Bayesian Logistic Regression

SP4R01d01.sas

Babies with low birth weights (defined to be less than 2500 grams) are a concern because of their potential medical problems. Health researchers want to identify possible contributing factors to low birth weight and recommend strategies to reduce the number of low birth weight babies. The data is named **birth** and includes the following variables:

ID	Identification code
LOW	Low birth weight (0 = birth weight >=2500g and 1 = birth weight <2500g)
AGE	Age of mother in years
LWT	Weight in pounds at the last menstrual period
ETH	Ethnicity
SMOKE	Smoking status during pregnancy (1 = Yes, 0 = No)
PTL	History of premature labor (0 = None, 1 = One, and so on)
HT	History of hypertension (1 = Yes, 0 = No)
UI	Presence of uterine irritability (1 = Yes, 0 = No)
FTV	Number of physician visits during the first trimester (0 = None, 1 = One, and so on)
BWT	Birth weight in grams

1. Read in the **birth** data set. Use a DATA step and create formats and labels to create a more informative analysis.

Partial DATA Step Code

```
proc format;
   value yesnofmt
        0="No"
        1="Yes";
   value ftvfmt
        0="0"
        1="1"
        2-high="2+";
   value ptlfmt
        0="0"
        1-high="1+";
run;

/*
LIST OF VARIABLES:

Columns    Variable                                          Abbreviation
-------------------------------------------------------------------------
2-4     Identification Code                                      ID

10      Low Birth Weight (0 = Birth Weight >= 2500g,             LOW
```

```
                              1 = Birth Weight < 2500g)

17-18   Age of the Mother in Years                                    AGE

23-25   Weight in Pounds at the Last Menstrual Period                 LWT

32      Ethnicity                                                     ETH

40      Smoking Status During Pregnancy (1 = Yes, 0 = No)             SMOKE

48      History of Premature Labor (0 = None  1 = One, etc.)         PTL

55      History of Hypertension (1 = Yes, 0 = No)                     HT

61      Presence of Uterine Irritability (1 = Yes, 0 = No)            UI

67      Number of Physician Visits During the First Trimester         FTV
              (0 = None, 1 = One, 2 = Two, etc.)

73-76   Birth Weight in Grams                                         BWT
--------------------------------------------------------------------------
*/

data work.birth;
   input ID LOW AGE LWT ETH SMOKE PTL HT UI FTV BWT;
   if FTV>2 then FTV=2;
   if PTL>1 then PTL=1;
   label
      ID="ID Code"
      LOW="Birth Weight < 2500 Grams"
      AGE="Mom's Age"
      LWT="Mom's Weight Last Menstrual Period"
      ETH="Ethnicity"
      SMOKE="Smoking Status"
      PTL="Hx of Premature Labor"
      HT="Hx of Hypertension"
      UI="Hx of Uterine Irritability"
      FTV="MD Visits 1st Trimester"
      BWT="Birth Weight, Grams"
   ;
   format LOW SMOKE HT UI yesnofmt. PTL ptlfmt. ftv ftvfmt.;
   datalines;
85    0     19    182    2       0       0      0     1     0     2523
86    0     33    155    3       0       0      0     0     3     2551
...
   ;
run;
```

