Assignment 1

Arthur Yu

1) For each of the situations below, identify the response and explanatory variables, variable types, and the generalized linear model that is well-suited to model the data. Make sure to justify your answer. Write down the linear predictor as well as the link function.

- **The effect of age, sex, height, daily food intake and daily exercise on a person's weight.**

  Response variable: W = Weight (Continuos, Normally distributed)

  Explanatory variables: A = Age (Discrete), S = Sex (Nominal), H = Height (Continuous), F = Daily food intake (Continuos), E = Daily excercise (Continuous)

  Because response variable is continuous and explanatory variables contain numeric and nominal variable, we should choose multiple regression.

  Which is like: $E(W) = \beta_0 + \beta_A x_A + \beta_S x_S + \beta_H x_H + \beta_F x_F + \beta_E x_E$

  Link function: $\eta = \mu$

- **The percentage of full-time graduate students that find employment upon graduation. For each student, sex, age, grades, major, prior years of work experience, and prior income levels are available.**

  Response variable: P = Percentage of full-time graduate students that find employment upon graduation (Count data as proportion)

  Explanatory variables: S = Sex (Nominal), A = Age (Discrete), G = Grades (Ordinal), M = Major (Nominal), W = Prior years of work experience (Discrete), I = Prior income levels (Ordinal)

  Because response variable is count data as proportion and there are different types of explanatory variables, we should use Binomial regression.

  Which is like: $P = \dfrac{exp(\beta_0 + \beta_S x_S + \beta_A x_A + \beta_G x_G + \beta_M x_M + \beta_W x_W + \beta_I x_I)}{1 + exp(\beta_0 + \beta_S x_S + \beta_A x_A + \beta_G x_G + \beta_M x_M + \beta_W x_W + \beta_I x_I)};$

  Link function: $\eta = ln(\dfrac{\mu}{1-\mu})$

- **The number of mortgage loan defaults in a given year by different counties across the United States. For each household/borrower information on income, loan interest rate, age, debt, loan to value at origination are available.**

  Response variable: D = Number of defaults. (Count data)

  Explanatory variables: I = Income (Continuous), L = Loan interest rate (Continuous), A = Age (Discrete), B = Debt (Continuous), O = loan to value at origination (Continuous)

  Because response variable is count data, we should use Poisson regression

Which is like: $ln\lambda = (\beta_0 + \beta_I x_I + \beta_L x_L + \beta_A x_A + \beta_B x_B + \beta_O x_O)$; $D \sim Poisson(\lambda)$

Link function: $\eta = ln(\mu)$

2) Show that the probability distributions below belong to the exponential family. Determine the functions a( ), b( ), c( ), and d( ) from the general exponential distribution for each of the probability distributions below.

- Pareto distribution:

$$\begin{aligned} f(y;\theta) &= \theta y^{-\theta-1} \\ &= exp\{ln[\theta y^{-(\theta+1)}]\} \\ &= exp[ln\theta - (\theta+1)lny] \\ &= exp(ln\theta - \theta lny - lny) \end{aligned}$$

$$So, a(y) = lny, b(\theta) = -\theta, c(\theta) = ln\theta, d(y) = -lny$$

- Exponential distribution:

$$\begin{aligned} f(y;\theta) &= \theta e^{-y\theta} \\ &= exp(ln\theta - y\theta) \end{aligned}$$

$$So, a(y) = y, b(\theta) = -\theta, c(\theta) = ln\theta, d(y) = 0$$

- Negative binomial:

$$\begin{aligned} f(y;\theta) &= \binom{y+r-1}{r-1}\theta^r(1-\theta)^y \\ &= exp\{ln[\binom{y+r-1}{r-1}\theta^r(1-\theta)^y]\} \\ &= exp[ln\binom{y+r-1}{r-1} + rln\theta + yln(1-\theta)] \end{aligned}$$

$$So, a(y) = y, b(\theta) = ln(1-\theta), c(\theta) = rln\theta, d(y) = ln\binom{y+r-1}{r-1}$$

3) Show that the gamma distribution with a scale parameter θ and nuisance parameter ϕ belongs to the exponential family of distributions. Using the properties of the exponential distributions, find $E[Y]$ and $VAR[Y]$.

$$\begin{aligned} f(y;\theta) &= \frac{y^{\phi-1}\theta^{\phi}e^{-y\theta}}{\Gamma(\phi)} \\ &= exp[lny^{(\phi-1)} + ln\theta^{\phi} + lne^{-y\theta} - ln\Gamma(\phi)] \\ &= exp[(\phi-1)lny + \phi ln\theta - y\theta - ln\Gamma(\phi)] \end{aligned}$$

$$So, a(y) = y, b(\theta) = -\theta, c(\theta) = \phi ln\theta - ln\Gamma(\phi), d(y) = (\phi-1)lny$$

$$\begin{aligned} E[a(Y)] &= -\frac{c'(\theta)}{b'(\theta)} \\ &= -\frac{(\phi ln\theta - ln\Gamma(\phi))'}{-\theta'} \\ &= \frac{\phi}{\theta} \end{aligned}$$

$$\begin{aligned} VAR[a(Y)] &= \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \\ &= \frac{(-\theta)''(\phi ln\theta - ln\Gamma(\phi))' - (\phi ln\theta - ln\Gamma(\phi))''(-\theta)'}{(-\theta')^3} \\ &= \frac{0 - (-\frac{\phi}{\theta^2})(-1)}{(-1)^3} \\ &= \frac{\phi}{\theta^2} \end{aligned}$$