frac{x_i\beta}{\beta^2} = \frac{x_i}{\beta},$$

$$U(\beta) = \sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n}\left(\frac{Y_i}{\beta} - x_i\right).$$

$$I(\beta) = \sum_{i=1}^{n} I_i(\beta) = \frac{\sum x_i}{\beta}$$

(b) (4 points) Suppose now that the following data are observed:

x_i	1	2	3	4	5
Y_i	2	4	6	8	10

Compute the maximum likelihood estimate of β and the corresponding 95% confidence interval.

- $\sum y_i = 30 \quad \sum x_i = 15$

$0 = U(\hat{\beta}) = \frac{30}{\hat{\beta}} - 15 \Rightarrow \hat{\beta} = 2.$

$I(\hat{\beta}) = \frac{15}{2} = 7.5$

- 95% CI : $\hat{\beta} \pm 1.96\sqrt{I(\hat{\beta})^{-1}} = 2 \pm 1.96\sqrt{1/7.5}$

$= (1.28, 2.72)$

(c) (4 points) Carry out a score test for $H_0 : \beta = 2.5$ versus $H_1 : \beta \neq 2.5$.

$X_S^2 = U(2.5)\, I^{-1}(2.5)\, U(2.5)$

$U(2.5) = \frac{30}{2.5} - 15 = -3$

$I(2.5) = 6$

$\therefore X_S^2 = \frac{9}{6} = 1.5 < \chi^2_{1,0.05} = 3.84$

We cannot reject H_0 at level 0.05.

3. (23 points, total) Researchers are interested in whether a new anti-viral drug can decrease the number of viral RNA copies. 30 patients were randomly assigned to drug ($X_1 = 1$) and placebo groups ($X_1 = 0$), and the viral RNA count per mL (Y) was measured using RNA-sequencing. In addition, age in years (X_2) was measured as a covariate. The following regression model has been proposed for this analysis

$$log(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

where Y_i follows the poisson distribution

$$f(Y_i|\lambda_i) = \frac{\lambda_i^{Y_i} e^{-\lambda_i}}{Y_i!}.$$

The estimate of the inverse of the Fisher information matrix is

$$I(\widehat{\beta})^{-1} = \begin{pmatrix} 0.0008553 & -0.000054 & -0.000035 \\ -0.000054 & 0.0002458 & -3.745 \times 10^{-6} \\ -0.000035 & -3.745 \times 10^{-6} & 1.7028 \times 10^{-6} \end{pmatrix}$$

Subset of the SAS outputs are provided in a separate document.

(a) (4 points) Interpret β_1 and β_2

β_1 : difference in log mean viral RNA count between drug and placebo groups, adjusting for age

β_2 : difference in log mean viral RNA count for a one-year increase in age.

(b) (4 points) Estimate $exp(\beta_1)$ and its 95% confidence interval.

$\hat{\beta}_1 = 0.1867$

$SE(\hat{\beta}_1) = \sqrt{0.0002458} = 0.0157.$

- $exp(\hat{\beta}_1) = exp(0.1867) = 1.21$
- 95% C.I. $exp(0.1867 \pm 1.96 \times 0.0157)$

$= (1.17, 1.24)$

(c) (4 points) Derive the likelihood ratio test for the drug effect.

* Write out full and reduced models.

full model : $\log \lambda_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

reduced model : $\log \lambda_i = \beta_0 + \beta_2 X_2$.

* Calculate the likelihood ratio test (LRT) statistic using SAS outputs.

$LRT = D_0 - D_1 = 158.83 - 16.34 = 142.49.$

* What is a conclusion based on the LRT?

Since $142.49 > \chi^2_{1,0.05} = 3.84$, we can reject H_0.

$\Rightarrow$ drug has a strong effect on viral RNA count.

(d) (5 points) Suppose that the first individual has $(Y_1, X_{i1}, X_{i2}) = (530, 0, 22)$ and the corresponding leverage $h = 0.3$. Obtain the standardized Pearson residual and the Cook's distance. Is this a high leverage observation (YES/NO)? or high influence observation (YES/NO)? Please justify your answer.

$\hat{\lambda}_i = \exp(6.0198 + 0.0103 \times 22) = 516.15$

- $\hat{r}^P = \dfrac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} = \dfrac{530 - 516.15}{\sqrt{516.15}} = 0.6096$
- Standardized pearson residual :

$\hat{r}^{PS} = \dfrac{0.6096}{\sqrt{1-0.3}} = 0.7286$

- Cook's $D = \dfrac{1}{3} \dfrac{0.3}{0.7} \times 0.7286^2 = 0.0758 < 1$

- high leverage point since $h = 0.3 > 2 \times \frac{9}{n} = 0.2$
- not a high influence point since Cook's $D < 1$

(e) (4 points) Carry out a Goodness of Fit test and state your conclusion.

- H_0 : model fits the data well
- test statistic : Pearson χ^2 = 16.37.
  (You can also use deviance = 16.34)
- Null dist : χ^2_{27}.

  Since $16.37 < \chi^2_{27,0.05} = 40.113$, we cannot reject H_0 at level 0.05.

(f) (4 points) Researchers want to investigate multicollinearity among covariates. Since proc genmod does not calculate the variance inflation factor (VIF), they plan to use proc reg with the following command. Is it a right way to calculate VIF in this model (YES/NO)? Please justify your answer.

```
[SAS Code]
proc reg data=ViralRNA;
model Y = X1  X2/ vif;
run;
```

No. GIM uses weighted least squares to estimate parameters, so weight should be specified.