2. Generate summary statistics for the **BWT** variable along with a histogram and QQPlot.

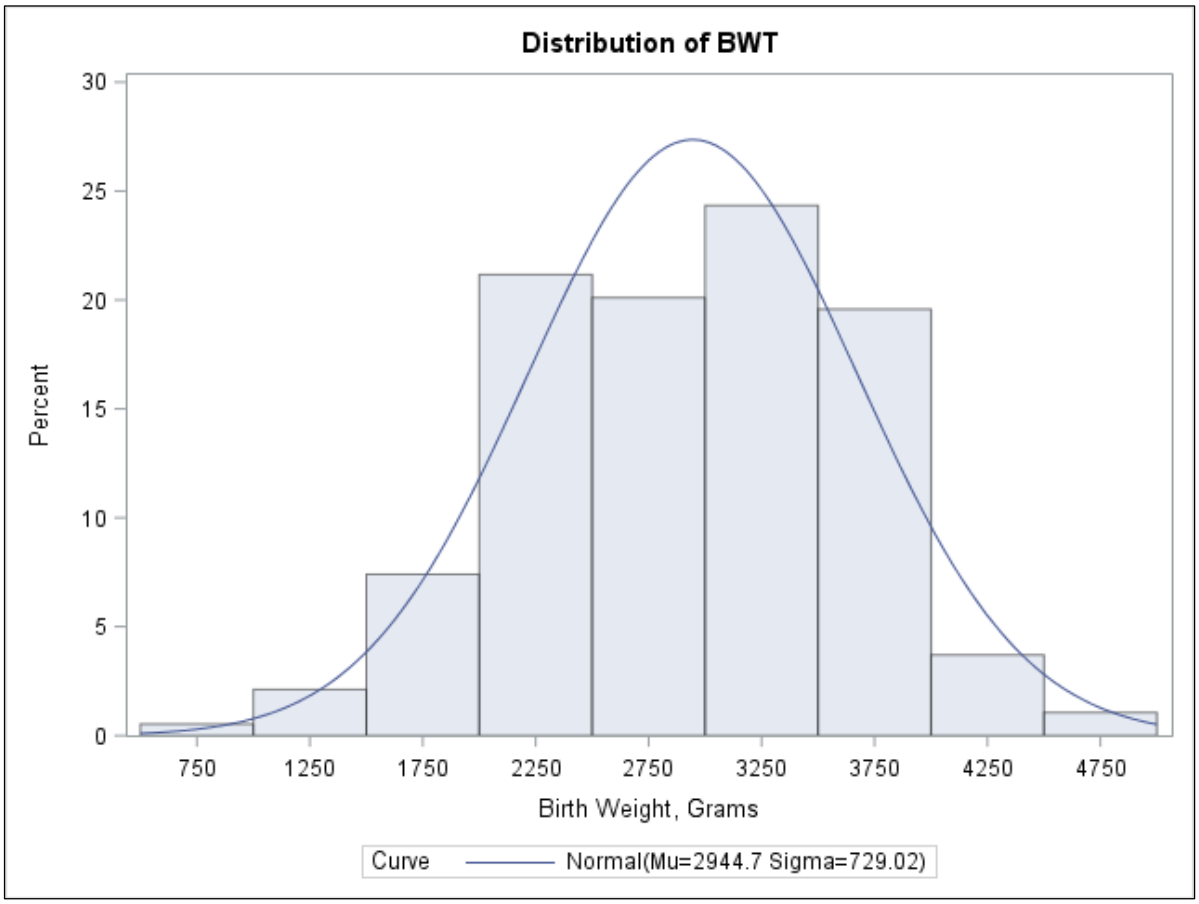
```
ods select basicmeasures histogram qqplot;
proc univariate data=work.birth;
   var bwt;
   histogram bwt / normal(mu=est sigma=est);
   qqplot bwt / normal(mu=est sigma=est);
run;
```

