ISYE 6420
Bayesian Statistics
Abhijeet Gaurav
Dec 2, 2019

INVESTIGATING RELATIONSHIPS BETWEEN AVERAGE WEEKLY EARNINGS OF FOREIGN BORN WORKERS WITH RESPECT TO THEIR LITERACY, ENGLISH SPEAKING ABILITY AND TIME IN UNITED STATES USING BAYESIAN APPROACH.

INTRODUCTION

It’s very interesting to find out, how the income of migrants varies based on their literacy, english speaking ability and their time in the United States.
I was not able to find any recent dataset with information. Therefore, I picked a very old dataset from 1909 by (Higgs, R). In 1909, Bayesian regression wasn’t very popular among the scientific community because of the complexity involved in calculating likelihood and multiplying with prior for huge datasets. At that time, we didn’t have amazing Bayesian tools such as WinBUGS, PYMC, MATLAB, R, etc. for applying Bayesian regression easily. Therefore, I thought it would be very interesting to explore Bayesian regression on this old dataset (Higgs, R) and compare the findings with the approach taken in the original paper.

The paper (Higgs, R) has a dataset which has information on ‘Characteristic of an adult, male, for foreign-born workers in mining and manufacturing occupations’. Earnings of males who work in the United States but were originally born in a different country were reported. The data collected which is given in the paper has details on ‘Birth Country’; ‘Average weekly earnings in dollars’; ‘Percentage who speak English language’; ‘Percentage Literate’; ‘Percentage Residing in the United States for 5 years or more’. The goal of the original research in the paper was to determine how ‘Average weekly earnings in dollars’ varied with the other covariates given in the dataset.

In the original paper, the author applied multiple linear regression on this dataset. The dependent variable ‘Average weekly earnings in dollars’ (WE) was found to depend on the explanatory variables ‘Percentage who speak English language’ (PSE) and ‘Percentage Literate’ (PL) in the form of the following estimated equation:

$WE = 2.55 + 0.383 * PSE + 0.0796 * PL$

In this project, my approach would be to determine how Bayesian regression performed on this dataset and how the coefficients calculated differ from the coefficients found in the original paper (Higgs, R). I would be doing Bayesian Multiple Regression to be precise. The Bayesian regression would help in this case because:

1) Bayesian models are more flexible, and they handle more complex models.
2) Bayesian model selection is superior (BIC/AIC).
3) Bayesian hierarchical models are easier to extend to many levels.
4) Bayesian analysis is more accurate for small samples.
5) Bayesian models can incorporate prior information.

To apply Bayesian regression to this dataset, we’ll do the following:

1) We'll let OpenBUGS figure out the likelihood function of the data.
2) We'll choose a prior normal distribution over all unknown parameters.
3) We'll let OpenBUGS to apply Bayes theorem to find the posterior distribution over all parameters.

ORIGINAL DATASET:

