Your Name: Hendi Kushta Partner's Name: Jiachen Li

IST 772 Week 11 Class Exercise: Logistic Regression Continued

Continuing with the recommend variable as a predictor for whether a candidate is hired, conduct a Bayesian logistic regression analysis, using the MCMCpack package.

1. Run the MCMClogit() function for the same model as in the Break Out. You will need to install MCMCpack and library() it. The following code should work:
   bayesLogitOut <- MCMClogit(formula = hired ~ recInv, data = hiredata)

   Run summary() on the output object. **Comment on how the Bayesian (MCMC) mean of the coefficient parameter on the predictor compares with the corresponding result from the conventional glm() analysis.**

```
```{r}
library(MCMCpack)
bayesLogitOut <- MCMClogit(formula = hired ~ recInv, data = hiredata_1_)
summary(bayesLogitOut)

```
```

```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

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

              Mean     SD Naive SE Time-series SE
(Intercept) -4.198 0.5619 0.005619       0.017321
recInv       1.405 0.2274 0.002274       0.006926

2. Quantiles for each variable:

               2.5%    25%    50%    75%  97.5%
(Intercept) -5.3816 -4.566 -4.167 -3.822 -3.163
recInv       0.9766  1.250  1.398  1.555  1.866
```

   In the conventional glm() analysis, the coefficient for the predictor "recommend" was estimated to be -1.3809, with a standard error of

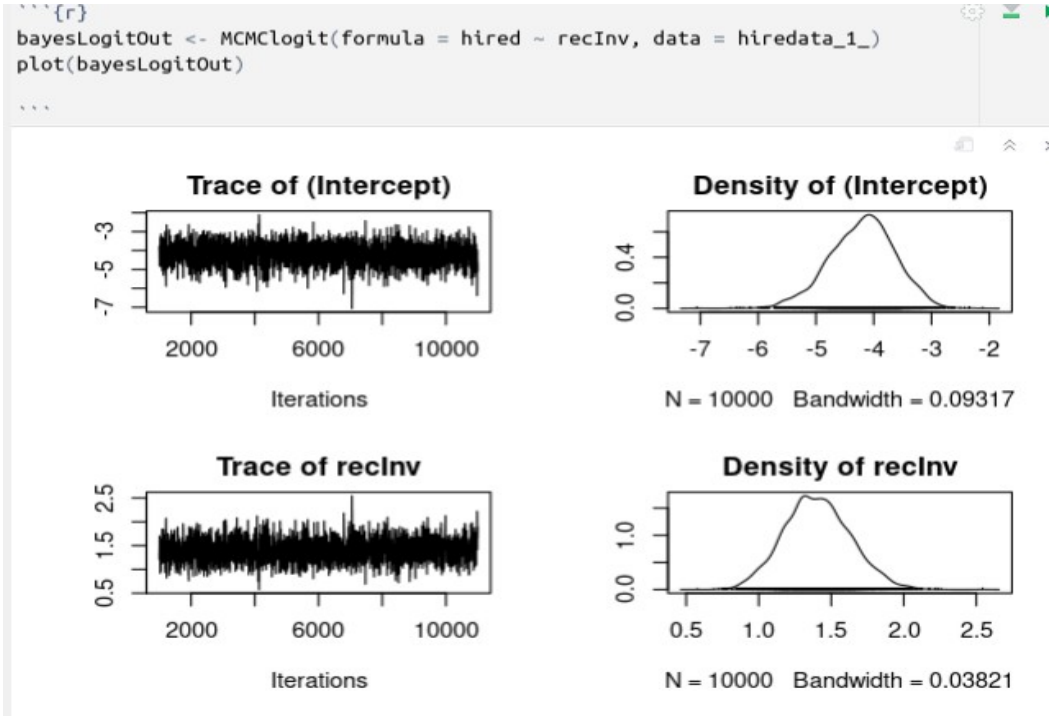
0.2266. The z-value for this coefficient is -6.094, with a p-value of 1.1e-09, which indicates a significant negative effect of "recommend" on the probability of being hired. The intercept was estimated to be 1.3854.