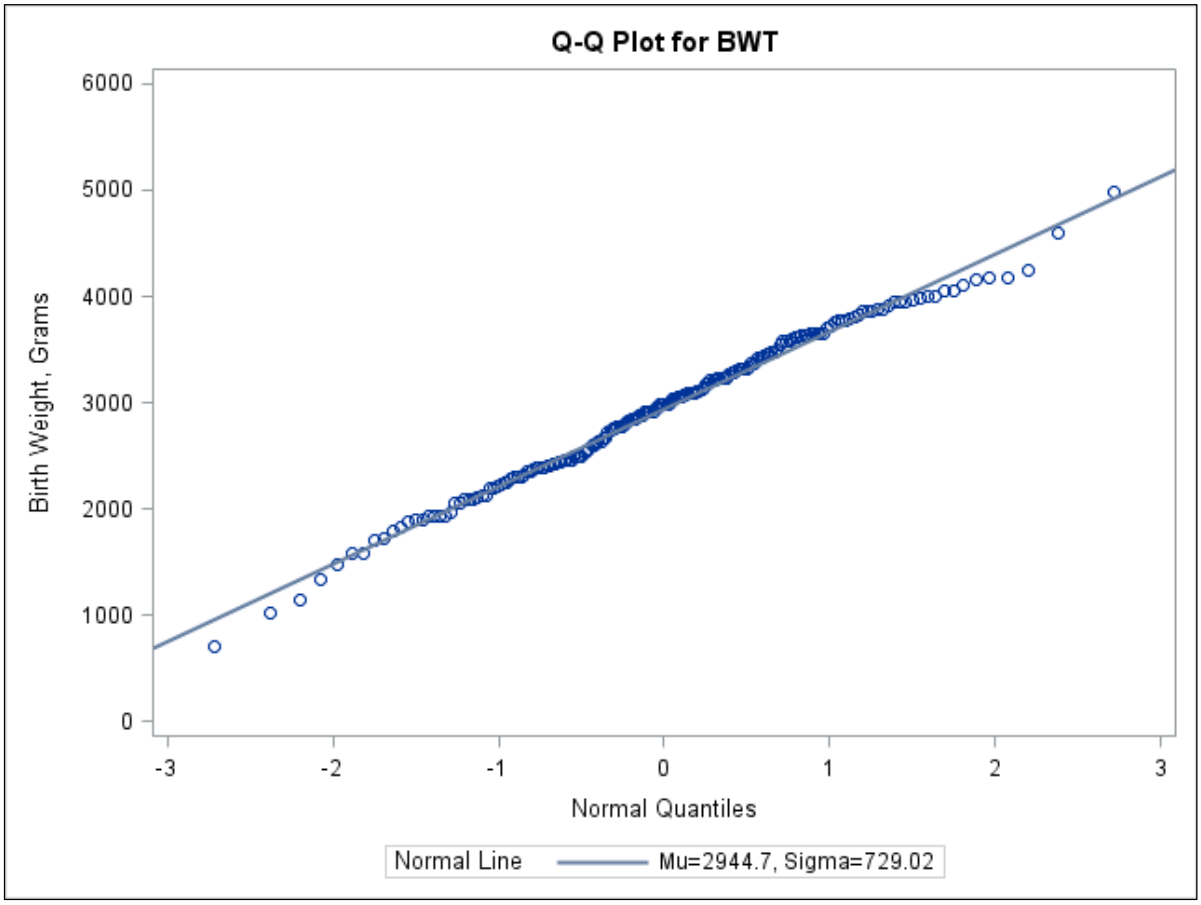
The UNIVARIATE Procedure
Variable: BWT (Birth Weight, Grams)

Basic Statistical Measures

Location		Variability	
Mean	2944.656	Std Deviation	729.02242
Median	2977.000	Variance	531474
Mode	2495.000	Range	4281
		Interquartile Range	1061

Note: The mode displayed is the smallest of 4 modes with a count of 4.





3. Create univariate contingency tables for the variables **Low**, **Smoke**, **HT**, and **PTL**.

```
proc freq data=work.birth;
   table low smoke ht ptl;
run;
```

```
                        The FREQ Procedure

                    Birth Weight < 2500 Grams

                                                Cumulative    Cumulative
LOW      Frequency     Percent     Frequency      Percent
-----------------------------------------------------------
No             130       68.78           130        68.78
Yes             59       31.22           189       100.00


                          Smoking Status

                                                Cumulative    Cumulative
SMOKE    Frequency     Percent     Frequency      Percent
-----------------------------------------------------------
No             115       60.85           115        60.85
Yes             74       39.15           189       100.00
```

Hx of Hypertension

HT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	177	93.65	177	93.65
Yes	12	6.35	189	100.00

Hx of Premature Labor

PTL	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	159	84.13	159	84.13
1+	30	15.87	189	100.00

4. Use the SAS MCMC procedure (Markov Chain Monte Carlo) to create a Bayesian logistic regression model with LOW as the dependent variable and SMOKE, HT, LWT, and PTL as the independent variables.

```
ods select nobs parameters postsummaries postintervals autocorr
tadpanel;
proc mcmc data=work.birth outpost=birthout diag=all dic propcov=quanew
   nbi=5000 ntu=5000 nmc=200000 thin=10 mchistory=brief
      plots(smooth)=all seed=27513 stats=all;
   parms (beta0 beta1 beta2 beta3 beta4) 0;
   prior beta: ~ normal(0, var=100);
   p=logistic(beta0+beta1*smoke+beta2*ht+ beta3*lwt+beta4*ptl);
   model low ~ binary(p);
   title "Bayesian Analysis of Low Birth Weight Data";
run;
title;
```

The PARMS statement specifies the five parameters with initial values of 0. The PRIOR statement specifies a normal prior distribution with a mean of 0 and a variance of 100 for each parameter. The *p* assignment statement computes the probability of low birth weight using the parameter estimates, data values, and the logit link transformation (with the SAS function LOGISTIC). The MODEL statement specifies that the response variable **low** has a binary distribution with a parameter *p*.

