MLDS 401/IEMS404: Homework 8
Due: November 27, 3pm
Professor Malthouse

The first three problems attempt to show you some of the key relationships between cross tabulations, Poisson count models logistic regression and the softmax function, which is central to classification problems, especially with neural networks.

1. (10 points) This problem and the next give you a taste of an important use of Poisson models, the loglinear model. This problem is a review of the chi-square test of independence, which is commonly covered in undergrad statistics. The next problem shows its relationship to the loglinear model. The eastern factory had 28 accidents last year, out of a work force of 673. The western factory had 31 accidents during this period, out of 1,306 workers. Thus follows this cross tabulation (the n notation will be used in the next problem):

Factory	No Accident (0)	Accident (1)	Total
East (0)	$n_{00} = 645$	$n_{01} = 28$	$n_{0+} = 673$
West (1)	$n_{10} = 1275$	$n_{11} = 31$	$n_{1+} = 1306$
Total	$n_{+0} = 1920$	$n_{+1} = 59$	$n = 1979$

Here are two ways to store the data in R:

```
dat = expand.grid(factory=c("East", "West"), accident=c("No", "Yes"))
dat$y = c(645,1275, 28,31)
tab = matrix(dat$y, nrow=2,
  dimnames=list(factory=c("East", "West"), accident=c("No", "Yes")))
```

(a) (2 points) If accidents were independent of factory, how many accidents would you expect in the west? Show work. Hint: events A and B are independent $\iff P(A \cap B) = P(A)P(B)$, then multiply by n to get the expected count. *Answer:* $59 \times 1306/1979 = 38.94$

(b) (2 points) Find all four expected cell counts. Hint:

```
chisq.test(tab)$expected

        accident
factory         No      Yes
   East   652.9358 20.06417
   West  1267.0642 38.93583
```

(c) (2 points) Let m_{ij} be the expected count in factory i and accident status j. Let π_{i+} be the marginal probability of a randomly selected person coming from factory i (e.g., $\pi_{1+} = P(\text{west})$) and π_{+j} be the probability of being in accident state j (e.g., $\pi_{+1} = P(\text{accident})$). Generalizing the previous part, write out an expression for m_{ij} as a function of n, π_{i+} and π_{+j}. *Answer:* $m_{ij} = n\pi_{i+}\pi_{+j}$

(d) (2 points) Take logs of both sides of the expression from the previous part and write the log of the product as the sum of the logs of individual terms. (You should recognize that, under independence, the log expected cell counts are an *additive* function consisting of a row effect, a column effect and a constant (intercept). You should find this fact exciting.) *Answer:* $\log m_{ij} = \log n + \log \pi_{i+} + \log \pi_{+j}$

(e) (2 points) Continuing to assume independence, write out $\log \pi_{ij}$ (π_{ij} is the joint probability) as a function of π_{i+} and π_{+j}. *Answer:* $\log \pi_{ij} = \log \pi_{i+} + \log \pi_{+j}$

2. (20 points) Continuing the previous problem, we will estimate the *log cell counts* as a dependent variable from the factory and whether or not there was an accident. Let `west` be a dummy variable that takes the value 1 if the factory is the west 0 for the east. Let `accident` equal 1 if there was an accident and 0 if not. So, n_{ij} is the observed number of workers in factory i (0=east, 1=west) with accident status j (0=no, 1=yes). The expected cell counts (or means) are still m_{ij}.

(a) (2 points) Estimate the following "main-effects" model with Poisson errors.

$$\log(m_{ij}) = \alpha + \beta_1 \texttt{west} + \beta_2 \texttt{accident}$$

```
glm(y ~ factory + accident, poisson, dat)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.48148    0.03875  167.27   <2e-16 ***
west         0.66298    0.04745   13.97   <2e-16 ***
accident    -3.48254    0.13217  -26.35   <2e-16 ***

    Null deviance: 2423.492  on 3  degrees of freedom
Residual deviance:    4.678  on 1  degrees of freedom
```

(b) (2 points) Use the main-effects model to estimate the unlogged number of accidents in the west. Show work. (You should deduce that the main-effects model gives the expected cell counts if accidents were independent of factory.) *Answer:* $\log m_{11} = 6.48 + 0.66 - 3.48 = 3.66192$, $\hat{m}_{11} = e^{3.66} = 38.94$.

