Lab 7

matthew.mullins33

March 2018

Note: I did not know that stargazer was not compatible with rjags until after completing all of my work in R and I was forced to paste the images into the document instead of using LATEX. My apologies!

1 Regression Analysis

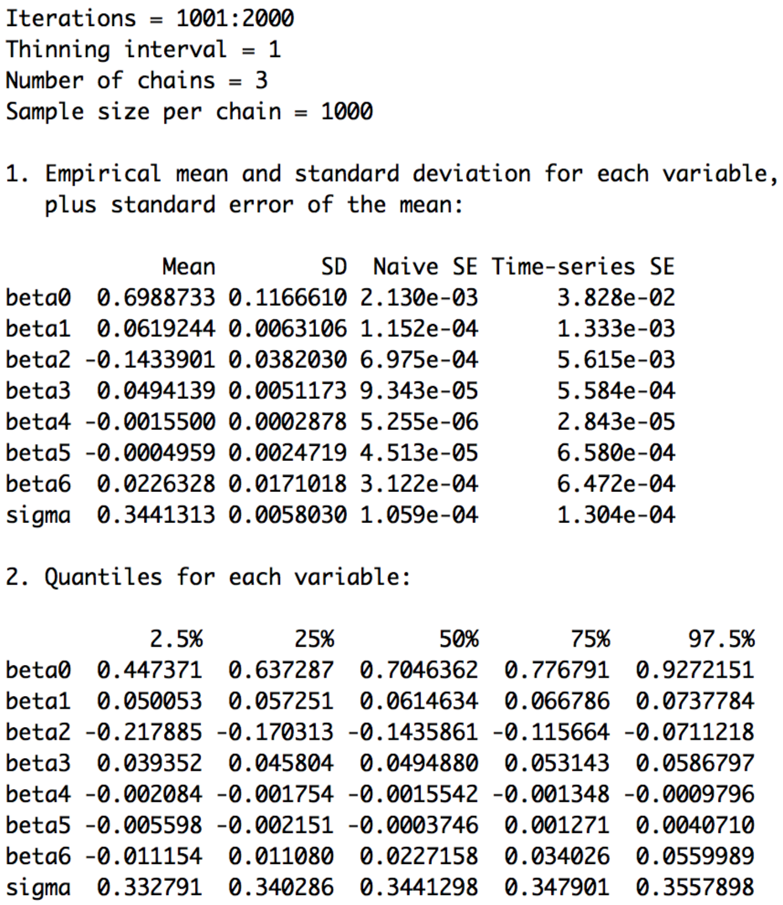
Original Data after factor conversion:

Table 1:

Statistic	N	Mean	St. Dev.	Min	Max
logwage	1,669	1.625	0.386	0.005	2.261
hgc	2,229	13.101	2.524	0	18
college	2,229	0.238	0.426	0	1
tenure	2,229	5.971	5.507	0.000	25.917
age	2,229	39.152	3.062	34	46
married	2,229	0.642	0.480	0	1

Log wages seem to be missing in every 1/5 (560 NAs total). I beleive that the logwage variable is most likely missing not at random. It is most likely the case that the sample includes a good portion of people who are not employed for various reasons and is not contradictory to current US unemployment rates because this sample could include individuals who are not actively seeking work, which could explain the high percentage of those not reporting a salary.