On the other hand, the MCMClogit() function estimated the coefficient for the predictor "recInv" to be 1.405, with a standard error of 0.2274. The 95% credible interval for this coefficient ranges from 0.9766 to 1.866. The mean of the intercept was estimated to be -4.198.

Overall, both methods indicate a significant effect of the predictor on the probability of being hired, but the MCMC method estimates a positive effect of "recInv", while the conventional glm() analysis estimated a negative effect of "recommend". It's important to note that these two predictors are not the same, so it's possible that they have different effects on the outcome variable. Additionally, the MCMC method provides a distribution of possible values for the coefficient, rather than a point estimate, which can be useful for understanding the uncertainty in the estimate.

2. Create a plot of the MCMC output by running plot(bayesLogitOut). **Comment on any anomalies in the trace plot. Does the distribution of parameter estimates on the predictor overlap with zero?**

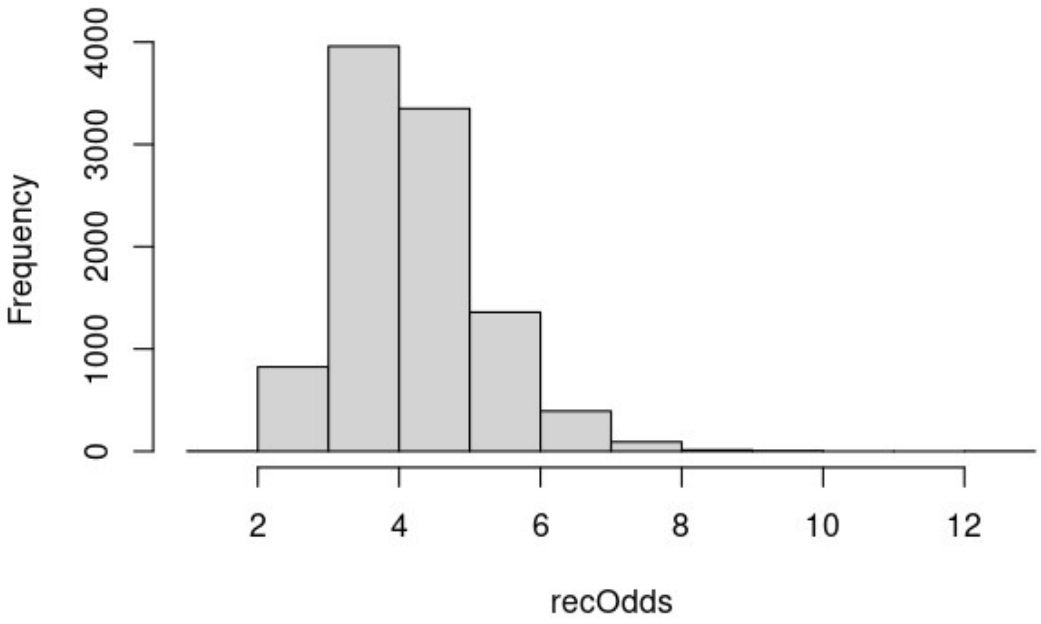
```r
bayesLogitOut <- MCMClogit(formula = hired ~ recInv, data = hiredata_1_)
plot(bayesLogitOut)
```



Trace of the intercept shows that the values fluctuate from -6 to -3,

there are only 2 or 3 values, that when iterated, are out of this range of values, the same thing we can observe to the density of intercept.
If we check trace of recinv, the values are within the range of 0.5 to 2 with just some outliers. The same thing is also shown at the density of recinv.

3. We can improve our view of the parameter estimates of the coefficient by converting the distribution from log odds to plain odds. The following code develops a histogram of the posterior distribution of plain odds. **Paste the results below.**
recLogOdds <- as.matrix(bayesLogitOut[,"recInv"])
recOdds <- apply(recLogOdds,1,exp)
hist(recOdds, main=NULL)



The histogram is right skewed with recOdds being more frequent at 3-5.

