ISYE 6420 Fall 2020 Final

Xiao Nan
GT Account: nxiao30
GT ID: 903472104

1. Vasoconstriction. The data give the presence or absence (y_i = 1 or 0) of vasoconstriction in the skin of the fingers following inhalation of a certain volume of air (v_i) at a certain average rate (r_i). Total number of records is 39. The candidate models for analyzing the relationship are the usual logit, probit, cloglog, loglog, and cauchyit models.

Data are given as follows.

y:1,1,1,1,1,1,0,0,0,0,0,0,0,1,1,1,1,1,
0,1,0,0,0,0,1,0,1,0,1,0,1,0,0,1,1,1,0,0,1

v:3.7, 3.5, 1.25, 0.75, 0.8, 0.7, 0.6, 1.1, 0.9, 0.9, 0.8, 0.55, 0.6, 1.4, 0.75, 2.3, 3.2, 0.85, 1.7, 1.8, 0.4, 0.95, 1.35, 1.5, 1.6, 0.6, 1.8, 0.95, 1.9, 1.6, 2.7, 2.35, 1.1, 1.1, 1.2, 0.8, 0.95, 0.75, 1.3

r: 0.825, 1.09, 2.5, 1.5, 3.2, 3.5, 0.75, 1.7, 0.75, 0.45, 0.57, 2.75, 3, 2.33, 3.75, 1.64, 1.6, 1.415,
1.06, 1.8, 2, 1.36, 1.35, 1.36, 1.78, 1.5, 1.5, 1.9, 0.95, 0.4, 0.75, 0.3, 1.83, 2.2, 2, 3.33, 1.9, 1.9, 1.625

(a) Transform covariates v and r as

$$x_1 = log(10 \times v), x_2 = log(10 \times r).$$

(b) Estimate posterior means for coefficients in the logit model. Use noninformative priors on all coefficients.

(c) For a subject with v = r = 1.5, find the probability of vasoconstriction.

(d) Compare with the result of probit model. Which has smaller deviance?

ANSWER

(b)

Logit:

$$\log \frac{p}{1-p} = F^{-1}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

F is logistic cdf.

From the matlab code attached, we can find that

$$\beta_0 = -25.6083, \beta_1 = 5.2205, \beta_2 = 4.6312$$

So

$$F^{-1}(p) = -25.6083 + 5.2205 \times x_1 + 4.6312 \times x_2$$

(c)

$$ypred = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 \times x_1^* + \beta_2 \times x_2^*)\}}$$
$$= \frac{1}{1 + \exp\{-(-25.6083 + 5.2205 \times \log(10 \times 1.5) + 4.6312 \times \log(10 \times 1.5))\}}$$
$$= 0.7447$$

(d)
Probit:

$$F^{-1}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

F is normal cdf

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 attached matlab code,

$$Logit\ Deviance = 29.2640$$
$$Probit\ Deviance = 29.3215$$

Logit model has a smaller deviance.

Matlab code

```
%%% problem 1
y =
[1,1,1,1,1,1,0,0,0,0,0,0,0,1,1,1,1,1,0,1,0,0,0,0,1,0,1,0,1,0,1,0,0,1,1,1,
0,0,1];
v = [3.7, 3.5, 1.25, 0.75, 0.8, 0.7, 0.6, 1.1, 0.9, 0.9, 0.8, 0.55, 0.6,
1.4, 0.75, 2.3, 3.2, 0.85, 1.7, 1.8, 0.4, 0.95, 1.35, 1.5, 1.6, 0.6, 1.8,
0.95, 1.9, 1.6, 2.7, 2.35, 1.1, 1.1, 1.2, 0.8, 0.95, 0.75, 1.3];
r= [0.825, 1.09, 2.5, 1.5, 3.2, 3.5, 0.75, 1.7, 0.75, 0.45, 0.57, 2.75,
3, 2.33, 3.75, 1.64, 1.6, 1.415, 1.06, 1.8, 2, 1.36, 1.35, 1.36, 1.78,
1.5, 1.5, 1.9, 0.95, 0.4, 0.75, 0.3, 1.83, 2.2, 2, 3.33, 1.9, 1.9,
1.625];

x1 = log(10*v);
x2 = log(10*r);

X = [x1' x2'];
Xdes =[ones(size(y')) x1' x2'];
n = length(y');

[b,dev,stats] = glmfit(X,y','binomial','logit');
logitFit = glmval(b,X,'logit');

% (b) get betas
b

% (c) prediction
xnew = [log(10*1.5) log(10*1.5)];
ypred = 1 / (1 + exp(-(b(1)+b(2)*xnew(1)+b(3)*xnew(2))))

% (d) probit
[bp,devp,statsp] = glmfit(X,y','binomial','probit');
dev
devp
```