2 Regression 1 Results

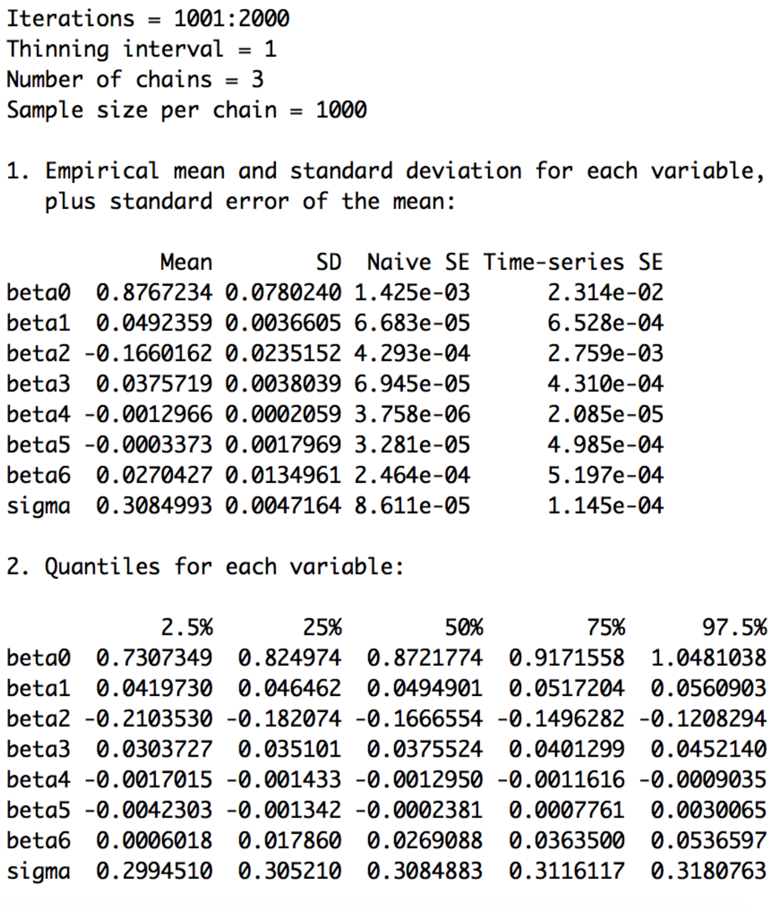


```
Iterations = 1001:2000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

            Mean        SD  Naive SE Time-series SE
beta0  0.6988733 0.1166610 2.130e-03      3.828e-02
beta1  0.0619244 0.0063106 1.152e-04      1.333e-03
beta2 -0.1433901 0.0382030 6.975e-04      5.615e-03
beta3  0.0494139 0.0051173 9.343e-05      5.584e-04
beta4 -0.0015500 0.0002878 5.255e-06      2.843e-05
beta5 -0.0004959 0.0024719 4.513e-05      6.580e-04
beta6  0.0226328 0.0171018 3.122e-04      6.472e-04
sigma  0.3441313 0.0058030 1.059e-04      1.304e-04

2. Quantiles for each variable:

           2.5%       25%        50%       75%      97.5%
beta0  0.447371  0.637287  0.7046362  0.776791  0.9272151
beta1  0.050053  0.057251  0.0614634  0.066786  0.0737784
beta2 -0.217885 -0.170313 -0.1435861 -0.115664 -0.0711218
beta3  0.039352  0.045804  0.0494880  0.053143  0.0586797
beta4 -0.002084 -0.001754 -0.0015542 -0.001348 -0.0009796
beta5 -0.005598 -0.002151 -0.0003746  0.001271  0.0040710
beta6 -0.011154  0.011080  0.0227158  0.034026  0.0559989
sigma  0.332791  0.340286  0.3441298  0.347901  0.3557898
```

