# Homework #3

Dustin Leatherman

February 2, 2020

## Contents

# 1 1

## 1.1 a

> Estimate the dispersion parameter.

For Binomial or Poisson distributions, the dispersion parameter can be estimated by the *Pearson Chi-Squared Statistic*

$$\hat{\phi} = \frac{X^2}{n-p} = 1.2927 \tag{1}$$

## 1.2 b

> Compute Full Log Likelihood and BIC for the model (hint AIC is given)

$$\begin{aligned} AIC =& -2l(\hat{\pi}; y) + 2p \\ BIC =& -2l(\hat{\pi}; y) + p \times log(n) \end{aligned} \tag{2}$$

**Log Likelihood**

$$\begin{aligned} 92.2094 =& -2\ l(\hat{\pi}; y) + 2p \\ 46.1047 =& -\ l(\hat{\pi}; y) + 2 \\ -44.1047 =& l(\hat{\pi}; y) \end{aligned} \tag{3}$$

**BIC**

$$-44.1047 \times -2 + 2 \times log(22) = 94.39148 \tag{4}$$

# 2 2

Let Y = number of ACFs in the rat colons and x = sacrificed times (endtime, 6, 12, and 18). Compute the predicted probabilities for Y = 2, 4, 8, and x = 12.

| | Y | P(y/ x = 12) |
|---|---|---|
| 1 | 2 | |
| 2 | 4 | |
| 3 | 8 | |

$$\begin{aligned} P(y) =& \frac{e^{\nu}}{1+e^{\nu}} \\ \nu =& \beta_0 + \Sigma\beta_i x_i \end{aligned} \tag{5}$$

$$\begin{aligned} P(y) =& 0.8807971,\ \ \nu = 2 \\ P(y) =& 0.9820138,\ \ \nu = 4 \\ P(y) =& 0.9996646,\ \ \nu = 8 \\ P(y) =& 0.8610574,\ \ \nu = -0.3215 + 0.1192 \times 12 \end{aligned} \tag{6}$$

## 2.1 a

How do we interpret $\hat{\beta}_1 = 0.1192$?

The log-odds of the number of ACFs in the rat colons increases by 0.1192 for each increase in sacrifice time.

# 3 3

To study factors that affect the recurrence of heart attacks (HA), an investigator collected data from 20 HA victims. The investigator fit a logistic regression model with an indicator of a second HA within one year (1 = HA; 0 = no HA) as the binary outcome. There are two predictors: $x_1 = 1$ if the patient completed an anger management program; 0 else. $x_2$ = anxiety score (0 = low, 100 = high). Computer output is given below:

|  | Estimate | Std Err | Z value | P Value |
|---|---|---|---|---|
| Intercept | -6.36 | 3.21 | -1.98 | 0.05 |
| X1 | -1.02 | 1.17 | -0.88 | 0.38 |
| X2 | 0.12 | 0.06 | 2.17 | 0.03 |

### 3.1 a

In terms of $x_1$ and $x_2$, what are the odds of a patient having a second heart attack?

$$\begin{aligned} \omega_{AB} =& \frac{\omega_A}{\omega_B} \\ =& \frac{e^{X0+X1\times1+X2\times A}}{e^{X0+X1\times0+X2\times B}} \\ =& e^{X1(1-0)+X2(A=B)} \\ =& e^{X1+X2(A-B)} \end{aligned} \tag{7}$$

### 3.2 b

What is the probability of a second heart attack for a patient that has completed an anger management program and scored a 100 on the anxiety test?

$$\begin{aligned} \pi =& \frac{e^{\eta}}{1+e^{\eta}} \\ =& \frac{e^{-6.36-1.02\times1+0.12\times100}}{1+e^{-6.36-1.02\times1+0.12\times100}} \\ =& 0.9902433 \end{aligned} \tag{8}$$

### 3.3 c

For patients that have completed the anger management program, is high anxiety associated with an increased probability of a second heart attack?

Regardless of whether or not a patient has completed the anger management program, there is moderate evidence that a higher anxiety score is associated with an increased risk of a second heart attack (p-value = 0.03).

### 3.4 d

> Is there statistical evidence that an anger management program is associated with a reduction in the probability of a second heart attack? Explain.

There is no evidence that completion of the anger management program is associated with reduced probability of a second heart attack (p-value = 0.38). The confidence interval for the anger management predictor encompasses 0 indicating that there is no conclusive effect on the estimated predicted probability.

### 3.5 e

> Explain why linear regression is in appropriate for modeling the probability of a second heart attack.

Linear Regression may yield invalid values, in this case, an estimated probability using Linear Regression may be outside the interval of $[0, 1]$. Linear Regression could be used to estimate a relative score or value that could be interpreted in a similar fashion but there would be no guarantee on the range of scores that would occur; whereas when modeling probabilities, the probability is guaranteed to be between 0 and 1.

