Bayesian Regression Model For Predicting Income

—Bayesian Statistics Project Final Report

Hai Liu, Yi-Lung Kuo, Yi He, Jin Liu

December 3, 2006

Introduction

Background

Undoubtedly, to some extent, income is considered as one index of a successful life. It is generally believed that different factors would influence the amount of income earned after leaving school, such as intelligence, gender, race, and educational levels. For example, with longitudinal data taken from National Longitudinal Survey of Youth (NLSY), Murray (1997, 1998) categorized intelligence into five levels from very dull to very bright, and illustrated that adults with higher intelligence would have higher income. Terman spent most of his life conducting the landmark longitudinal study of gifted students for almost forty years from 1921 (Terman & Oden, 1959). He found that earned income of male gifted adults was higher than that of female gifted adults. Another longitudinal study of gifted students (Subotnik, Karp, & Morgan, 1989) showed the same result. Additionally, Terman's study revealed that different educational levels conditional on gender would result in different amounts of income. In short, although this study was confined to the group of gifted adults, it implied that gender and educational level might be factors influencing income in the adulthood.

The 1998 data from U.S. Census Bureau renders us some indication that demographic information, such as gender, race, and educational levels, might have effects on yearly income. Tables 1 through 4 in the appendix show mean income of the year 1998 for all applicable people in the United States. These data will be used to set informative priors when Bayesian regression model is used in our study.

Our research questions is: how would these different factors affect the income on average? Would the gifted tend to earn more than the non-gifted? Here by gifted we refer to those who have ever been in a gifted/talented program.

Dataset

The data to be used in our project is from the National Education Longitudinal Study (NELS 1988-2000), which was conducted by the National Center for Education Statistics (NCES). The data collection experienced five waves, including the base year in the spring of 1988 (BY), and four follow-ups in 1990(F1), 1992(F2), 1994(F3), and 2000(F4). The subjects were nationally representatively sampled eighth-grade students who were surveyed in the base year and from F1 to F4. The major topics in the questionnaire consisted of school, work, and home experiences; educational resources and support; the role in education of their parents and peers; neighborhood characteristics; educational and occupational aspirations;

and other student perceptions. We are going to just pick some variables from this huge data set to do our statistical analysis under the Bayesian framework.

Model Specification

Our purpose is to set up a model to find if gifted students are more likely to have higher income than the non-gifted and if gender, race, education level, and working hours will have significant influence on income. So in our model, we use income as the response variable, which could be treated as a continuous random variable. The predictor variables are included are:

Gifted Ever enrolled in a gifted program

Gender Male/Female

Race Race of respondent

1. American Indian or Alaska Native
2. Asian or Pacific Islander
3. Black, not Hispanic
4. White, not Hispanic
5. Hispanic or Latino
6. More than one race

Education Highest PSE degree attained as of 2000

1. Some PSE, no degree attained
2. Certificate/license
3. Associate's degree
4. Bachelor's degree
5. Master's degree/equivalent
6. Ph.D or a professional degree

Working Hours Number of hours the respondents worked in the year of 1999

The working hours can be treated as a continuous variable. All the other predictor variables are categorical, so just like the frequentist approach, dummy variables are used in setting up the model in WinBUGS. Another issue is that after checking the data we find that the response income is very skewed to the left. So before fitting the model we should first make some transformation to the income variable. A natural choice is to use the logarithm of income instead of raw income as the response. So the model has the form:

$$\log(\texttt{Income}) \sim \texttt{1 + Gifted + Gender + Race + Education + Working Hours}$$

Because an overall intercept is included in the model, the number of dummy variables within each predictor should be one fewer than the level of that predictor. For instance, there are 6 levels of education as described before, so there should be 5 dummy variables associated with the predictor education.

First, noninformative priors for all parameters are used and then some informative priors based on previous studies are considered. Also the parameters of the dummy variables are considered to be independent. The model specification in WinBUGS is listed in the appendix.

Some Computing Output

Noninformative Prior

The following is some computing output from the model with noninformative priors. Notice that the response is log(`Income`).