[export]adjustbox

Model 1: Results from regression after list-wise deletion

3 Regression 2 Results

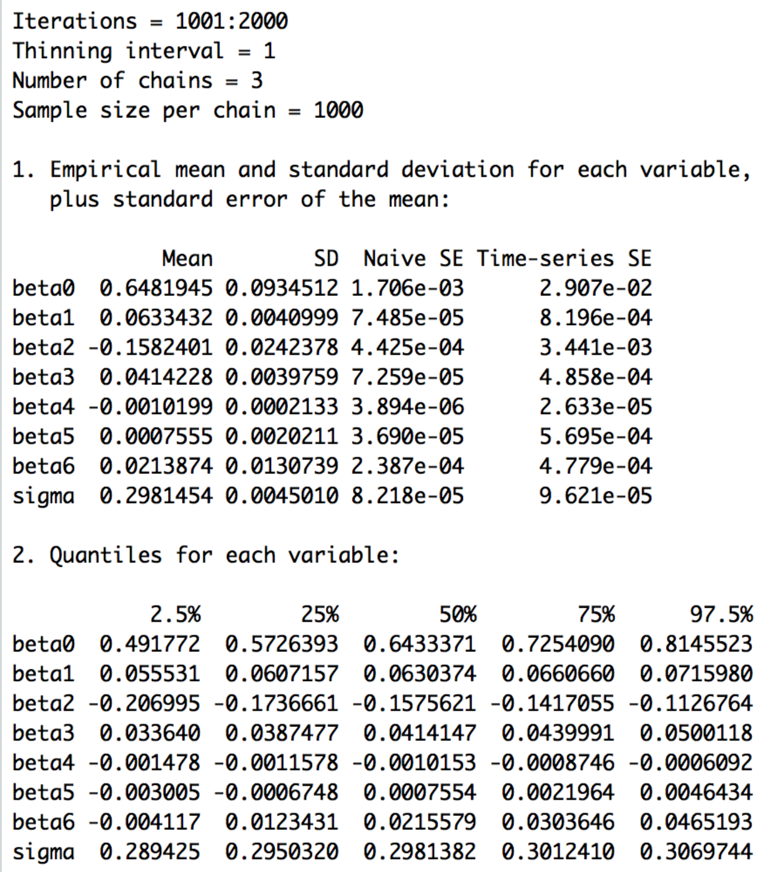


```
Iterations = 1001:2000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

            Mean        SD  Naive SE Time-series SE
beta0  0.8767234 0.0780240 1.425e-03      2.314e-02
beta1  0.0492359 0.0036605 6.683e-05      6.528e-04
beta2 -0.1660162 0.0235152 4.293e-04      2.759e-03
beta3  0.0375719 0.0038039 6.945e-05      4.310e-04
beta4 -0.0012966 0.0002059 3.758e-06      2.085e-05
beta5 -0.0003373 0.0017969 3.281e-05      4.985e-04
beta6  0.0270427 0.0134961 2.464e-04      5.197e-04
sigma  0.3084993 0.0047164 8.611e-05      1.145e-04

2. Quantiles for each variable:

            2.5%       25%        50%        75%      97.5%
beta0  0.7307349  0.824974  0.8721774  0.9171558  1.0481038
beta1  0.0419730  0.046462  0.0494901  0.0517204  0.0560903
beta2 -0.2103530 -0.182074 -0.1666554 -0.1496282 -0.1208294
beta3  0.0303727  0.035101  0.0375524  0.0401299  0.0452140
beta4 -0.0017015 -0.001433 -0.0012950 -0.0011616 -0.0009035
beta5 -0.0042303 -0.001342 -0.0002381  0.0007761  0.0030065
beta6  0.0006018  0.017860  0.0269088  0.0363500  0.0536597
sigma  0.2994510  0.305210  0.3084883  0.3116117  0.3180763
```

[export]adjustbox

Model 2: Results from regression after mean imputation

4 Regression Model 3

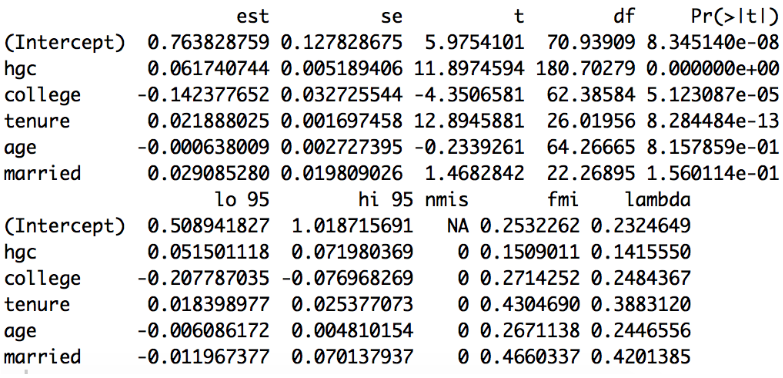


```
Iterations = 1001:2000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

            Mean        SD  Naive SE Time-series SE
beta0  0.6481945 0.0934512 1.706e-03      2.907e-02
beta1  0.0633432 0.0040999 7.485e-05      8.196e-04
beta2 -0.1582401 0.0242378 4.425e-04      3.441e-03
beta3  0.0414228 0.0039759 7.259e-05      4.858e-04
beta4 -0.0010199 0.0002133 3.894e-06      2.633e-05
beta5  0.0007555 0.0020211 3.690e-05      5.695e-04
beta6  0.0213874 0.0130739 2.387e-04      4.779e-04
sigma  0.2981454 0.0045010 8.218e-05      9.621e-05

2. Quantiles for each variable:

           2.5%        25%        50%        75%      97.5%
beta0  0.491772  0.5726393  0.6433371  0.7254090  0.8145523
beta1  0.055531  0.0607157  0.0630374  0.0660660  0.0715980
beta2 -0.206995 -0.1736661 -0.1575621 -0.1417055 -0.1126764
beta3  0.033640  0.0387477  0.0414147  0.0439991  0.0500118
beta4 -0.001478 -0.0011578 -0.0010153 -0.0008746 -0.0006092
beta5 -0.003005 -0.0006748  0.0007554  0.0021964  0.0046434
beta6 -0.004117  0.0123431  0.0215579  0.0303646  0.0465193
sigma  0.289425  0.2950320  0.2981382  0.3012410  0.3069744
```