## 4 4

> Let Y be a binomial distribution. Show taht Y has the exponential distribution of the form:
>
> $$f(y;\theta) = s(y)y(\theta)exp(a(y)b(\theta));$$
>
> this can be rewritten
>
> $$f(y;\theta) = exp(a(y)(\theta) + c(\theta) + d(y))$$

### 4.1 a

> Clearly identify the link function, $b(\theta)$

$$
\begin{aligned}
f(y;\pi) =&\binom{n}{y}\pi^y(1-\pi)^{n-y}\\
=&exp(y\ log(\pi)+(n-y)log(1-\pi)+log(\binom{n}{y}))\\
=&exp(y\ log(\pi)+n\ log(1-\pi)-y\ log(1-\pi)+log(\binom{n}{y}))\\
=&exp(y\ (log(\pi)-log(1-\pi))+n\ log(1-\pi)+log(\binom{n}{y}))\\
=&exp(y\ log(\frac{\pi}{1-\pi})+n\ log(1-\pi)+log(\binom{n}{y}))
\end{aligned}
\tag{9}
$$

$a(y) = y$
$b(\theta) = log(\frac{\pi}{1-\pi})$
$c(\theta) = nlog(1-\pi)$
$d(y) = log(\binom{n}{y})$

# 5 5

For games in baseball's National League during nine decades: The following table shows the percentage of times that the starting pitcher pitched a complete game.

| | $\text{Decade}_{\text{complete}}$ | Percent |
|---|---|---|
| 1 | 1900-1909 | 72.7 |
| 2 | 1910-1919 | 63.4 |
| 3 | 1920-1939 | 50 |
| 4 | 1930-1939 | 44.3 |
| 5 | 1940-1949 | 41.6 |
| 6 | 1950-1959 | 32.8 |
| 7 | 1960-1969 | 27.2 |
| 8 | 1970-1979 | 22.5 |
| 9 | 1980-1989 | 13.3 |

## 5.1 a

Treating the number of games as the same in each decade, the ML fit of the linear probability model is $\hat{p} = 0.7578 - 0.0694x$, where x = decade [1:9]. Interpret

Each additional decade starting at 1900 is associated with a 6.94% *decrease* in the percentage of times that the starting pitcher pitched a complete game.

### 5.2 b

Substituting $x = 10, 11, 12$, predict the percentage of complete games for the next three decades. Are these predictions plausible? Why?

$$\begin{aligned} 0.7578 - 0.0694 \times 11 &= -0.0056 \\ 0.7578 - 0.0694 \times 12 &= -0.075 \\ 0.7578 - 0.0694 \times 13 &= -0.1444 \end{aligned} \tag{10}$$

These predictions are not plausible because they fall outside the range between 0 and 1. This is one of the reasons why linear regression is not suitable for predicting probabilities.

### 5.3 c

The ML Fit with logistic regression is

$$\hat{p} = exp(1.148 - 0.315x)/(1 + exp(1.148 - 0.315x))$$

Obtain for $x = 10, 11, 12$. Are these more plausible?

$$\begin{aligned} exp(1.148 - 0.315 \times 10)/(1 + exp(1.148 - 0.315 \times 10)) &= 0.1189931 \\ exp(1.148 - 0.315 \times 11)/(1 + exp(1.148 - 0.315 \times 11)) &= 0.08972478 \\ exp(1.148 - 0.315 \times 12)/(1 + exp(1.148 - 0.315 \times 12)) &= 0.06710713 \end{aligned} \tag{11}$$

These are more plausible since the values are valid (between 0 and 1) and still show a decreasing probability over time.

## 6 6

Show that the following probability density functions belong to the exponential family:

### 6.1 a

Pareto distribution

$$f(y : \theta) =^{-\theta - 1}$$

$$\begin{aligned} f(y;\theta) &= \theta Y^{-\theta-1} \\ &= exp((-\theta - 1)\ log(y) + log(\theta)) \\ &= exp(-\theta log(y) - log(y) + log(\theta)) \end{aligned} \tag{12}$$

$$\begin{aligned} a(y) &= \ - log(y) \\ b(\theta) &= \theta \\ c(\theta) &= log(\theta) \\ d(y) &= \ - log(y) \end{aligned} \tag{13}$$

### 6.2 b

Exponential distribution

$$f(y;\theta) = \theta\ exp(-y\theta)$$

$$\begin{aligned} f(y;\theta) &= \theta\ exp(-y\theta) \\ &= exp(log(\theta) - y\theta) \end{aligned} \tag{14}$$

$$\begin{aligned} a(y) &= \ - y \\ b(\theta) &= \theta \\ c(\theta) &= log(\theta) \\ d(y) &= 0 \end{aligned} \tag{15}$$

## 7 7

The following associations can be described by generalized linear models. For each one:

1. Identify the response variable and the explanatory variables
2. Select a probability distribution for the response (justifying your choice)
3. Write down the linear component

## 7.1 a

The effect of age, sex, height, mean daily food intake, and mean daily energy expenditure on a person's weight.

1. A person's weight.

2. t-distribution since weight is a nominal value with no inherent limitations in terms of range of values.

3. $\hat{weight} = \beta_0 + \beta_1\ age + \beta_2\ isMale + \beta_3\ height + \beta_4\ avgDailyFoodIntake + \beta_5\ avgDailyEnergyExpend$

## 7.2 b

The proportion of laboratory mice that become infected after exposure to bacteria when five different exposure levels are used and 20 mice are exposed at each level.

1. Proportion of infected laboratory mice

2. Binomial Distribution since a mouse can either be infected or not infected.

3. $\hat{infected} = \beta_0 + \beta_1\ exp1 + \beta_2\ exp2 + \beta_3\ exp3 + \beta_4\ exp4 + \beta_5\ exp5$ where exp1 through exp5 are indicator variables (1 when exposed at a given level; 0 otherwise).

## 7.3 c

The association between the number of trips per week to the supermarket for a household and the number of people in the household, the household income, and the distance of the supermarket.

1. The number of trips per week to the supermarket.

2. Poisson or Negative Binomial Distribution. If a Poisson model is fit and there is over-dispersion, then a Negative Binomial Distribution may be a better fit. Both Poisson and Negative Binomial are useful distributions for modeling *count* data, which this response variable is.

3. $\hat{tripsPerWeek} = \beta_0 + \beta_1 \ numPeople + \beta_2 \ income + \beta_3 \ distance$