2. Magnesium Ammonium Phosphate and Chrysanthemums. Walpole et al. (2007) provide data from a study on the effect of magnesium ammonium phosphate on the height of chrysanthemums, which was conducted at George Mason University in order to determine a possible optimum level of fertilization, based on the enhanced vertical growth response of the chrysanthemums. Forty chrysanthemum seedlings were assigned to 4 groups, each containing 10 plants. Each was planted in a similar pot containing a uniform growth medium. An increasing concentration of $MgNH_4PO_4$, measured in grams per bushel, was added to each plant. The 4 groups of plants were grown under uniform conditions in a greenhouse for a period of 4 weeks. The treatments and the respective changes in heights, measured in centimeters, are given in the following table:

Solve the problem as a Bayesian one-way ANOVA. Use STZ constraints on treatment effects.

50g/bu	100g/bu	200g/bu	400g/bu
13.2	16	7.8	21
12.4	12.6	14.4	14.8
12.8	14.8	20	19.1
17.2	13	15.8	15.8
13	14	17	18
14	23.6	27	26
14.2	14	19.6	21.1
21.6	17	18	22
15	22.2	20.2	25
20	24.4	23.2	18.2

(a) Do different concentrations of $MgNH_4PO_4$ affect the average attained height of chrysanthemums? Look at the 95% credible sets for the differences between treatment effects.

(b) Find the 95% credible set for the contrast $\mu_1 - \mu_2 - \mu_3 + \mu_4$.

ANSWER

(a)
Null hypothesis: H_0: all population means μ_i are equal, different concentrations have no effect.

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

H1 : $(H_0)^c$ (or $\mu_i \neq \mu_j$, for at least one pair i, j).
Since the sample sizes are the same, the ANOVA is balanced.

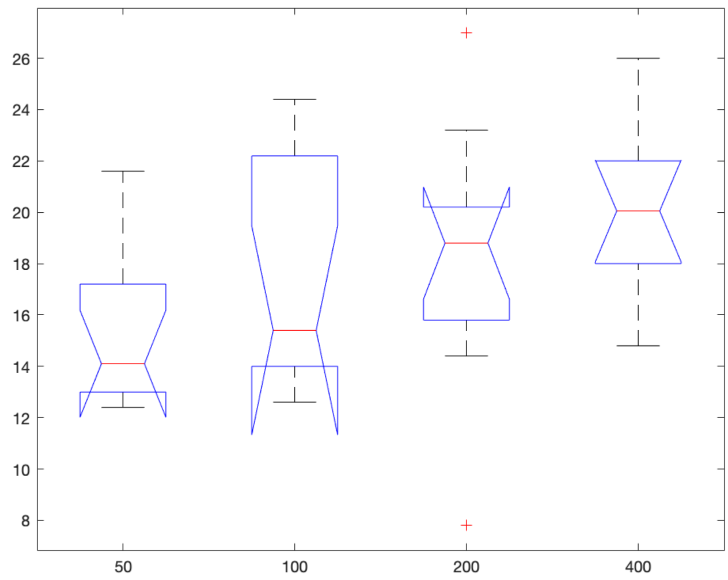
Using STZ,

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0, \ \mu_i = \mu + \alpha_i$$

$$\sum_i \alpha_i = 0$$

Using matlab code we get,

Source	SS	df	MS	F	Prob>F
Groups	119.787	3	39.929	2.25	0.0989
Error	638.248	36	17.7291		
Total	758.035	39			



The observed F = 2.25 < critical value finv(0.95,3,36) = 2.8863. And the p value is 0.0989 > 0.05, we failed to reject null hypothesis. Thus different concentrations of $MgNH_4PO_4$ does not affect the average attained height of chrysanthemums.