[export]adjustbox

Model 3: Results from regression after missing at random assumption

5 Regression Model 4



```
                    est          se          t        df     Pr(>|t|)
(Intercept)  0.763828759 0.127828675  5.9754101  70.93909 8.345140e-08
hgc          0.061740744 0.005189406 11.8974594 180.70279 0.000000e+00
college     -0.142377652 0.032725544 -4.3506581  62.38584 5.123087e-05
tenure       0.021888025 0.001697458 12.8945881  26.01956 8.284484e-13
age         -0.000638009 0.002727395 -0.2339261  64.26665 8.157859e-01
married      0.029085280 0.019809026  1.4682842  22.26895 1.560114e-01
                  lo 95        hi 95 nmis       fmi    lambda
(Intercept)  0.508941827  1.018715691   NA 0.2532262 0.2324649
hgc          0.051501118  0.071980369    0 0.1509011 0.1415550
college     -0.207787035 -0.076968269    0 0.2714252 0.2484367
tenure       0.018398977  0.025377073    0 0.4304690 0.3883120
age         -0.006086172  0.004810154    0 0.2671138 0.2446556
married     -0.011967377  0.070137937    0 0.4660337 0.4201385
```

[export]adjustbox

Model 4: Results from regression after mice multiple imputation

6 Results

Since the true value of Beta1 hat is 0.093, it appears that my linear regression models have all significantly underestimated the significance of Beta1 on Salary. This may be the result of using a Gibbs Sampling (JAGS) approach to the regression analysis (see R code), but it could also mean that my data was not as carefully cleaned, or oppositely, was adjusted too much. It is worth noting that the regression using the mice package and not a Gibbs sampler returns a slightly higher Beta1 in my various trials, not all listed, but may further suggest that the Bayesain approach is not ideal. I definitely see a convergence to around .05-0.6 for Beta1, all positive, signifying that there is a positive return on investment for schooling. However, this percentage is not nearly as enticing as the expected .09-.1 ROI.

I was really interested specifically in why Beta2 would be negative, indicating that college had a negative impact on salary, but combing through the data I noticed that a significant portion of the sample that did not report their wages also were the ones who reported being college educated. This may suggest that they prefer to keep higher or lower salaries private and means that the model is not representing the sample accurately.

Using the different imputation methods and seeing the impact each has on the created model has made me much more aware of the importance of first understanding your data before just diving in to clean and analyze it. If you are going to go as far as to remove complete rows of data you should have a very concrete reason for doing so or it will seem as if the data has been skewed to find a specific result. The other two methods can have similiar effects, but I do not think they would be as severe.

7 Project Results

For my project I have been collecting data from the past 7 years on the movements of free agents, team wins, scoring, total championships, best ranked places to live, coach success, etc. Most of this data is relatively easy to collect from pro-sport sites like ESPN.com. Just looking at the data I can already make some hypothesis about the types of results I am going to get.

I am still deciding on my modeling approaches, but would definitely like to incorporate both Bayesian and more traditional regression models into my work. A few models that come to mind are using Gibbs sampling for a quadratic and SLR model. Additionally, I think I will also try to use a GLS model to account for heteroskedasticity.