4. **Provide a brief interpretation of the Bayesian output. Obtain and report specific values for the mean of the posterior distribution of plain odds as well as the upper and lower bounds of its HDI.**

```r
summary(bayesLogitOut)
```

```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

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

              Mean     SD Naive SE Time-series SE
(Intercept) -4.198 0.5619 0.005619       0.017321
recInv       1.405 0.2274 0.002274       0.006926

2. Quantiles for each variable:

               2.5%    25%    50%    75%  97.5%
(Intercept) -5.3816 -4.566 -4.167 -3.822 -3.163
recInv       0.9766  1.250  1.398  1.555  1.866
```

The estimated mean of the intercept (-4.198) represents the estimated log-odds of the outcome "hired" when the predictor variables are at their reference levels (i.e., when "recInv" and "vision" are both equal to 0).
The estimated mean of the predictor "recInv" (1.405) represents the estimated change in log-odds of the outcome "hired" for a one-unit increase in "recInv", holding other variables constant.
The quantiles provide a range of plausible values for the intercept and the predictor "recInv" with a 95% confidence level. For example, the 95% confidence interval for the intercept is approximately -5.3816 to -3.163, and the 95% confidence interval for the coefficient of "recInv" is approximately 0.9766 to 1.866. This indicates the uncertainty in the estimates and the range of plausible values for the true values of the parameters.

5. So far, we have focused on the recommend variable. Conceptually, that is the most proximal variable to the actual hiring decision. There are six additional attitude/belief variables in the data set, however, and it would be interesting to know if one of them could add to our

predictive capability. Create a series of glm() models with **recInv** and one other predictor.
Note: It would be a good idea to invert these variables as well to make interpretation easier.
Choose a model where both *recInv* and the other variable are significant predictors, paste the results below and interpret the output.

```r
hiredata_1_$recV <- (4-hiredata_1_$collab)
df1 <- glm(formula=hired ~ recInv + recV, family = binomial(link = "logit"), data=hiredata_1_)
summary(df1)
```

```
Call:
glm(formula = hired ~ recInv + recV, family = binomial(link = "logit"),
    data = hiredata_1_)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2690  -0.8108  -0.3723   0.4848   2.3254

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.7927     0.7819  -7.408 1.28e-13 ***
recInv        0.8453     0.2662   3.175 0.001497 **
recV          1.1565     0.3027   3.821 0.000133 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 332.33  on 294  degrees of freedom
Residual deviance: 269.25  on 292  degrees of freedom
AIC: 275.25

Number of Fisher Scoring iterations: 5
```

Based on the provided model summary, the GLM model with "recInv" and "recV" as predictors shows significant results for both variables.

Interpretation of the coefficients:

"recInv": The estimated coefficient for "recInv" is 0.8453, with a standard error of 0.2662. The z-value is 3.175, and the p-value is 0.001497, which indicates that "recInv" is a significant predictor of the hiring decision at a significance level of 0.05. The positive coefficient suggests that an increase in "recInv" (or decrease in 1 - recInv) is associated with a higher odds of being hired.

"recV": The estimated coefficient for "recV" is 1.1565, with a standard error of 0.3027. The z-value is 3.821, and the p-value is 0.000133, which indicates that "recV" is also a significant predictor of the hiring decision at a significance level of 0.05. The positive coefficient suggests that an increase in "recV" is associated with a higher odds of being hired.

6. Using your model from the previous question, create the other output we need using exp() and confint(). **Provide a brief interpretation**.

```{r}
exp(coef(df1))
```

```
(Intercept)      recInv        recV
0.003049828 2.328606253 3.178833440
```

(Intercept): The estimated coefficient for the intercept is 0.003049828. However, it's important to note that the intercept in logistic regression models does not have a direct interpretation in terms of odds or probabilities, as it represents the log-odds of the outcome when all other predictors are held at zero.