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
a	4.67E-4	8.26E-6	1.266E-7	4.51E-4	4.67E-4	4.83E-4	1001	5000
alpha	9.0840	0.05768	8.912E-4	8.973	9.084	9.198	1001	5000
beta	0.0545	0.01091	1.516E-4	0.03271	0.05463	0.0762	1001	5000
e2	-0.0160	0.01628	2.331E-4	-0.04755	-0.0162	0.0162	1001	5000
e3	0.0519	0.01650	2.046E-4	0.01944	0.05191	0.0843	1001	5000
e4	0.2055	0.01058	1.483E-4	0.1849	0.2056	0.226	1001	5000
e5	0.2637	0.02261	2.981E-4	0.2188	0.2634	0.3089	1001	5000
e6	0.6288	0.05971	8.349E-4	0.5139	0.6281	0.7475	1001	5000
gamma	-0.1802	0.00922	1.263E-4	-0.1984	-0.1803	-0.1623	1001	5000
r2	0.3612	0.05765	9.484E-4	0.2509	0.3611	0.4749	1001	5000
r3	0.1832	0.05653	9.74E-4	0.07506	0.1825	0.2959	1001	5000
r4	0.2621	0.05481	8.718E-4	0.1571	0.2613	0.3706	1001	5000
r5	0.2362	0.05601	8.972E-4	0.127	0.2359	0.3461	1001	5000
r6	0.1933	0.06054	0.001005	0.07671	0.1927	0.3094	1001	5000
sigma	0.3176	0.00319	4.766E-5	0.3116	0.3175	0.3242	1001	5000

`Alpha` is the overall intercept, `beta` is the coefficient associated with gifted, and `gamma` with female; `e[i]`'s are the coefficients associated with education; `r[i]`'s associated with race; `sigma` is the variance of the random error; and `a` is the coefficient of the only continuous predictor working hours.

Only one chain is run. There are totally 6000 posterior samples and the first 1000 iterations are discarded as the burn-in stage by checking the history plots of all the parameters. Also the autoregressive plots look good.

Some interesting conclusions could be obtained by looking at the output. The gifted tend to have a little higher income than the non-gifted; Males are likely to make much more money(about 1.2 times) than females. The income tends to increase as the education level increases and not surprisingly, those who got a bachelor's degree or higher seem to have much higher income than those with lower education level. Also the race will influence the income somehow. Especially the Asian or Pacific Islander seem to make more money than other races. Most the 95% confidence intervals exclude 0 except `e2` which means in some sense most effects we are interested in are significant. And also, one's income significantly increases as his working hours increases. Hence the conclusion based on this model is consistent with what we expect.

Informative Prior

We have some information about the mean income distribution of US in the year of 1998 as described in the introduction. That might help us give some informative priors to the

parameters we are trying to estimate. In particular, from those information, we know that males earn almost twice as much as females on average; those having higher IQ tend to have higher average income; those with higher level of education usually have more income; and different ethnicities have different mean income. So we use these to build a second Bayesian regression model with some informative priors. It turns out that the results do not change very much. That is because the dataset that out model is based on is so huge with over $5,000$ observations that prior information will only contribute a little to the model fitting compared to the information contained in the data.

Because there is very little difference between the models with informative and noninformative priors, we still stick with the noninformative prior model to do the following model checking and model comparison.

Model Checking

DIC

After fitting the regression model, the next step is to check whether this model is good or not in terms of how well we can use it to do prediction. Before that we may still be not quite sure if the model is overparametrized, e.g., if we could remove some predictors and still get a good model. So first we compare some possible nested models using *DIC*.

```
Full Model:
                      Dbar        Dhat        pD       DIC
          y       2738.190    2723.210    14.980  2753.170
          total   2738.190    2723.210    14.980  2753.170

Model without gender:
                      Dbar        Dhat        pD       DIC
          y       3108.550    3094.550    14.002  3122.550
          total   3108.550    3094.550    14.002  3122.550

Model without gifted:
                      Dbar        Dhat        pD       DIC
          y       2762.030    2748.020    14.002  2776.030
          total   2762.030    2748.020    14.002  2776.030
```

We only list some of the reduced models. Hence based on *DIC* the full model is the best.

