140.753 Final Exam

(7 problems, 12 pages, 100 points in total)

Time: 10:35 – 11:50am

Name:

Department:

1. [10 points] Consider data from a prospective study and a logistic regression model

$$\log \frac{\pi_i}{1-\pi_i} = \theta_0 + \theta_1 x_i$$

(1) [5 pt] The parameter θ_0 can be interpreted as

(a) an odds

(b) a log odds

(c) a log odds ratio

(d) an odds ratio

(2) [5 pt] $\exp(\theta_1)$ can be interpreted as

(a) an odds

(b) a log odds

(c) a log odds ratio

(d) an odds ratio

2. [10 points] Consider data from a retrospective case-control study and a logistic regression

$$logit(\pi_i) = \theta_0 + \theta_1 x_i.$$

(1) [5 pt] θ_0 can be interpreted as

(a) a probability

(b) an odds as in a prospective study

(c) a log odds as in a prospective study

(d) none of the above

(2) [5 pt] $\exp(\theta_1)$ can be interpreted as

(a) an odds

(b) a log odds

(c) a log odds ratio

(d) an odds ratio

3. [10 points] Consider logistic regression fitted using data from a prospective study

$$\log \frac{\pi}{1-\pi} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2.$$

exp (θ_3) can be interpreted as

(a) an odds

(b) an odds ratio

(c) a ratio of odds ratio

(d) none of the above

4. [10 points] In a matched case-control study, data are collected to investigate how various factors affect the risk of diabetes. Each diabetes case is matched with three normal controls with similar age, gender and weight. In addition to these matching variables, two other covariates – dietary fats and smoking status – are also collected for each individual. Which one of the following statements is true:

(a) One can use these data to study how gender affects diabetes risk.

(b) One can use the conditional likelihood approach to study how diabetes risk depends on weight.

(c) One can use these data to study how diabetes risk depends on dietary fats.

(d) The data can be used to estimate the odds of getting diabetes for a woman in the population given her age, weight, dietary fat and smoking status.

5. [20 points] In order to study the association between cardiovascular (heart) disease (CVD) and gender, 240 subjects (130 males and 110 females) are randomly sampled from the population, and their CVD status is summarized in the 2×2 table below.

	Normal	CVD
Male	102	28
Female	95	15

(1) [5 pt] Use a binomial logistic regression to test whether the CVD risk is associated with gender. Write down the model and statistical test you use, and provide the fitted model and conclusions.

Let $CVD_i = 1$ or 0 indicate whether subject i has CVD or not.

Let $Fem_i = 1$ or 0 indicate whether subject i is female or male.

Define $\pi_i = Pr(CVD_i = 1 \mid Fem_i)$

Model: $\text{logit}\, \pi_i = \beta_0 + \beta_1 Fem_i$

The fitted model is

$$\text{logit}\, \pi_i = -1.29 - 0.55\, Fem_i$$

SE (0.21) (0.35)

To test association, the hypotheses are

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_1: \beta_1 \neq 0$$

Wald* test: $Z = -1.579$, p-value $= 0.114$

or Drop-in-Deviance test: $D = 2.57$ p-value $= 0.109$ > 0.05

Cannot reject H_0. CVD risk is not associated with gender.

(2) [10 pt] For the same data, can one use Poisson log-linear model to test the association between CVD and gender? If your answer is yes, use the Poisson regression to test the association. If your answer is no, explain why.

Yes.

Let Y_{ij} be the count in each cell, and μ_{ij} be its mean.

Fit model $\log \mu_{ij} = \lambda + \lambda_i^{Fem} + \lambda_j^{CVD}$

[Note that technically, one can vectorize Ys into $\begin{pmatrix} 102 \\ 28 \\ 95 \\ 15 \end{pmatrix}$

The design matrix is $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$

The regression coefficients are $\beta = \begin{pmatrix} \lambda \\ \lambda^{Fem} \\ \lambda^{CVD} \end{pmatrix}$

Thus $\log \mu = X\beta$].

The residual deviance of this model can be used to test association.

Residual deviance $= 2.57$

d.f. $= 1$

p-value based on χ_1^2 distribution $= \underline{0.109}$

This is the same as the result from the drop-in-deviance test in (1)

(3) [5 pt] Suppose instead of sampling 240 random subjects, the data in the above table were obtained using a retrospective case-control design. 10% of CVD patients and 1% of normal people in a small city were randomly recruited to the study, resulting in 197 normal controls and 43 CVD patients. Their genders were then observed. Using logistic regression, estimate the (prospective) probability that a female randomly sampled from the city has CVD.