(b)

The test for a contrast,

$$H_0\colon \sum_{i=1}^{k} c_i\mu_i = 0 \text{ versus } H_1\colon \sum_{i=1}^{k} c_i\mu_i <, \neq, > 0$$

$(1-\alpha)100\%$ confidence interval,

$$\left[\sum_{i=1}^{k} c_i\bar{y}_i - t_{N-k,1-\alpha/2}\cdot s\cdot\sqrt{\sum_{i=1}^{k}\frac{c_i^2}{n_i}}, \sum_{i=1}^{k} c_i\bar{y}_i + t_{N-k,1-\alpha/2}\cdot s\cdot\sqrt{\sum_{i=1}^{k}\frac{c_i^2}{n_i}}\right]$$

From the matlab code we have,

$H_0\colon \mu_1 - \mu_2 - \mu_3 + \mu_4 = 0$ is not rejected, and the p-value is 0.4970. The 95% creditable set for contrast $\mu_1 - \mu_2 - \mu_3 + \mu_4$ is [-5.5750   5.5350].

Matlab code:

```
%% problem2
% (a)
heights = [13.2 12.4    12.8    17.2    13  14  14.2    21.6    15  20
16  12.6    14.8    13  14  23.6    14  17  22.2    24.4    7.8 14.4
20  15.8    17  27  19.6    18  20.2    23.2    21  14.8    19.1    15.8
18  26  21.1    22  25  18.2];
gbu = [50   50  50  50  50  50  50  50  50  50  100 100 100 100 100 100
100 100 100 100 200 200 200 200 200 200 200 200 200 200 400 400 400 400
400 400 400 400 400 400];
[p,table,stats] = anova1(heights, gbu,'on');
stats
%    gnames: {4√ó1 cell}
%         n: [10 10 10 10]
%    source: 'anova1'
%     means: [15.3400 17.1600 18.3000 20.1000]
%        df: 36
%         s: 4.2106
fcrit = finv(0.95,3,36)
% fcrit = 2.8663
%(b)
m = stats.means
% 15.3400   17.1600   18.3000   20.1000
c = [1 -1 -1 1];
L = c(1)*m(1) + c(2)*m(2)+c(3)*m(3) + c(4)*m(4)
% L = -0.02
LL= m * c'
% LL = -0.02
stdL = stats.s * sqrt(c(1)^2/10+c(2)^2/10+c(3)^2/10+c(4)^2/10)
% stdL = 2.6630
t = LL/stdL
% t = -0.0075

% p-value
tcdf(t, 36)
% 0.4970

% 95% confidence interval for population contrast
[LL - tinv(0.975, 20)*stdL, LL + tinv(0.975, 20)*stdL]
% -5.5750    5.5350
```

3. Hocking–Pendleton Data. This popular data set was constructed by Hocking and Pendelton (1982) to illustrate influential and outlier observations in regression. The data are organized as a matrix of size 26 × 4; the predictors x_1, x_2, and x_3 are the first three columns, and the response y is the fourth column. The data are given in hockpend.dat.

(a) Fit the linear regression model with the three covariates, report the parameter estimates and Bayesian R^2

(b) Is any of the 26 observations influential or outlier (in the sense of CPO and cumulative)?

(c) Find the mean response and prediction response for a new observation with covariates $x^*_1 = 10$, $x^*_2 = 5$, and $x^*_3 = 5$. Report the corresponding 95% credible sets

ANSWER

(a)
Linear Regression,

$$y = X\beta + \epsilon$$

From dataset,

$$n = 43, p = 2$$

Least square estimator,

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

Let

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

We have

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

From the Matlab code we have,

$$\beta = [8.855, 3.420, -1.451, 0.334]$$

Thus,

$$y = 8.855 + 3.42x_1 - 1.451x_2 + 0.334x_3$$
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 0.8628$$

(b)
CPO

$$(CPO)_i = f(y_i \mid y_{-i})$$
$$= \int f(y_i \mid \theta)\pi(\theta \mid y_{-i})d\theta$$
$$(CPO)_i^{-1} = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{f(y_i \mid \theta^b)}$$
$$(CPO)_i < 0.02 \rightarrow potential\ outlier$$

Cumulative

If F is correct distribution for y_i, then $F(y_i) \sim$ Uniform(0,1)