Prediction

In this section we try to use our regression model to make predictions for the all respondents in our dataset. If our model is a good one, the predicted values should look similar to the true response log(`Income`). *R* is used to generate simulated dataset with some replications. For each simulated dataset several quantiles are calculated and then the mean quantiles through all replications are compared to those from the true dataset. Here are some results:

	1%	25%	50%	75%	99%
True	9.210	10.043	10.309	10.593	11.513
Predicted	9.229	10.018	10.313	10.602	11.324

We could see that the simulated quantiles are very close to the true values, which means our model performs well in prediction. The code for the simulation is also listed in the Appendix.

Model Comparison with Frequentist

Finally, we want to compare the Bayesian regression model with the model under the frequentist framework. The frequentist model is fitted using *R*. The same model structure is used although by default in *R* some dummy variables may be chosen slightly differently.

R output:

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            9.137e+00  5.828e-02 156.792  < 2e-16 ***
as.factor(Gifted)2    -5.446e-02  1.093e-02  -4.981 6.54e-07 ***
as.factor(Sex)2       -1.801e-01  9.190e-03 -19.603  < 2e-16 ***
as.factor(Race)2       3.629e-01  5.783e-02   6.276 3.77e-10 ***
as.factor(Race)3       1.846e-01  5.679e-02   3.250  0.00116 **
as.factor(Race)4       2.635e-01  5.481e-02   4.807 1.58e-06 ***
as.factor(Race)5       2.374e-01  5.596e-02   4.241 2.26e-05 ***
as.factor(Race)6       1.949e-01  6.081e-02   3.205  0.00136 **
as.factor(Education)2 -1.563e-02  1.654e-02  -0.945  0.34454
as.factor(Education)3  5.181e-02  1.651e-02   3.138  0.00171 **
as.factor(Education)4  2.056e-01  1.058e-02  19.428  < 2e-16 ***
as.factor(Education)5  2.638e-01  2.245e-02  11.752  < 2e-16 ***
as.factor(Education)6  6.280e-01  5.871e-02  10.696  < 2e-16 ***
WH                     4.667e-04  8.483e-06  55.013  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3175 on 5024 degrees of freedom
Multiple R-Squared: 0.4849,     Adjusted R-squared: 0.4836
F-statistic: 363.8 on 13 and 5024 DF,  p-value: < 2.2e-16
```

Compared to the previous output from WinBUGS, all the coefficients are very close to each other. For example, the gifted effect is .0545 under both models. The overall intercepts are different because the baselines are chosen differently. Again as in the Bayesian model, the level 2 effect with education is not significant.

Therefore, the results from both Bayesian and frequentist approaches are quite consistent with each other, which is not surprising because both models use the same data and same structure, just treating the parameters(coefficients) from two different points of view.

Appendix A

Table 1: IQ and Income

IQ	Median income in 1992 (dollars)
Very Bright (125+)	36,000
Bright (110-124)	27,000
Normal (90-109)	20,000
Dull (75-89)	12,400
Very Dull (less than 75)	5,000

Source: Data are from Murray (1997)

Table 2: Race and Income

Race	Mean Income (dollars)
White	30,391
Black	20,609
Asian and Pacific Islander	28,748
Hispanic	20,105

Source: U.S. Census Bureau

Table 3: Educational levels and Income

Education	Mean Income (dollars)
Less than 9th Grade	1,3431
9th to 12th Grade (No Diploma)	1,6576
High School Graduate (Includes Equivalency)	2,3363
Some College, No Degree	2,9643
Associate Degree	3,2759
bachelor's degree	4,4962
master's degree	5,4421
Professional Degree	9,3358
Doctorate Degree	7,7195

Source: U.S. Census Bureau

Table 4: Gender and Income

Gender	Mean Income (dollars)
Male	24,054
Female	12,311

Source: U.S. Census Bureau

Appendix B

WinBUGS program:

