Biometry 711 - Categorical Data Analysis
Summer 2015
Homework 2
Due Wednesday 1 July 2015, 6pm

1. Let $Y_i \sim \exp(\lambda_i)$.
   (a) Show that Y_i is a member of the scaled exponential family. Identify each of the following:
      - The canonical parameter θ_i
      - $b(\theta_i)$
      - $a(\phi)$ and the corresponding scale parameter ϕ
      - $c(y_i; \phi)$

   (b) Derive the mean of Y_i using your answer to question 1a.
   (c) Derive the variance of Y_i using your answer to question 1a.
   (d) What is the canonical link function?

2. Let X and Y be dichotomous random variables. The SAS code to compute the (marginal) $X - Y$ odds ratio using PROC FREQ is shown below.

```
proc freq data = one;
     table X*Y/or;
run;
```

Let Z be a categorical variable with q levels. In your epidemiology courses, you learned how to estimate the $X - Y$ odds ratio *adjusted for the confounding effects of variable* Z by conducting a stratified analysis whereby the $X - Y$ odds ratios for each of the q partial tables are combined to produce the Cochran-Mantel-Haenszel (CMH) adjusted odds ratio. The SAS code to conduct this analysis using PROC FREQ is shown below.

```
proc freq data = one;
     table Z*X*Y/cmh;
run;
```

   (a) Use the IMPACT data set and SAS' PROC FREQ to construct the odds ratio and corresponding 95% CI for being drug free at 12 months comparing white subjects to those of other races. Interpret this odds ratio using a complete sentence.
   (b) Using PROC FREQ, construct the CMH odds ratio and corresponding 95% CI for being drug free at 12 months comparing white subjects to those of other races, controlling for the effect of injecting drug use history. Interpret the adjusted odds ratio using a complete sentence. Based on the estimated odds ratios in questions 2a and 2b, is there evidence that injecting drug use history confounds the association between being drug free at 12 months and race? Support your answer.
   (c) The CMH option also produces the Breslow-Day test. Read about this test in the SAS help file (SAS Help and Documentation > SAS/STAT 9.3 User's Guide > The FREQ Procedure > Details: FREQ Procedure > Statistical Computations > Cochran-Mantel-Haenszel Statistics), and on page 242 of your text. Provide an explicit expression for the null and alternative hypothesis for the Breslow-Day test for the variables in question 2b, and state your conclusion.

(d) Use the `OR` option in `PROC FREQ`'s `TABLE` statement to obtain the partial odds ratios and their corresponding 95% CIs. Interpret each partial odds ratio. Explain how their values are supportive of your findings in question 2c.

(e) An equivalent approach to this stratified analysis is based on appropriately selected nested logistic regression models. Provide explicit expressions for these models, and then fit them using `PROC LOGISTIC`. Report the parameter estimates for each model and demonstrate they yield the same marginal, adjusted and partial ORs as those computed in questions 2a, 2b and 2d.

(f) Suggest two different ways to test the null hypothesis stated in question 2c based on the logistic regression models fit in question 2e. Conduct both tests. Report the p-values and state your conclusions.

(g) When reporting measures of association for being drug free at 12 months and subject's race, is it more appropriate to report the marginal, adjusted or partial odds ratios? Support your answer?

(h) Repeat questions 2a - 2g where interest now is in the association between remaining drug free at 12 months and subject's race, controlling for treatment site.

3. In class, we discussed the iterative reweighted least squares (IRWLS) procedure for fitting a GLM. For $Y_1, \ldots, Y_n \overset{\text{ind}}{\sim} \text{Poisson}(\mu_i)$ and observed values $y_1, \ldots, y_n$, provide explicit expressions for the IRWLS steps for fitting a GLM with p predictors $\mathbf{X}_1, \ldots, \mathbf{X}_p$. That is to say, for $\boldsymbol{\beta}^\ell$ equal to the estimate of $\boldsymbol{\beta}$ at the ℓth iteration or the IRWLS algorithm, provide explicit expressions for: $\boldsymbol{\eta}^\ell$, $\boldsymbol{\mu}^\ell$, $\mathbf{Z}^\ell$, $\mathbf{W}^\ell$, and $\boldsymbol{\beta}^{\ell+1}$.

4. Read the paper by Hu and Smyth. We will focus primarily on Section 4 of the paper, but you should read the rest of the paper to provide context.

(a) In Sections 4.1 and 4.2, what do p_i, λ and d_i represent?

(b) In Section 4.2, the authors state: "We now show the Poisson assumption is not required. Suppose that the number of cells d_i in each culture is counted exactly, by some experimental means. In that case, d_i is a fixed quantity, and the use of the Poisson probabilities is inappropriate." What result shown in class and stated in your text are the authors referring to? What is the appropriate distributional model if the d_i are fixed?

(c) In Sections 4.1 and 4.2, two different expression for $1 - p_i$ are stated. Explain how the two different expressions for $1 - p_i$ are determined based on the stated assumptions. Then show how the two results stated in terms of the complementary log-log transformation are obtained. (N.b. There is an error in the second formula shown in Section 4.2. It should read: $\log(1 - p_i) = d_i \log(1 - \lambda)$.)

(d) Derive the expressions for $\hat{\lambda}$ shown in Sections 4.1 and 4.2 based on fitting the stated simple regression model. Why is $1/\hat{\lambda}$ a sensible estimator for "... the active cell frequency ..." (Section 4.1)?

(e) Using the data shown in Table 5, fit an appropriate GLM and state the parameter estimate for the model. Then estimate both λ and the active cell frequency and corresponding 95% CIs under the assumed models in Sections 4.1 and 4.2. (N.b. As stated in Section 4.1, "The regression slope β is equal to one. An important point is that, when the regression model is estimated, the slope is kept fixed at this value rather than being estimated de novo, so α is the only unknown quantity. (In GLM terminology, x_i is known as an offset.))" In `SAS`' `PROC LOGISTIC`, use the `OFFSET = <`*`variable name`*`>` option in the `MODEL` statement.

5. Exercise 4.18

6. Exercise 4.27