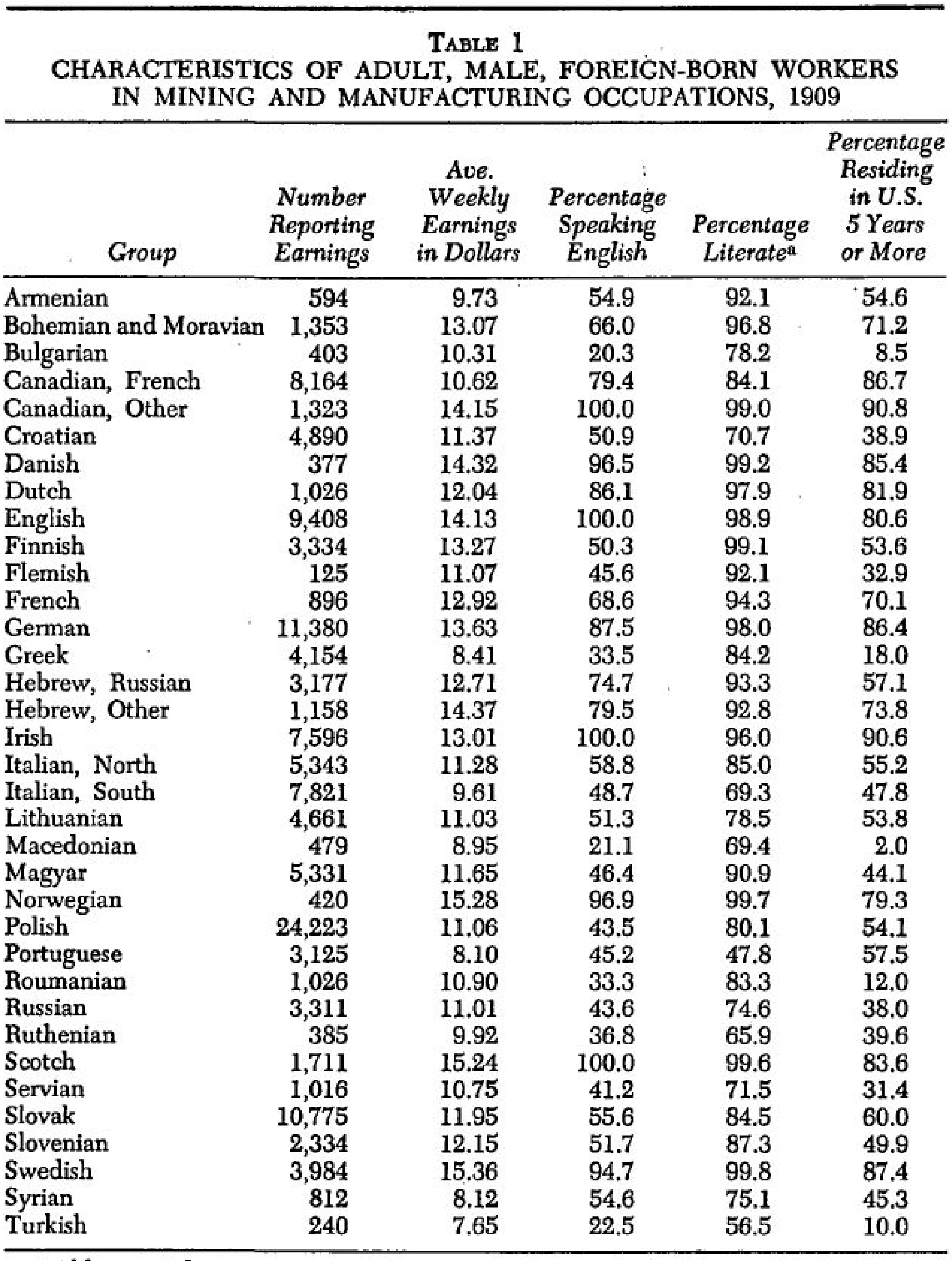


TABLE 1
CHARACTERISTICS OF ADULT, MALE, FOREIGN-BORN WORKERS IN MINING AND MANUFACTURING OCCUPATIONS, 1909

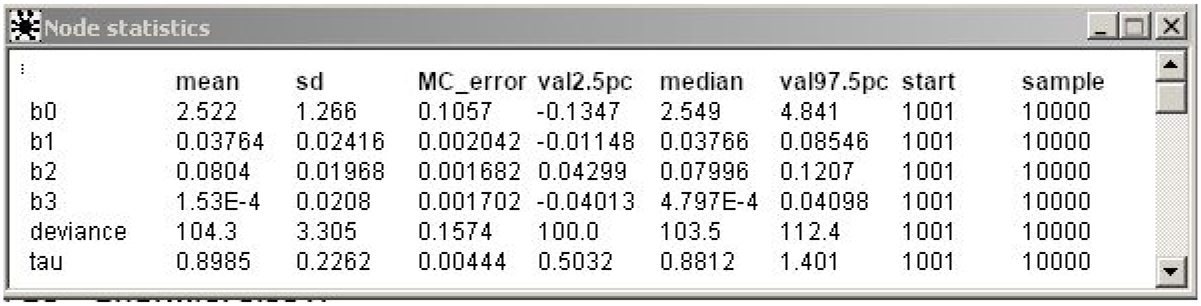
Group	*Number Reporting Earnings*	*Ave. Weekly Earnings in Dollars*	*Percentage Speaking English*	*Percentage Literate*[a]	*Percentage Residing in U.S. 5 Years or More*
Armenian	594	9.73	54.9	92.1	54.6
Bohemian and Moravian	1,353	13.07	66.0	96.8	71.2
Bulgarian	403	10.31	20.3	78.2	8.5
Canadian, French	8,164	10.62	79.4	84.1	86.7
Canadian, Other	1,323	14.15	100.0	99.0	90.8
Croatian	4,890	11.37	50.9	70.7	38.9
Danish	377	14.32	96.5	99.2	85.4
Dutch	1,026	12.04	86.1	97.9	81.9
English	9,408	14.13	100.0	98.9	80.6
Finnish	3,334	13.27	50.3	99.1	53.6
Flemish	125	11.07	45.6	92.1	32.9
French	896	12.92	68.6	94.3	70.1
German	11,380	13.63	87.5	98.0	86.4
Greek	4,154	8.41	33.5	84.2	18.0
Hebrew, Russian	3,177	12.71	74.7	93.3	57.1
Hebrew, Other	1,158	14.37	79.5	92.8	73.8
Irish	7,596	13.01	100.0	96.0	90.6
Italian, North	5,343	11.28	58.8	85.0	55.2
Italian, South	7,821	9.61	48.7	69.3	47.8
Lithuanian	4,661	11.03	51.3	78.5	53.8
Macedonian	479	8.95	21.1	69.4	2.0
Magyar	5,331	11.65	46.4	90.9	44.1
Norwegian	420	15.28	96.9	99.7	79.3
Polish	24,223	11.06	43.5	80.1	54.1
Portuguese	3,125	8.10	45.2	47.8	57.5
Roumanian	1,026	10.90	33.3	83.3	12.0
Russian	3,311	11.01	43.6	74.6	38.0
Ruthenian	385	9.92	36.8	65.9	39.6
Scotch	1,711	15.24	100.0	99.6	83.6
Servian	1,016	10.75	41.2	71.5	31.4
Slovak	10,775	11.95	55.6	84.5	60.0
Slovenian	2,334	12.15	51.7	87.3	49.9
Swedish	3,984	15.36	94.7	99.8	87.4
