MSiA 401: Homework 7
Due: December 3, noon
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) 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.

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

```
chisq.test(tab)$expected
```

(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}.

(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.)

(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}.

2. (22 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)
   ```

   (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.)

   (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)
   ```

   (2 points) Use the interaction model to estimate the unlogged number of accidents in the west. Show work.

   (d) (2 points) Explain briefly why the residual deviance of the interaction model is 0 (and thus the model fits perfectly).

   (e) (2 points) Test whether the interaction (in the second model) is significant using the z-value given in the output.

   (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?

   (g) (2 points) We could alternatively use the likelihood ratio test to evaluate the interaction. Give the test statistic and P-value.

(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.

(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.)

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

(k) (2 points) If accidents were independent of factory, what would you expect the estimated logistic regression to be?

3. Suppose we have a sample of size n where observation i consists of dependent variable Y_i, a multinomial RV taking values $\{1, \ldots, K\}$, and $(p+1)$-vector of predictor variables $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^\mathsf{T}$. Let $\boldsymbol{\alpha}_k$ be a $(p+1)$-vector of regression coefficients. Let $\pi_{ik} = \mathsf{P}(Y_i = k)$ for $k = 1, \ldots, K$ and

$$\log \pi_{ik} = \boldsymbol{\alpha}_k^\mathsf{T}\mathbf{x}_i - \log Z, \qquad (k = 1, \ldots, K)$$

where log is the natural log function and the term $\log Z$ ensures that the probabilities some to one, i.e., $\sum_{k=1}^{K} \pi_{ik} = 1$.

(a) Show that $Z = \sum_{k=1}^{K} \exp(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i)$.

(b) Show that $\pi_{ik} = \exp(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i)/Z$. This is called the softmax function.

(c) The usual formulation of the multinomial logit from class picks a base category (WLOG class 1) and assumes:

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \boldsymbol{\beta}_k^\mathsf{T} \mathbf{x}_i, \qquad (k = 2, \dots, K)$$

How is $\boldsymbol{\beta}_k$ related to $\boldsymbol{\alpha}_k$?

4. This problem studies news deserts. You have data for nearly every county in the US:

- `numPub`: number of newspapers published for the county. This count is the dependent variable.
- `age`: average age in county
- `BAhigher`: percentage of county residence with a BA degree or higher (this is a measure of education level)
- `pop`: population of the county
- `income`: median income in county
- `raceBlack`: percent of county that is black
- `raceHisp`: percent of county that is Hispanic
- `digDistress`: a measure of "digital distress," where larger values indicate lower broadband penetration, lower internet speeds, poorer cell phone connections, etc.

Build a Poisson model to answer the question, how the other variables affect `numPub`? Here are some things to consider:

- You will likely have to transform at least some of the predictor variables.
- Digital distress may considered a pipe, where broadband and cellular providers concentrate their service on large cites (i.e., large `pop`) and wealthy counties, i.e., pop and income cause (lower) digital distress.