1. **Fisher Information**. For a GLM, the estimate of the expected Fisher information is

$$\mathbf{F}(\hat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{W}\mathbf{X}$$

where $\mathbf{W}$ is a diagonal matrix with i-th diagonal entry:

$$w_i = \frac{\left[\frac{\delta h}{\delta \eta}(\hat{\eta}_i)\right]^2}{\hat{\text{var}}(y_i)}, \quad \hat{\text{var}}_i(y_i) = \text{var}(y_i|\hat{\eta}_i)$$

   (a) Find an expression for w_i for a Bernoulli GLM with canonical logit link.

   (b) Use your expression from (a) to replicate the p-values in the following linear regression (see "LinearRegression.r" for details)

```
admissions <-read.csv(''http://www.ats.ucla.edu/stat/data/binary.csv'')
fit=glm(admit~gre+gpa+rank,family=binomial,data=admissions)
summary(fit)
```

2. **Smoking** The "smoking.csv" file contains observational data on death rates for smokers and non-smokers of different ages. The recorded items are as follows:

   - agecat. Age category 1: 35-44, 2: 45-54, 3: 55-64, 4: 65-74, 5: 75-84
   - smokes. Indicator variable (1=smokes, 0=does not).
   - deaths. Number of deaths recorded in the observational window.
   - pyears. Person-years of observation for each category. This item is a record of the total amount of observation time of all people in the category.

   Provide an estimate and a 95% confidence interval of the expected number of deaths in one year for (1) 1000 smokers aged 55-64 and (2) 1000 non-smokers aged 55-64. Clearly describe all models and methods used.

3. **Patents**. On page 8 of the FKLM textbook is a description of the data found in "patent.csv". For this problem, ONLY CONSIDER THE FOLLOWING ITEMS IN THE DATASET:

   - ncit
   - biopharm
   - year

   The "ncit" variable gives the number of citations from the year the patent was granted through 1998. Of interest are the following two questions.

   How does the average number of citations per year of a patent vary between biotech / pharmaceutical patents (biopharm=1) and patents from other industries?

   Are more recent patents cited more or less often than older patents?

   Describe all models and methods used to answer these two questions.

4. **Mistletoe** The "mistletoe.csv" file contains forest stand information from black spruce stands in Minnesota forests. The recorded items are as follows:

   - csize. Cover size class, indicating overall size of canopy trees. Values range from 1 (small canopy trees) to 5 (large canopy trees).
   - cdense. Cover density class, indicating overall density of canopy trees. Values range from 1 (sparse canopy) to 5 (dense canopy)
   - usize. Understory size class, indicating overall size of understory trees. Values range from 1 (small trees) to 5 (large trees).
   - udense. Understory density class, indicating overall density of understory. Values range from 1 (sparse) to 5 (dense)
   - si. Site Index: This is a conventional forestry measure of the timber-growing quality of a site for the primary species occupying it. It represents the average height attained by a tree of that species at 50 years of age.
   - phys. Water drainage through the soil of the stand. Ranges from 1 (very dry, water drains quickly) to 5 (very wet, water doesn't drain well).
   - age. Average age of trees in stand.
   - ba. Basal area per acre. A measure of how densely the stand is stocked.
   - dbh. Average diameter at breast height for all trees in the stand.
   - height. Average height of trees in the stand.
   - volume. A measure of the total volume of trees in the stand.
   - mortal. Mortality. Indicator variable coded as 1 if there are visible dead trees in the stand and 0 if not.
   - dense. Another measure of stand density (stems per acre)
   - x,y. "x" and "y" coordinates of the stand's spatial location (in UTMs).
   - infected.mndnr. Indicator variable coded as 1 if the MNDNR survey found mistletoe in the stand and 0 otherwise.
   - infected.usu. Indicator variable coded as 1 if the USU survey found mistletoe in the stand and 0 otherwise. NA values indicate that the USU survey did not survey this stand.

   Data on all forest stands were collected as part of the Minnesota Department of Natural Resources (MNDNR) forest inventory. The "infected.mndnr" item indicates whether or not dwarf mistletoe was found in the stand. The MNDNR survey was broadly focused and collected data on a large number of stand characteristics. Because of this, there is reason to believe that the MNDNR survey is not always accurate at finding dwarf mistletoe when it is there.

   A separate, smaller survey focused primarily on mistletoe presence or absence was conducted by scientists from Utah State University (USU). This survey is likely to be very accurate, but only covers a small fraction of the stands surveyed by the MNDNR. The "infected.usu" item indicates whether or not dwarf mistletoe was found in a stand by the USU survey. For the purposes of this homework, we will consider the "infected.usu" to be the true presence / absence of mistletoe in the stand.

   (a) Conduct an analysis to identify stand characteristics that are linked to errors in the MNDNR survey (relative to the USU survey). Are there stand characteristics that make it hard to detect mistletoe when it is present? Are there stand characteristics that are common when mistletoe is erroneously reported in the MNDNR survey? Support any statements you make with statistical evidence.

   (b) Use the stands that were surveyed by both the USU and MNDNR surveys to predict what the USU survey would have found if it were able to survey all stands. Your goal here is only prediction. You are not interested in hypothesis testing. Try multiple approaches to finding a good predictive model, including ridge regression and lasso regression. Provide some justification of the accuracy of your predictive model, and predict the percent of stands where mistletoe would be found if if the USU survey were conducted at every stand.