Syrian	812	8.12	54.6	75.1	45.3
Turkish	240	7.65	22.5	56.5	10.0

APPLYING BAYESIAN REGRESSION TO THIS DATASET

I used the OpenBUGS code that I wrote for this problem (attached the ODC file named ProjectBayesianMLR.odc in the submission).

I used the prior of Normal distribution (0, 0.001) for all b0, b1, b2, and b3. I made tau to use prior of Gamma distribution(0.001, 0.001).

With my Beyasian code applied on this dataset, I first burned in 1000 samples. I found the following results in the form of Node Statistics in the next 10000 samples.



```
            mean      sd        MC_error  val2.5pc  median    val97.5pc start     sample
b0          2.522     1.266     0.1057    -0.1347   2.549     4.841     1001      10000
b1          0.03764   0.02416   0.002042  -0.01148  0.03766   0.08546   1001      10000
b2          0.0804    0.01968   0.001682  0.04299   0.07996   0.1207    1001      10000
b3          1.53E-4   0.0208    0.001702  -0.04013  4.797E-4  0.04098   1001      10000
deviance    104.3     3.305     0.1574    100.0     103.5     112.4     1001      10000
tau         0.8985    0.2262    0.00444   0.5032    0.8812    1.401     1001      10000
```

Using the Bayesian regression, I found that the dependent variable 'Average weekly earnings in dollars' (WE) was dependent on the explanatory variables 'Percentage who speak English language' (PSE); 'Percentage Literate' (PL) but not dependent on 'Percentage Residing in the US for 5 years or more' (PR) in the form of the following estimated equation:

WE = b0 + b1 * PSE + b2 * PL + b3 * PR.