(c) Include an interaction between `west` and `accident`:

$$\log(m_{ij}) = \alpha + \beta_1 \texttt{west} + \beta_2 \texttt{accident} + \beta_3 \texttt{west} \times \texttt{accident}$$

```
fit2 = glm(y ~ factory*accident, poisson, dat)
```

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      6.46925    0.03937 164.299   <2e-16 ***
west             0.68145    0.04832  14.103   <2e-16 ***
accident        -3.13705    0.19304 -16.251   <2e-16 ***
west:accident   -0.57967    0.26515  -2.186   0.0288 *

    Null deviance: 2423.492 on 3  degrees of freedom
Residual deviance: 0  on 0  degrees of freedom
```

(2 points) Use the interaction model to estimate the unlogged number of accidents in the west. Show work. *Answer:* $\log m_{11} = 6.47 + 0.68 - 3.137 - 0.5797 = 3.434$, $\hat{m}_{11} = e^{3.43} = 31$.

(d) (2 points) Explain briefly why the residual deviance of the interaction model is 0 (and thus the model fits perfectly). *Answer: This is the saturated model—you are using 4 parameters to represent 4 numbers.*

(e) (2 points) Test whether the interaction (in the second model) is significant using the z-value given in the output. *Answer:* $H_0 : \beta_3 = 0$ *versus* $H_1 : \beta_3 \neq 0$, $P = 0.0288 < 5\%$ *so reject* H_0.

(f) (2 points) When you can reject the null hypothesis in the previous part, what does it tell you about whether factory is independent of accidents? *Answer: They are not independent.*

(g) (2 points) We could alternatively use the likelihood ratio test to evaluate the interaction. Give the test statistic and P-value. *Answer: Use* `drop1` *to find the LRT test statistic to be* $4.679 - 0 = 4.678$ *with* $P = 0.03055$.

(h) (4 points) Estimate the log odds of an accident in the east using the parameter estimates from the *interaction* model. Separately, estimate the log odd of an accident in the west. Hint: $\log[\pi_{1|i}/(1 - \pi_{1|i})] = \log(m_{i1}/m_{i0})$, using notation defined in the next part. *Answer:*

- *East:* $\log[\pi/(1-\pi)] = \beta_2 = -3.137$
- *West:* $\log[\pi/(1-\pi)] = \beta_2 + \beta_3 = -3.137 - 0.580 = -3.72$

(i) (4 points) Let $\pi_{1|i} = \pi_{i1}/\pi_{i+}$ be the conditional probability that an accident occurs in factory i. Use the results from the previous problem to find values c and d so that

$$\log\left(\frac{\pi_{1|i}}{1 - \pi_{1|i}}\right) = \log\left(\frac{\pi_{1|i}}{\pi_{0|i}}\right) = c + d \times \texttt{west}$$

(You should note that this is a logistic regression of accident on factory! Logistic regression and log-linear models are thus closely related.) *Answer:* $\log[\pi/(1 - \pi)] = \beta_2 + \beta_3\texttt{west} = -3.137 - 0.580\texttt{west}$.

(j) (2 points) Confirm your answer to the previous part by regressing `accident` on `factory` using logistic regression.

```
glm(accident ~ factory, family=binomial, data=dat, weights=y)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.1370     0.1930 -16.251   <2e-16 ***
factoryWest  -0.5797     0.2651  -2.186   0.0288 *

    Null deviance: 530.73  on 3  degrees of freedom
Residual deviance: 526.06  on 2  degrees of freedom
AIC: 530.06
```

3. This problem uses the news desert data from the previous homework assignment.

```
> poisson1 = glm(Cpub2023 ~ age + SES21 + Lpopdens2021 + Lblack2021
    + Lhisp2021, family=poisson, data=dat)
> summary(poisson1) # part b
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.95154    0.13639  -6.976 3.03e-12 ***
age           0.01054    0.00273   3.861 0.000113 ***
SES21         0.14832    0.01462  10.146  < 2e-16 ***
Lpopdens2021  0.28120    0.01038  27.088  < 2e-16 ***
Lblack2021   -0.07448    0.01401  -5.317 1.06e-07 ***
Lhisp2021     0.19242    0.01513  12.719  < 2e-16 ***

    Null deviance: 5943.5  on 3139  degrees of freedom
Residual deviance: 3813.3  on 3134  degrees of freedom
AIC: 11003

> poisson2 = glm(Cpub2018 ~ log(Cpub2018+1) + age + SES21 + Lpopdens2021
    + Lblack2021 + Lhisp2021, family=poisson, data=dat)
> summary(poisson2) # part c

                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.5554788  0.1353769  -4.103 4.07e-05 ***
log(Cpub2018 + 1)  1.2278642  0.0184001  66.732  < 2e-16 ***
age              -0.0027386  0.0027999  -0.978    0.328
SES21             0.0202752  0.0149226   1.359    0.174
Lpopdens2021      0.0007736  0.0117123   0.066    0.947
Lblack2021       -0.0161786  0.0140533  -1.151    0.250
Lhisp2021        -0.0145631  0.0150506  -0.968    0.333

