IYSE 6420 Fall 2020 Homework5

Xiao Nan
GT Account: nxiao30
GT ID: 903472104

1. Paddy Soil Adhesion.
Pan and Lu (1998) provide measurements of adhesion on 43 pairs of samples of paddy soil to steel and rubber. From 1974 to 1983, during the rice-growing season, the adhesion of soils to steel and to rubber were measured in situ simultaneously in paddy fields in South China. As steel and rubber have long been the most important mate- rials used for wetland running gears such as wheel and track, it is expected that the adhesion to them would be roughly the same. The adhesion was measured with an adhesometer.
Data set paddy.dat has two columns: (1) adhesion to steel, and (2) adhesion to rubber. Both measurements are given in kPa.
(a) Fit the linear regression model where the response variable y is adhesion to rubber. Report the parameter estimates and Bayesian R^2
(b) What adhesion with rubber do you predict in paddy soil for which adhesion to steel was 2. Find 95% credible set for a single predictive response.

(a)
Linear Regression,

$$y = \mathrm{X}\beta + \epsilon$$

From dataset,

$$n = 43, p = 2$$

Least square estimator,

$$\hat{\beta} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T y$$

Let

$$C = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1}$$

We have

$$SSE = \sum (y - \mathrm{X}\hat{\beta})^2$$
$$SST = \sum (y - \bar{y}C)^2$$

From the Matlab code we can see that,

$$\beta = 0.7693, \epsilon = 0.3974$$

Thus,

$$\bar{y} = 0.7693x + 0.3974$$

Coefficient of determination,

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{(y - \mathrm{X}\hat{\beta})^T(y - \mathrm{X}\hat{\beta})}{(y - \bar{y}C)^T(y - \bar{y}C)} = 1 - \frac{6.723}{21.743} = 0.691$$

(b)
For new data point,

$$x_n = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$\hat{y}_n = 0.7693 \times 2 + 0.3974 = 1.936$$
$$\hat{\sigma}^2 = MSE = \frac{SSE}{n - p}$$
$$SE\{\hat{y}_n\} = \sqrt{MSE} \times \sqrt{1 + x_n^T(X^TX)x_n}$$

$$(1-\alpha)100\%\ CI: \hat{y}_n \pm t_{\alpha/2,n-p} SE\{\hat{y}_n\}$$

From the Matlab code, the 95% credible set is:

$$[1.1025, 2.7697]$$

Matlab Code

```
%% (a)
load paddy.dat;
x = paddy(:,1);
y = paddy(:,2);
n = length(x);
C = ones(n,1);
X=[C x];
p = size(X,2);
params = inv(X'*X)*X'*y;
epsilon = params(1)
beta = params(2)
y_pred = epsilon + beta * x;
SSE = sum((y - y_pred).^2)
SST = sum((y - mean(y)).^2)
R2 = 1 - SSE/SST
```

```
%% (b)
xn = [1 2];
yn = params' * xn'
se_yn = sqrt(SSE/(n-p)) * sqrt(1 + xn*inv(X'*X)*xn');
alpha = 0.05;
lb = yn - tinv(1-alpha/2, n-p) * se_yn;
rb = yn + tinv(1-alpha/2, n-p) * se_yn;
[lb rb]
```

2. Third-degree Burns.

The data for this exercise, discussed in Fan et al. (1995), refer to n = 435 adults who were treated for third-degree burns by the University of Southern California General Hospital Burn Center. The patients were grouped according to the area of third-degree burns on the body. For each midpoint of the groupings "log(area + 1)," the number of patients in the corresponding group who survived and the number who died from the burns was recorded:

Log(area+1)	Survived	Died
1.35	13	0
1.60	19	0
1.75	67	2
1.85	45	5
1.95	71	8
2.05	50	20
2.15	35	31
2.25	7	49
2.35	1	12

(a) Fit the logistic regression on the probability the covariate x = log(area+1). What is the deviance?

(b) Using your model, find the posterior probability of survival for a person for which log(area + 1) equals 2.

(a)

Logistic regression

$$\mathrm{F}^{-1}(\mathrm{p}) = \log\left(\frac{p}{1-p}\right) = b_0 + b_1 X$$

Using the attached Matlab code we get,

$$b_0 = 22.822$$
$$b_1 = -10.736$$
$$\mathrm{F}^{-1}(\mathrm{p}) = \log\left(\frac{p}{1-p}\right) = b_0 + b_1 X = 22.822 - 10.736X$$

Deviance

$$D = -2log\frac{likelihood\ of\ the\ fitted\ model}{likelihood\ of\ the\ saturated\ model} = -2\sum_{i=1}^{k}[y_i log\widehat{p_i} + (1-y_i)\log(1-\widehat{p_i})]$$