Potential outlier

$$F(y_i) < \frac{c}{n} \text{ and } F(y_i) > 1 - \frac{c}{n}$$

Using the OpenBUGS code attached, we can find

	mean	sd	sample	CPO	CPO < 0.02
icpo[1]	9.252	4.814	100000	0.108084738	0
icpo[2]	10.43	3.618	100000	0.095877277	0
icpo[3]	6.465	1.006	100000	0.154679041	0
icpo[4]	6.662	1.035	100000	0.150105074	0
icpo[5]	6.481	1.016	100000	0.154297176	0
icpo[6]	8.859	3.388	100000	0.112879558	0
icpo[7]	7.819	1.724	100000	0.127893593	0

icpo[8]	7.996	4.349	100000	0.125062531	0
icpo[9]	7.868	2.413	100000	0.127097102	0
icpo[10]	8.265	2.705	100000	0.120992136	0
icpo[11]	10.88	6.148	100000	0.091911765	0
icpo[12]	6.501	1.021	100000	0.153822489	0
icpo[13]	7.418	2.005	100000	0.134807226	0
icpo[14]	7.485	2.152	100000	0.133600534	0
icpo[15]	20830	951200	100000	4.80077E-05	1
icpo[16]	7.568	2.247	100000	0.132135307	0
icpo[17]	34.41	21.55	100000	0.029061319	0
icpo[18]	124.7	324.5	100000	0.008019246	1
icpo[19]	8.28	2.858	100000	0.120772947	0
icpo[20]	9.869	3.589	100000	0.101327389	0
icpo[21]	6.757	1.25	100000	0.147994672	0
icpo[22]	7.499	1.336	100000	0.133351113	0
icpo[23]	6.603	1.057	100000	0.151446312	0
icpo[24]	27.81	927.4	100000	0.035958288	0
icpo[25]	6.879	1.432	100000	0.145369967	0
icpo[26]	6.859	1.278	100000	0.145793847	0

So sample 15 and 18 are outliers.

(c)

For new data point,

$$x_n = [1, 10, 5, 5]$$

$$\hat{y}_n = x_n \times \beta$$

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-p}$$

$$SE\{\hat{y}_n\} = \sqrt{MSE} \times \sqrt{1 + x_n^T (X^T X) x_n}$$

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

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

$$[30.1487, 44.7874]$$

Matlab Code

```
% problem 3
%(a)
data = importdata('hockpend.dat');
x1 = data(:,1);
x2 = data(:,2);
x3 = data(:,3);
Y = data(:,4);

vecones = ones(size(Y));
X =[vecones x1 x2 x3];
```

```
[n, p] = size(X);
b = inv(X' * X) * X'* Y;  % [8.855;3.420;-1.451;0.334]
H = X * inv(X' * X) * X';
Yhat=H*Y; %or Yhat =X*b;
J=ones(n); I = eye(n);
SSR = Y' * (H - 1/n * J) * Y;
SSE = Y' * (I - H) * Y;
SST = Y' * (I - 1/n * J) * Y;
MSR = SSR/(p-1);
MSE = SSE/(n-p);
F = MSR/MSE;
pval = 1-fcdf(F, p-1, n-p);
Rsq = 1 - SSE/SST;  % 0.8628
Rsqadj = 1 - (n-1)/(n-p) * SSE/SST;
s = sqrt(MSE);

%(c)
Xh=[1, 10, 5, 5];
Yh=Xh*b; % 37.4681
sig2h=MSE* Xh *inv(X'*X) *Xh';
sig2hpre=MSE*(1+Xh *inv(X'*X) *Xh');
sigh = sqrt(sig2h);
sighpre = sqrt(sig2hpre);
%95% CI,Äôs on the individual responses
[Yh-tinv(0.975, n-p)*sighpre, Yh+tinv(0.975, n-p)*sighpre]
% 30.1487   44.7874
```