Let $S_i = 1$ or 0 indicate whether subject i has been sampled or not,

$$\pi_1 = Pr(S_i = 1 \mid CVD_i = 1) = 0.1$$

$$\pi_0 = Pr(S_i = 1 \mid CVD_i = 0) = 0.01$$

In a retrospective case-control study, the model we fitted is

$$\text{logit } Pr(CVD_i = 1 \mid S_i = 1, Fem_i) = -1.29 - 0.55\, Fem_i \quad (\text{From } (1)).$$

In a prospective study, the logistic regression would be

$$\text{logit } Pr(CVD_i = 1 \mid Fem_i) = \beta_0 - 0.55\, Fem_i$$

Here $\beta_0 = -1.29 - \log \frac{\pi_1}{\pi_0} = -1.29 - \log \frac{0.1}{0.01} = -3.59$

Thus $Pr(CVD_i = 1 \mid Fem_i = 1) = \dfrac{e^{-3.59 - 0.55}}{1 + e^{-3.59 - 0.55}} = 0.016$

(Note: In fact, in this simple case, you can quickly check whether your answer is correct using

$$\frac{15 \times 1}{95 \times 10 + 15 \times 1} = \frac{15}{965} = 0.016\ )$$

6. [20 points] Researchers are interested in studying whether the number of daily car accidents depends on the weather condition. For each day in 2013, they have collected the following data from a city:

- *acc*: number of accidents
- *rain*: a binary indicator for whether it rained (1) or not (0).
- *weekday*: a binary indicator. 1=Monday-Friday; 0=Saturday and Sunday.

They have fitted two Poisson log-linear models using these data. The results are summarized below.

Model A: $\log(\mu_{acc}) = \beta_0 + \beta_1 rain + \beta_2 weekday$

```
Call:
glm(formula = acc ~ rain + weekday, family = poisson)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-5.216  -1.562  -0.381   1.135   6.725

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.94338    0.03532   55.02   <2e-16 ***
rain         0.66709    0.02834   23.54   <2e-16 ***
weekday      0.81558    0.03713   21.96   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2561.7  on 364  degrees of freedom
Residual deviance: 1462.2  on 362  degrees of freedom
AIC: 3054.9

Number of Fisher Scoring iterations: 5
```

Model B: $\log(\mu_{acc}) = \beta_0 + \beta_1 rain + \beta_2 weekday + \beta_3 rain \times weekday$

```
Call:
glm(formula = acc ~ rain * weekday, family = poisson)
```

```
Deviance Residuals:
   Min       1Q   Median       3Q      Max
-5.014  -1.516  -0.410   1.210   6.557

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.00940    0.03972  50.595  < 2e-16 ***
rain          0.43066    0.07851   5.485 4.13e-08 ***
weekday       0.73726    0.04339  16.991  < 2e-16 ***
rain:weekday  0.27422    0.08421   3.256  0.00113 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2561.7  on 364  degrees of freedom
Residual deviance: 1451.3  on 361  degrees of freedom
AIC: 3046

Number of Fisher Scoring iterations: 5
```

(1) [5pt] Is overdispersion a concern when analyzing this data set? Why?

Yes. Check Model B (Full model).

Residual deviance = 1451.3 > 361 (Note: $E\chi^2_{361} = 361$).

Dispersion $\hat{\sigma}^2 = \frac{1451.3}{361} = 4.02$

(2) [10pt] Is it necessary to include the rain:weekday interaction term in the model? Why?

When there is overdispersion, use the following F-test

$$F = \frac{\frac{1462.2 - 1451.3}{1}}{\frac{1451.3}{361}} = \frac{10.9}{4.02} = 2.71$$

p-value based on $F_{1,361} = 0.10$

It is unnecessary to include the interaction term.

[Note: If you use t-test with SE inflated by $\hat{\sigma}$, you will also get credits].

(3) [5pt] Based on Model A, calculate the percent increase in the mean number of accidents comparing a Monday with rain to a Monday without rain. Also provide a 95% confidence interval.

