**Statistics 347, Homework 5, due Feb 14**

Discussion of homework problems among students is encouraged. However, all material handed in for credit must be your own work.

**Please hand in each problem in a separate file with your name on it.**

1. Let $Y$ be Poisson with mean $\mu$. Recall that $EY = \mu, Var(Y) = E(Y-\mu)^2 = \mu, E(Y-\mu)^3 = \mu$.

   (a) Show that

   $$E\sqrt{Y} \sim \sqrt{\mu}\left(1 - \frac{1}{8\mu}\right), \text{and } Var(\sqrt{Y}) \sim \frac{1}{4}.$$

   Hint: Define $Z = (Y-\mu)/\mu$, write $Y = \mu \cdot (1+Z)$ and use the Taylor expansion.
   This is called a *variance stabilizing* transformation. The variance of the transformed Poisson no longer depends on the mean.

   (b) Use the data `warpbreaks` in the R datasets. This provides the number of breaks in different types of wool at different levels of tension. Use a Poisson glm to fit a main effects model `mod1`. Plot the residuals and summarize the conclusions of the model.

   (c) Use the variance stabilizing transformation `sqrt` and fit a linear model `mod2` with the same predictors. Again plot the residuals and summarize the conclusions of this model on the transformed data.
   Compare the fitted values of `mod1,mod2`.

2. Recall the gamma distribution has the form:

   $$G(\mu,\nu)(y) = \frac{\nu^\nu}{\Gamma(\nu)\mu^\nu} y^{\nu-1} \exp\left(-\frac{\nu}{\mu}y\right)$$

   (a) The coefficient of variation is the ratio of the Sd to the mean of a variable. Denote by $\sigma^2 = var(Y)/(E(Y))^2$. Show that $\sigma^2 = 1/\nu$.

   (b) Show that if $Y \sim G(y;\mu,\nu)$ then $E\log Y \sim \log\mu - \sigma^2/2$ and $Var(\log(Y)) \sim \sigma^2$. (Hint: expand log around $\mu$.) This is a variance stabilizing transformation for the Gamma distribution.

   (c) Using the moment generating function of $\log Y$, show the following exact equalities:
   $E\log Y = \log\mu + \psi(\nu) - \log(\nu)$
   $Var(\log Y) = \psi'(\nu)$,
   where $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$.

   fixed v

   (d) Suppose you have independent observations $Y_i \sim G(\mu_i, \nu)$, with log-link: $\log(\mu_i) = \alpha + x_i^t\beta$ for $i = 1,\ldots,n$.
   Write out the score equations and the information matrix **J** at the estimated $\hat\beta$. Don't use the general formulas of glms' do this directly on the log-likelihood.

   (e) Now instead of fitting the generalized linear model you fit the linear model: $E\log(Y_i) = \alpha^* + x_i^t\beta^*$.
   What is the covariance matrix of $\hat\beta^*$.
   How does this compare to the covariance matrix of $\hat\beta$?

   (f) Load the data ('fly.Rdata'). We are given information on the duration `duration` of the embryonic period of flies subjected to different temperatures `temp`. The experiment was done with batches of flies and only the mean `duration` and standard deviation of the duration `dur.sd` is reported for each batch. The number of flies in the batch is given in `batch`.

      i. Plot `dur.sd` on `duration` and explain why a Gamma glm would make sense for this data.

ii. Plot `log(duration)` against `temp` and `1/temp` explain why you would try to fit `duration ~ 1/temp`.

iii. Fit `mod1` with duration regressed on `1/temp` using the Gamma glm, `log` link, with `weights=batch` to adjust for the size of each batch. Plot the Pearson residuals against temp and comment on the fit.

iv. On the plot of the data plot the curve defined by your estimate. What do you notice for large values of temperature.

v. To fix this problem we make two modifications to the model. First we add an additional predictor `temp` and we change `1/temp` to `1/(temp-d)`. Explain why we subtract the constant `d`.

vi. For the range of values $d \in [45, 100]$ fit the glm with the two predictors `temp,1/(temp-d)` and see at which $d^*$ gives the smallest deviance. Provide the summary of the model `mod2` with the optimal $d^*$. What are the residual degrees of freedom? Estimate the $\sigma$ for this model and plot the standardized residuals against `temp`. Plot the estimated curve on top of the data. How does this compare to `mod1`?

vii. We saw above that the log transform stabilizes the variance. So now fit `log(duration)` using a linear model with the same predictors as `mod2`. How do the coefficients of this model compare to `mod2`.