OpenBUGS code

```
A: Model
model {
for (i in 1:26) {
y[i] ~ dnorm(m[i],tau)
m[i] <- b[1]+b[2]*x1[i]+b[3]*x2[i]+b[4]*x3[i]
r[i] <- y[i]-m[i]
f[i] <- sqrt(tau/6.2832)*exp(-0.5*tau*r[i]*r[i])   #2*pi approx 6.2832
icpo[i] <- 1/f[i]}
# take inverses of average (over a smulation run) of icpo
# to get estimate of CPO (outside WinBUGS)
for (j in 1:4) {b[j] ~ dnorm(0,0.00001)}
tau ~ dgamma(1,0.001)
s2 <- 1/tau}

A: Data
list(x1=c(12.98,14.295,15.531,15.133,15.342,17.149,15.462,12.801,17.039,13.172,16.125,14.34,12.923,14.231,15.222,15.74,14.958,14.125,16.391,16.452,13.535,14.199,15.837,16.565,13.322,15.949),
x2=c(0.317,2.028,5.305,4.738,7.038,5.982,2.737,10.663,5.132,2.039,2.271,4.077,2.643,10.401,1.22,10.612,4.815,3.153,9.698,3.912,7.625,4.474,5.753,8.546,8.598,8.29),
x3=c(9.998,6.776,2.947,4.201,2.053,-0.055,4.657,3.048,0.257,8.738,2.101,5.545,9.331,1.041,6.149,-1.691,4.111,8.453,-1.714,2.145,3.851,5.112,2.087,8.974,4.011,-0.248),

y=c(57.702,59.295,55.166,55.767,51.722,60.446,60.715,37.447,60.974,55.27,59.289,54.027,53.199,41.896,53.254,45.798,58.699,50.086,48.89,62.213,45.625,53.923,55.799,56.741,43.145,50.706
))

A: Inits
list(tau=1,b=c(0,0,0,0))
```

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
icpo[1]	9.252	4.814	0.06426	5.395	7.961	20.98	1	100000
icpo[2]	10.43	3.618	0.0404	6.338	9.556	19.65	1	100000
icpo[3]	6.465	1.006	0.009974	4.852	6.343	8.787	1	100000
icpo[4]	6.662	1.035	0.009697	5.008	6.533	9.043	1	100000
icpo[5]	6.481	1.016	0.009727	4.858	6.357	8.824	1	100000
icpo[6]	8.859	3.388	0.06029	5.497	8.006	17.39	1	100000
icpo[7]	7.819	1.724	0.01708	5.466	7.494	12.03	1	100000
icpo[8]	7.996	4.349	0.0699	5.087	7.122	16.26	1	100000
icpo[9]	7.868	2.413	0.03552	5.235	7.31	13.84	1	100000
icpo[10]	8.265	2.705	0.04222	5.371	7.605	15.15	1	100000
icpo[11]	10.88	6.148	0.06737	5.83	9.204	25.93	1	100000
icpo[12]	6.501	1.021	0.0107	4.871	6.375	8.855	1	100000
icpo[13]	7.418	2.005	0.02481	5.109	6.981	12.4	1	100000
icpo[14]	7.485	2.152	0.03093	5.105	7.008	12.76	1	100000
icpo[15]	20830.0	951200.0	2997.0	72.81	1110.0	79330.0	1	100000
icpo[16]	7.568	2.247	0.01504	5.113	7.054	13.26	1	100000
icpo[17]	34.41	21.55	0.1405	14.3	28.58	89.96	1	100000
icpo[18]	124.7	324.5	1.544	16.6	64.09	586.7	1	100000
icpo[19]	8.28	2.858	0.01832	5.326	7.566	15.68	1	100000
icpo[20]	9.869	3.589	0.05312	5.989	8.973	19.02	1	100000
icpo[21]	6.757	1.25	0.02085	4.938	6.563	9.71	1	100000
icpo[22]	7.499	1.336	0.02375	5.478	7.298	10.68	1	100000
icpo[23]	6.603	1.057	0.01065	4.934	6.47	9.055	1	100000
icpo[24]	27.81	927.4	8.06	5.214	8.171	84.66	1	100000
icpo[25]	6.879	1.432	0.02106	4.954	6.631	10.28	1	100000
icpo[26]	6.859	1.278	0.01049	4.992	6.656	9.899	1	100000

CPO=c(0.108084738,0.095877277,0.154679041,0.150105074,0.154297176,0.112879558,0.127893593,0.125062531,0.127097102,0.120992136,0.091911765,0.153822489,0.134807226,0.133600534,4.80077E-05,0.132135307,0.029061319,0.008019246,0.120772947,0.101327389,0.147994672,0.133351113,0.151446312,0.035958288,0.145369967,0.145793847)

Reference:

Textbook: ENGINEERING BIOSTATISTICS by Brani Vidakovic