```
## Regression Model

# Non-Informative Priors

model {
  for (i in 1:N) {
       y[i] ~ dnorm(mu[i], tau)
       ypred[i] ~ dnorm(mu[i], tau)
       resid[i] <- abs(y[i] - mu[i])
       respred[i] <- abs(ypred[i] - mu[i])
       mu[i] <- alpha + beta*gifted[i] + gamma*female[i]
                + r2*race2[i] + r3*race3[i] + r4*race4[i]
                + r5*race5[i] + r6*race6[i] + e2*edu2[i] + e3*edu3[i]
                + e4*edu4[i] + e5*edu5[i] + e6*edu6[i] + a*WH[i]
}
  sigma <- 1/sqrt(tau)
  alpha ~ dnorm(0, 1.0E-6)
  beta ~ dnorm(0, 1.0E-6)
  gamma ~ dnorm(0, 1.0E-6)
  r6 ~ dnorm(0, 1.0E-6)
  r2 ~ dnorm(0, 1.0E-6)
  r3 ~ dnorm(0, 1.0E-6)
  r4 ~ dnorm(0, 1.0E-6)
  r5 ~ dnorm(0, 1.0E-6)
  e6 ~ dnorm(0, 1.0E-6)
  e2 ~ dnorm(0, 1.0E-6)
  e3 ~ dnorm(0, 1.0E-6)
  e4 ~ dnorm(0, 1.0E-6)
  e5 ~ dnorm(0, 1.0E-6)
  a ~ dnorm(0, 1.0E-6)
  tau  ~ dgamma(0.001, 0.001)
}

# Informative Priors

model {
  for (i in 1:N) {
       y[i] ~ dnorm(mu[i], tau)
       ypred[i] ~ dnorm(mu[i], tau)
       resid[i] <- abs(y[i] - mu[i])
```

```
        respred[i] <- abs(ypred[i] - mu[i])
        mu[i] <- alpha + beta*gifted[i] + gamma*female[i]
                 + r2*race2[i] + r3*race3[i] + r4*race4[i]
                 + r5*race5[i] + r6*race6[i] + e2*edu2[i] + e3*edu3[i]
                 + e4*edu4[i] + e5*edu5[i] + e6*edu6[i] + a*WH[i]
}
  sigma <- 1/sqrt(tau)
  alpha ~ dnorm(0, 1.0E-6)
  beta ~ dnorm(0.40, 2.5)
  gamma ~ dnorm(-0.69, 2.8)
  r6 ~ dnorm(0, 1.0E-6)
  r2 ~ dnorm(0.37, 5)
  r3 ~ dnorm(0, 1.0E-6)
  r4 ~ dnorm(0.37, 5)
  r5 ~ dnorm(0, 1.0E-6)
  e6 ~ dnorm(1.5, 1.3)
  e2 ~ dnorm(0, 1.0E-6)
  e3 ~ dnorm(0, 1.0E-6)
  e4 ~ dnorm(1.0, 2)
  e5 ~ dnorm(1.2, 1.7)
  a ~ dnorm(0, 1.0E-6)
  tau  ~ dgamma(0.001, 0.001)
}
```

R Code:

```
model.full <- lm(log(Income) ~ as.factor(Gifted) + as.factor(Sex)
 + as.factor(Race) + as.factor(Education) + WH, data = gift)

rep <- 2000
q01 <- q25 <- q50 <- q75 <- q99 <- NULL

for(ii in 1:rep) {
  pred <- X %*% b + rnorm(5038, 0, sigma)
  q01[ii] <- quantile(pred, .01)
  q25[ii] <- quantile(pred, .25)
  q50[ii] <- quantile(pred, .50)
  q75[ii] <- quantile(pred, .75)
  q99[ii] <- quantile(pred, .99)
}

list(mean(q01), mean(q25),mean(q50),mean(q75),mean(q99))
list(quantile(y, c(.01,.25,.5,.75,.99)))
```

References

[1] Barringer, H. R., Takeuchi, D. T., Xenos, P. (1990). Educaiton, occupational prestige, and income of Asian Americans. Sociology of Education, 63(1), 27-43.

[2] Murray, C. (1997). IQ and economics success. Public Interest, summer 1997

[3] Murray, C. (1998). Income, inequality, and IQ. Washington, DC: American enterprise Institute.

[4] Spiegelhalter D J, Best N G, Carlin B P and van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J. Roy. Statist. Soc. B. 64, 583-640.

[5] Subotnik, R., Karp, D., Morgan, E. (1989). High IQ children at midlife: An investigation into the generalizability of Terman's genetic studies. Roeper Review, 11(3), 139-144.

[6] Terman, L. M., Oden, M. H. (Eds). (1959). Genetic studies of genius: V. The gifted group at mid-life. Stanford, CA: Stanford, CA: Stanford University Press.