WE = 2.522 + 0.03764 * PSE + 0.0804 * PL + 1.53E-4 * PR. (substituted b0, b1, b2, b3)

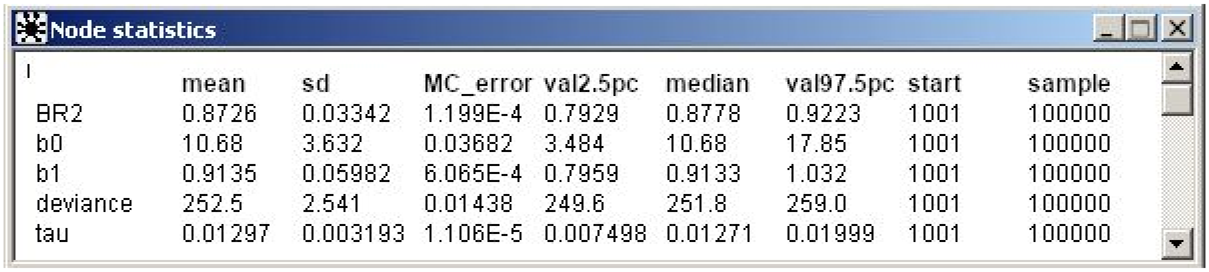
This equation seems to be very similar to the one Higgs proposed in his original research paper using the normal linear regression (without Bayesian) which is:

WE = 2.55 + 0.383 * PSE + 0.0796 * PL

Both of these equations say that 'Average weekly earnings in dollars' WE is not at all dependent on 'Percentage Residing in the US for 5 years or more' PR. In the Bayesian regression approach, I found that the value of b3 which is 1.53E-4 is quite low and can be said that it has a value of almost 0.

I again analyzed the reason behind this by looking at how each variable is correlated with one another by applying the simple linear regression on all the variables taking 2 at a time and found that PR is directly correlated with PSE. (attached the ODC file named ProjectBayesianSLR.odc in the submission). I found the following node statistics using simple linear regression between PSE and PR and obtained the correlation value (by taking R^2):

In the end, I found out the following relationship between PSE and PR:
PSE = b0 + b1 * PR



```
            mean      sd        MC_error  val2.5pc  median    val97.5pc  start  sample
 BR2        0.8726    0.03342   1.199E-4  0.7929    0.8778    0.9223     1001   100000
 b0         10.68     3.632     0.03682   3.484     10.68     17.85      1001   100000
 b1         0.9135    0.05982   6.065E-4  0.7959    0.9133    1.032      1001   100000
 deviance   252.5     2.541     0.01438   249.6     251.8     259.0      1001   100000
 tau        0.01297   0.003193  1.106E-5  0.007498  0.01271   0.01999    1001   100000
```

This means that:
PSE = 10.68 + 0.9135 * PR (substituted b0, b1 from above node statistic)
As seen in the above node statistics, the relationship between PR and PSE has a high Bayesian correlation value (BR2) of 0.8726 thus it means that they are dependent on one another.
Other covariates are not correlated with one another as much as PSE and PR are correlated with one another. The BR2 value for each of such relationships between other variables are not reported in this document because of the 5 pages limit constraint for this report.

CONCLUSION

Similar to the conclusion of the paper, I found that 'Average weekly earnings in dollars' (WE) is only dependent on covariates 'Percentage who speak English language' (PSE) and 'Percentage Literate' (PL). WE is directly dependent on the PSE and PL because literacy and better english speaking ability means that immigrants were more skilled and educated and thus had a higher salary. WE is not dependent much on 'Percentage Residing in the US for 5 years or more' (PR). This is because 'Percentage Residing in the US for 5 years or more' (PR) is directly correlated with 'Percentage who speak English language' (PSE) as seen from the high value of Bayesian R^2. This might be because people from countries who arrived in the US earlier (PR) gained english speaking skills quickly and thus had higher value of (PSE) which in turn made them more skillful. Thus, they had higher weekly earnings. Since the contribution of PSE is already counted in Bayesian multiple linear equations, there was a very minute/almost negligible weightage given to 'Percentage Residing in the US for 5 years or more' (PR).

REFERENCE:

Higgs, R. (1971). Race, Skills, and Earnings: American Immigrants in 1909. The Journal of Economic History, 31(2), 420-428. doi:10.1017/S002205070009094X