The 95% CI for β_{rain} :

$$\hat{\beta}_{rain} \pm t_{0.975,\, df=361} \cdot \hat{\sigma} \cdot \underbrace{SE(\hat{\beta}_{rain})}_{\text{From Model A fitting result}}$$

$$= 0.667 \pm 1.967 \cdot \sqrt{4.02} \cdot 0.0283$$

$$= 0.667 \pm 0.112$$

$\therefore$ 95% CI for β_{rain} : $[0.555, 0.779]$

95% CI for $e^{\beta_{rain}}$: $[1.742, 2.179]$

$\therefore$ The percent increase is $e^{\beta_{rain}} - 1 = e^{0.667} - 1 = 94.8\%$

95% CI : $[74\%, 118\%]$

7. [20 points] Consider data collected from n independent subjects: $\{(x_i, y_i) : i = 1, \ldots, n\}$. Treat Y as response and X as covariate. X is univariate. Define $\mu_i = E(Y_i|X_i)$ and $\sigma_i^2 = Var(Y_i|X_i)$. Assume $\log \mu_i = X_i\beta$ and $\sigma_i^2 = \phi\mu_i^{\frac{1}{2}}$.

(1) [5 pt] Derive the quasi-likelihood for β using this data set.

For one observation,

$$Q(\mu; y) = \int_y^\mu \frac{y-t}{\phi \cancel{\mu} t^{\frac{1}{2}}} dt$$

$$= \frac{1}{\phi}\left(2yt^{\frac{1}{2}} - \frac{2}{3}t^{\frac{3}{2}}\right)\Big|_y^\mu$$

$$= \frac{1}{\phi}\left(2y\mu^{\frac{1}{2}} - \frac{2}{3}\mu^{\frac{3}{2}} \underline{- \frac{4}{3}y^{\frac{3}{2}}}\right)$$

(Note: For inferring β, this part is irrelevant).

For the whole data set

$$Q(\vec{\mu}; \vec{y}) = \sum_i Q(\mu_i; y_i)$$

$$= \frac{1}{\phi}\sum_i \left(2y_i\mu_i^{\frac{1}{2}} - \frac{2}{3}\mu_i^{\frac{3}{2}}\right) + \text{Constant that does not involve } \beta.$$

(2) [10 pt] Provide a solution to estimate β using quasi-likelihood. Provide necessary details for implementing the solution.

$$U(\beta) = \frac{dQ}{d\beta} = D^T V^{-1}(y-\mu)/\phi$$

$$= \frac{1}{\phi}\sum_i x_i e^{x_i\beta}\left(\frac{y_i - \mu_i}{\mu_i^{\frac{1}{2}}}\right)$$

$$= \frac{1}{\phi}\sum_i x_i\left(y_i\mu_i^{\frac{1}{2}} - \mu_i^{\frac{3}{2}}\right)$$

Here $y - \mu = \begin{pmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{pmatrix}$, $V = \text{diag}\{\mu_1^{\frac{1}{2}}, \ldots, \mu_n^{\frac{1}{2}}\}$, $D = \begin{pmatrix} \frac{d\mu_1}{d\beta} \\ \vdots \\ \frac{d\mu_n}{d\beta} \end{pmatrix}$

$$-E\frac{dU(\beta)}{d\beta} = i_\beta = D^T V^{-1} D/\phi = \sum_i (x_i e^{x_i\beta})^2 / \phi\mu_i^{\frac{1}{2}} = \sum_i \frac{x_i^2 \mu_i^{\frac{3}{2}}}{\phi} = \sum_i \frac{x_i^2 [e^{x_i\beta}]^{\frac{3}{2}}}{\phi}$$

Using Fisher Scoring

$$\beta^{(t+1)} = \beta^{(t)} + \left[(D^T V^{-1} D)^{-1} D^T V^{-1}(y-\mu)\right]^{(t)}$$

$$= \beta^{(t)} + \left[\sum_i x_i^2 e^{\frac{3}{2}x_i\beta^{(t)}}\right]^{-1} \sum_i x_i\left(y_i e^{\frac{1}{2}x_i\beta^{(t)}} - e^{\frac{3}{2}x_i\beta^{(t)}}\right)$$

iterate until convergence.

(2) cont'd:

(3) [5 pt] Derive the asymptotic variance of the quasi-likelihood estimate $\hat{\beta}$.

$$i_\beta^{-1} = \phi (D^T V^{-1} D)^{-1}$$

$$= \phi \left[\sum_i x_i^2 e^{\frac{3}{2} x_i \beta} \right]^{-1}$$