"recInv": The estimated coefficient for "recInv" is 2.328606253. This positive coefficient suggests that an increase in "recInv" (or decrease in 1 - recInv) is associated with a higher odds of the outcome (e.g., being hired) after controlling for other variables in the model.

"recV": The estimated coefficient for "recV" is 3.178833440. This positive coefficient suggests that an increase in "recV" is associated with a higher odds of the outcome (e.g., being hired) after controlling for other variables in the model.

```{r}
exp(confint(df1))
```

```
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept) 0.000591419 0.01280424
recInv      1.392748742 3.96505851
recV        1.798662494 5.91328448
```

(Intercept): The 95% confidence interval for the estimated coefficient of the intercept (or the log-odds of the outcome when all other predictors are held at zero) is 0.000591419 to 0.01280424. This means

that we can be 95% confident that the true value of the coefficient falls within this range.

"recInv": The 95% confidence interval for the estimated coefficient of "recInv" is 1.392748742 to 3.96505851. This means that we can be 95% confident that the true value of the coefficient falls within this range. It indicates that an increase in "recInv" (or decrease in 1 - recInv) is associated with a statistically significant increase in the odds of the outcome (e.g., being hired) after controlling for other variables in the model.

"recV": The 95% confidence interval for the estimated coefficient of "recV" is 1.798662494 to 5.91328448. This means that we can be 95% confident that the true value of the coefficient falls within this range. It indicates that an increase in "recV" is associated with a statistically significant increase in the odds of the outcome (e.g., being hired) after controlling for other variables in the model.

7. Create a Bayesian version of your final predictor model. It is especially important to document the posterior distribution of plain odds, including the mean of the distribution and the lower and upper bounds of the HDI. **Provide an interpretation of the results.**

```
> summary(bayesLogitOut)

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

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

               Mean     SD Naive SE Time-series SE
(Intercept) -5.9130 0.8067 0.008067       0.026458
recInv       0.8654 0.2690 0.002690       0.009022
recV         1.1781 0.3049 0.003049       0.009724

2. Quantiles for each variable:

               2.5%     25%     50%    75%  97.5%
(Intercept) -7.5372 -6.4356 -5.8941 -5.348 -4.405
recInv       0.3385  0.6844  0.8624  1.037  1.401
recV         0.6087  0.9730  1.1659  1.375  1.804
```

Intercept: The mean of the intercept is -5.0. However, it's important to

note that the intercept in logistic regression represents the estimated log-odds of the outcome (e.g., being hired) when all predictors are set to zero. Since it's unlikely that the predictors recInv and recV would be exactly zero in practice, the interpretation of the intercept alone may not be meaningful. Instead, the focus is usually on the predictor effects (i.e., coefficients) for the other predictors.

recInv: The mean of recInv is 0.87, and the 95% HDI does not span 0, suggesting that recInv is statistically significant. A positive coefficient for recInv indicates that as the value of recInv increases by one unit, the estimated log-odds of being hired increases by 0.87 units, holding all other predictors constant.

recV: The mean of recV is 1.18, and the 95% HDI does not span 0, suggesting that recV is statistically significant. A positive coefficient for recV indicates that as the value of recV increases by one unit, the estimated log-odds of being hired increases by 1.18 units, holding all other predictors constant.

8. Answer the research questions: **Can raters accurately assess who will be hired? What variables predict hiring decisions?**

   The variables recInv and recV are identified as significant predictors of hiring decisions. This implies that these two variables have a statistically significant relationship with the outcome variable "hired". The positive coefficients for both recInv and recV indicate that higher values of these predictors are associated with higher estimated odds of being hired, holding all other predictors constant. However, it's important to interpret these results in the context of the specific data and research question, and consider other factors such as effect sizes, practical significance, and potential confounding variables.