From the Matlab code we get,

$$D = 0.1523$$

(b)

$$posterior = \frac{1}{1+\exp\{-(b_0 + b_1 x^*)\}} = \frac{1}{1+\exp\{-(22.822 - 10.736 \times 2)\}} = 0.794$$

Matlab Code

```
%% Problem 2
%% (a)
X = [1.35, 1.6, 1.75, 1.85, 1.95, 2.05, 2.15, 2.25, 2.35];
S = [13 19 67 45 71 50 35 7 1];
D = [0 0 2 5 8 20 31 49 12];
Y = S./(S+D);
[bs,stderr,phat,deviance] = logisticmle(Y,X);
% output
bs
deviance
```

```
%% logisticmle.m

function [bs,stderr,phat,deviance] = logisticmle(y,x)
%[bs,stderr,phat,deviance] = logisticmle(y,x)
%  Input:
% y - responses, a binary vector, values 0 and 1
% x - the covariate, as a vector
%
% Output:
% bs - estimators of beta0 and beta1
% stderr - standard error of the estimate
% phat - estimator of p= P(Y=1)
% deviance - deviance
%
%
x=x(:); y=y(:);
 %initialize at LS estimate if vector bstart not given
b1 =  sum((y-mean(y)).*(x-mean(x)))/sum((x-mean(x)).^2);
b0 =  mean(y) - b1 * mean(x);
bs= [b0; b1];
%============================================
% Refine bs by Newton-Raphson
diff = 1;   precision = 1.0E-6;
while diff > precision;
    bsold = bs;
    p = exp(bs(1)+bs(2)*x)./(1+exp(bs(1)+bs(2)*x));
    score =   [sum(y-p);  sum((y-p).*x)];
    Infmat =  [sum(p.*(1-p))                    sum(p.*(1-p).*x);
                      sum(p.*(1-p).*x)                 sum(p.*(1-
p).*x.*x)    ];
```

```
    bs =    bsold + Infmat\score;
    diff =         sum((bs-bsold).^2);
end
Covmat = inv(Infmat);
stderr = sqrt(diag(Covmat));
phat = exp(bs(1)+bs(2)*x)./(1+exp(bs(1)+bs(2)*x));
deviance = 2* sum( y.* log((y+eps)./phat)  + ...
    (1-y) .*log((1-y+eps)./(1-phat)) );
```

3. SO_2 , NO_2 , and Hospital Admissions.

Fan and Chen (1999) discuss a public health data set consisting of daily measurements of pollutants and other environmental factors in Hong Kong between January 1, 1994 and December 31, 1995. The association between levels of pollutants and the number of daily hospital admissions for circulation and respiratory problems is of particular interest.

The data file hospitaladmissions.dat consists of six columns: (1) year, (2) month, (3) day in month, (4) concentration of sulfur dioxide SO2, (5) concentration of pollutant nitrogen NO2, and (6) daily number of hospital admissions.

(a) Fit a Bayesian Poisson regression model, explaining how the expected number of hospital admissions varies with the levels of SO2 and NO2.

(b) Suppose that on a particular day the levels of SO2 and NO2 were measured as 44 and 100, respectively. Estimate the expected number of hospital admissions and report 95% credible set.

In this problem, only last 3 columns are needed. Use noninformative priors on all parameters in you model.

(a)

$$y_i \sim Poi(\lambda)$$
$$\theta = log\lambda = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$\lambda = Ey_i$$

Using the Matlab code to fit a poisson model we can get,

$$[\beta_0, \beta_1, \beta_2] = [5.4618, -0.0018, 0.0025]$$

Since x_1 is the concentration of SO_2 and x_2 is the concentration of NO_2,
the expected number of admissions has a positive correlation with concentration of NO_2, and a negative correlation with concentration of SO_2.

(b)

For the new point,

$$x_n = [1\ 44\ 100]$$

We have the prediction from Matlab code:

$$\widehat{y_n} = 278.4106$$

95% credible set

$$[273.4122, 283.5005]$$

Matlab Code

```
%% Problem 3
clear all
close all
%% (a)
load hospitaladmissions.dat;
x = hospitaladmissions(:,4:5);
y = hospitaladmissions(:,6);
xdes = [ones(size(y)) x];
[n p] = size(xdes);
```

```
[b dev stats] = glmfit(x,y,'poisson','link','log')
%% (b)
mdl =  fitglm(x,y,'linear','Distribution','poisson');
xn = [44 100];
[yhatn,cin] = predict(mdl,xn,'Alpha',0.05,'Simultaneous',true)
```

Reference:

[1] http://springer.bme.gatech.edu/Ch17.Logistic/Logisticmat/logisticmle.m
[2] http://springer.bme.gatech.edu/Ch17.Logistic/Logisticmat/ihga.m
[3] https://www.mathworks.com/help/stats/generalizedlinearmodel.predict.html