1. **Poisson Regression**

The Normal linear model assumes that the mean of the response variable is a linear function of one or more \( x \) variables, and that the errors are Normal. Generalized linear models allow for different error distributions and a more general “link function” between the mean of the response and the linear function of the \( x \)’s.

a) Explain how the exponential family parameterization suggests a link function for a given error distribution. Use the Normal, Poisson and Gamma distributions as examples.

b) Describe an algorithm for finding the maximum likelihood estimate for \( \beta \) in the case of Poisson regression (note that this is also the posterior mode if you assume a flat prior density for \( \beta \)). Show how to get standard errors for the estimates.

c) Simulate data to evaluate the performance of such estimates.

2. **Logistic regression**

Logistic regression may be the most commonly used generalized linear model. If responses are coded 0 or 1, the mean response is the probability of a 1. It wouldn’t make sense to model this as a linear function of \( x \) because you could end up fitting probabilities outside of \([0, 1]\).

a) Show that the Bernoulli and Binomial distributions are exponential families, and identify the canonical parameter and link function. Write out the likelihood function for \( \beta \).

b) Work out an algorithm for maximizing the likelihood (or posterior density) for \( \beta \) and demonstrate it for the NCAA tournament data. Estimate the probability of a team winning given the Vegas betting spread on the game.

c) Discuss how to decide whether or not to include predictors in a logistic regression model, and for other generalized linear models. For example, the difference in the seeds is also useful for predicting whether a team will win. But is it still useful after you know the Vegas spread?

3. **Latent Variables and Probit Regression**

Probit regression models the probabilities of success/failure outcomes as depending on one or more possibly continuous covariates. With one predictor variable \( x \), the Probit model can be written as:

\[
Y_i \mid \pi_i \overset{\text{indep}}{\sim} \text{Bin}(1, \pi_i), \quad \pi_i = \Phi(\beta_0 + \beta_1 x_i), \quad i = 1, \ldots, n,
\]

were \( \Phi \) is the CDF for the standard Normal distribution (the pnorm function in R). Probit regression differs from Logistic regression in that the logistic model assumes
logit(\(\pi_i\)) = X'_i \beta, \text{ while the probit model assumes } \Phi^{-1}(\pi_i) = X'_i \beta \text{ (the functions logit() and } \Phi^{-1}() \text{ are the link functions for the logistic and probit regression models).}

This model implicitly assumes a latent variable \(\theta_i \sim \text{i} \text{N}(\mu_i, 1)\), where \(\mu_i = X'_i \beta\). For \(i = 1, \ldots, n\), we observe \(Y_i = 1\) if \(\theta_i \geq 0\) and \(Y_i = 0\) if \(\theta_i < 0\).

a) Show that the latent variable assumption leads to the Probit model.

b) Write out the conditional density for \(\theta_i | Y_i, \beta\) and derive an expression for \(E(\theta_i | Y_i, \beta)\). It will involve \(\Phi()\) and \(\beta = (\beta_0, \beta_1)'\).

c) For the Probit model, the EM Algorithm alternates between taking the expectation of \(\sum X_i \theta_i\) (the “E-step”) given the \(Y_i\)’s and a current estimate of \(\beta\), and replacing the estimate of \(\beta\) with its maximum likelihood estimate (the “M-step”) given that the complete-data sufficient statistic \(S = X'\theta\) takes its expected value from the E-step. This algorithm converges to the marginal MLE of \(\beta\) (i.e., unconditional on the \(\theta_i\)’s). Note that if we assume a prior density \(p(\beta) \propto c\), then this is also the posterior mode for \(f(\beta | y)\). Implement this algorithm in R for estimating winning probabilities for NFL teams, given the Vegas betting spread. Verify that the marginal joint-Likelihood function \(L(\beta_0, \beta_1)\) increases at each step in the algorithm.

**Problem to turn in.** The file NFL2009.txt contains data for the 256 football games played during the 2009 NFL regular season. Each game has a home team (i.e., the host team) and a road team. The margin of victory (Actual H-R) is coded to be positive if the home team wins and negative if the road team wins. Similarly, the Vegas betting spread (Vegas H-R) is coded to be positive if the home team is favored and negative if the road team is favored. The Vegas spread may be considered a casino’s estimate of the mean number of points by which the home team will win (actually their goal is to come up with a number that half of betters will think is too high, and half will think is too low). The halftime spread is the home team’s winning margin at halftime.

a) The least squares fit for Actual H-R on Vegas H-R is pretty close to the line \(\hat{\mu}_i = x_i\). Verify that the model assumptions seem reasonable, and carry out an extra sum of squares \(F\) test of \(H_o : \beta_0 = 0 \cap \beta_1 = 1\), vs. \(H_a : \text{ “not } H_o.\)” Explain your conclusion in the context of this problem.

b) The fit for Actual H-R on halftime H-R is also very close to \(\hat{\mu}_i = x_i\). Test to see whether the Vegas H-R value is still useful as a predictor after you know the halftime score. Find a 95% CI for the coefficient for Vegas H-R in the multiple regression, and give a careful explanation of what it represents.

c) Suppose your favorite team was favored by 7 points (Vegas H-R = 7) but at halftime the score is tied (Halftime H-R = 0). Find a 95% prediction interval for the Actual H-R you can expect in this game. Also find a confidence interval for the mean Actual H-R in all games where the home team is favored by 7 but tied at halftime.