Selected PROC MCMC statement options:

OUTPOST= specifies an output data set that contains the posterior samples of all model parameters.

PROPCOV= specifies the method used in constructing the initial covariance matrix for the Metropolis-Hastings algorithm. The quasi-Newton optimization (QUANEW) and the Nelder-Mead simplex optimization (NMSIMP) methods find numerically approximated covariance matrices at the optimum of the posterior density function with respect to all continuous parameters. The optimization does not apply to discrete parameters. The tuning phase starts at the optimized values. In some problems, this can greatly increase convergence performance.

NTU= specifies the number of iterations to use in each proposal tuning phase. By default, NTU=500.

NMC= specifies the number of iterations in the main simulation loop. This is the MCMC sample size if there is no thinning. By default, NMC=1000.

THIN=*n* controls the thinning rate of the simulation. PROC MCMC keeps every nth simulation sample and discards the rest. All of the posterior statistics and diagnostics are calculated using the thinned samples. By default, THIN=1.

NBI= *n* specifies the number of burn-in iterations to perform before beginning to save parameter estimate chains. By default, NBI=1000.

MCHISTORY= controls the display of the Markov chain sampling history. The keyword BRIEF produces a summary output for the tuning, burn-in, and sampling history tables.

STATS= specifies options for posterior statistics. You can request all of the posterior statistics by specifying STATS=ALL. You can suppress all the calculations by specifying STATS=NONE.

Partial PROC MCMC Output

The MCMC Procedure

Number of Observations Read 189
Number of Observations Used 189

Parameters

Block	Parameter	Sampling Method	Initial Value	Prior Distribution
1	beta0	N-Metropolis	0	normal(0, var=100)
	beta1		0	normal(0, var=100)
	beta2		0	normal(0, var=100)
	beta3		0	normal(0, var=100)
	beta4		0	normal(0, var=100)

The first table that PROC MCMC produces is the Number of Observations table. This table lists the number of observations read from the input data set and the number of nonmissing observations used in the analysis. The Parameters table lists the names of the parameters, the blocking information, the sampling method used, the starting values, and the prior distributions. You should check this table to ensure that you have specified the parameters correctly, especially for complicated models.

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25	50	75
beta0	20000	0.8812	0.8756	0.2758	0.8693	1.4598
beta1	20000	0.5124	0.3461	0.2785	0.5111	0.7460
beta2	20000	1.9292	0.7443	1.4288	1.9108	2.4112
beta3	20000	-0.0179	0.00685	-0.0224	-0.0178	-0.0132
beta4	20000	1.3191	0.4447	1.0220	1.3155	1.6145

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
beta0	0.050	-0.7734	2.6647	-0.8075	2.6183
beta1	0.050	-0.1665	1.1824	-0.1533	1.1941

beta2	0.050	0.5218	3.4477	0.4763	3.3879
beta3	0.050	-0.0321	-0.00522	-0.0317	-0.00493
beta4	0.050	0.4448	2.1960	0.4384	2.1874

For each posterior distribution, PROC MCMC also reports summary statistics (posterior means and standard deviations) and interval statistics (95% highest posterior density credible intervals).

Posterior Autocorrelations

Parameter	Lag 1	Lag 5	Lag 10	Lag 50
beta0	0.2951	-0.0036	-0.0017	-0.0118
beta1	0.2993	-0.0027	0.0011	0.0055
beta2	0.3186	-0.0037	-0.0103	-0.0190
beta3	0.2975	-0.0037	-0.0060	-0.0134
beta4	0.3199	-0.0070	0.0058	0.0007

Partial Graphics Output