    Null deviance: 7149.86  on 3139  degrees of freedom
Residual deviance:  245.18  on 3133  degrees of freedom

> 1-logLik(poisson2) / logLik(glm(Cpub2018~1, poisson, data=dat))
```

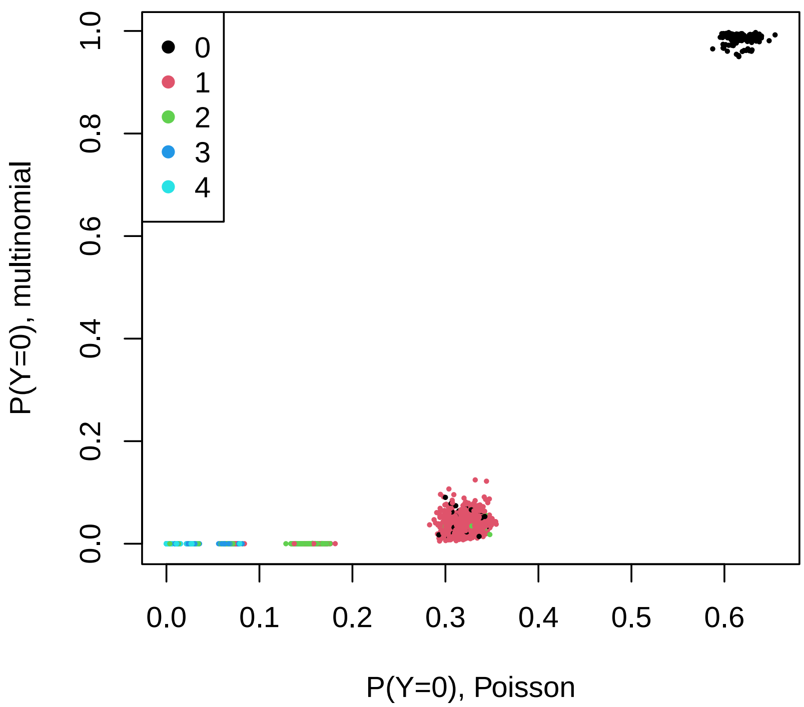
```
'log Lik.' 0.4719568 (df=7)  # Pseudo R^2

plot(P0, mult2$fitted.values[,1], xlab="P(Y=0), Poisson",
     ylab="P(Y=0), multinomial", pch=16, cex=.4,
     col=if_else(dat$Cpub2023>4, 4, dat$Cpub2023)+1)
legend("topleft", as.character(0:4), col=1:5, pch=16)
```

(a) Use a Poisson regression to predict `numPub23` from the demographic variables. Interpret the model. Which variables are are associated with more news organizations? Fewer? Not important? Compute Pseudo R^2. *Answer: See above for output. In desending order of z statistics, population density has the strongest positive association with count ($z = 27.1$), followed by Hispanic ($z = 12.7$), SES ($z = 10.1$), and age ($z = 3.9$). The only variable with a negative association is black ($z = -5.3$). The conclusions are similar to the logit model, although the sign for the age effect flips They are similar to the* $\log(2+/1)$ *multinomial model.*

(b) Now add `log(numPub18+1)` to the model (AR1+). Interpret the results. Compute Pseudo R^2 *Answer: As with the logit models, the only signifiant variable is* `log(Cpub2018+1)`. *Pseudo $R^2 = 47\%$.*

(c) For the AR1+ model, use the Poisson assumption to estimate the probability that $Y = 0$, i.e., news desert. Compute a scatterplot matrix with these probabilities and $\mathsf{P}(Y = 0)$ from the multinomial model in the previous homework. *Answer: See the plot below, which is a little more detailed than the question asks. The actual values are coded in colors. The Poisson model is a bit more conservative, e.g., the black clump of actual 0's in teh upper left have multinomial prbabilities near 1 and Poisson probabilities around 0.7. Likewise, the greens (actual=2) and blues (actual ≥ 3) have multinomials near 0 and Poissons that mostly separate the 2s, 3s, and 4+s; in this case the other multinomial logits should also separate the groups. The red clump of actual=1 is consistently in the middle.*


0
1
2
3
4
1.0
0.8
0.6
0.4
0.2
0.0
P(Y=0), multinomial
0.0
0.1
0.2
0.3
0.4
0.5
0.6
P(Y=0), Poisson