Bayesian Logistic Regression Analysis of Obesity Among Asian Americans in California

Jonathan King

UCLA Fielding School of Public Health

BIOSTAT M234: Applied Bayesian Inference

Dr. Robert Weiss

March 21, 2023

Analysis Goals and Dataset Description

Obesity continues to be a major public health issue in the United States. While there have been many research studies on this topic, few studies have explored obesity in Asian Americans. To address this research gap, a Bayesian analysis is conducted using a logistic regression to determine characteristics that affect risk of obesity among Asian Americans adults in California.

The data set used for this study is pooled data from the 2020 and 2021 California Health Information Survey (CHIS), the largest annual state health survey in the United States (UCLA Center for Health Policy Research, n.d.). This survey is jointly conducted by the UCLA Center for Health Policy Research, the California Department of Public Health, and the California Department of Health Care services. While women, children, and adults all have separate surveys, this analysis focuses on the adult survey. The sample is defined as the 6,818 Asian participants. The outcome variable, obesity, is dichotomized as 'obese' or 'not obese', with 'obese' being defined as having a body mass index (BMI) of 27.5 and higher, which is the cut point for Asians according to the World Health Organization (2004). Predictor variables of interest are categorical age (18–39 years, 40–59 years, and 60 years or older), sex, family income (0–99% of federal poverty line (FPL), 100–299% of FPL, or 300% of FPL and above), marital status (married, never married, or other), smoking status (current smoker or not), and education (high school or less, some college, or college graduate).

Model Specification

Since the outcome is a dichotomous variable, a Bayesian logistic regression model was fit. The reference group is defined as never-married male college graduates aged 18–39 years who

do not currently smoke and have an income 100–299% of the federal poverty line (FPL). With *i* indexing individuals from 1 to 6,818, y_i indicating if individual *i* is obese, p_i indicating the probability of individual *i* of being obese, '*female*' indicating female sex, '*middleage*' indicating that age is 40–59 years old, '*senior*' indicating that age is 60 years old or older, '*poor*' indicating that income is below 100% of the FPL, '*rich*' indicating that income is above 299% of the FPL, '*highorless*' indicating high school graduate or less education level, '*somecollege*' indicating past or current enrollment in college without graduating, '*married*' indicating marriage, '*othermaritial*' indicating marital statuses aside from either marriage or never-married, and '*smoking*' indicating current smoking, the model is specified as follows:

$y_i|p_i \sim Bernoulli(p_i)$, where $\mathrm{logit}(\mathrm{p_i}) = \beta_0 + \beta_1 * female_i + \beta_2 * middleage_i + \beta_3 * senior_i + \beta_4 * poor_i + \beta_5 * rich_i + \beta_6 * highorless_i + \beta_7 * somecollege_i + \beta_8 * married_i + \beta_9 * othermarital_i + \beta_{10} * smoking_i$.

Prior Specification

Normal priors were set for all regression coefficients in the model on the log odds scale. Priors were derived from various research studies, so means were set based on estimates from these studies, but variances were inflated by a factor that depended on how different prior study populations were from the target population of this analysis. To determine the intercept prior, reported proportions from an analysis of obesity among Chinese adults were converted to log odds. Since 15.66% of males aged 18-39 years were reported to obese, the mean of the corresponding coefficient was set to be -1.68, but the variance of approximately .0225 on the log odds scale was inflated by a factor of 50 to account for the major difference in population, so our prior was set to be $\beta_0 \sim N_o(-1.68, 0.318)$ (Chen et al., 2019). Priors for age, smoking, and

marital status were also set using a similar procedure on calculated log odds from the same study given counts of participants with obesity and corresponding percentages. These priors were set as $\beta_2 \sim N_0(-.015, 2.40)$, $\beta_3 \sim N_0(-0.31, 1.82)$, $\beta_8 \sim N_0(0.50, 1.17)$, $\beta_9 \sim N_0(0.50, 0.89)$, and $\beta_{10} \sim N_0(-.089,\ 6.56)$. The prior for female sex was set as $\beta_1 \sim N_o(-0.253, 4.13)$ by calculating log odds from results for non-Hispanic Asians from an analysis of 2017–18 National Health and Nutrition Examination Study (NHANES) data and inflating approximated variances by a factor of 10 (Liu et al., 2021). The Range Method was applied to results for non-Hispanic Asians from an analysis conducted on 2011-14 NHANES data to determine prior variances for income coefficients and education coefficients because specific counts were not reported, while prior means were calculated by calculating log odds from reported results (Ogden et al., 2017). These priors were set as $\beta_4 \sim N_o(0.34,\ 0.45)$, $\beta_5 \sim N_0(-0.051, 0.70)$, $\beta_6 \sim N_o(0.051, 0.94)$, and $\beta_7 \sim N_o(0.13,\ 1.20)$.

Results

Posterior estimates are reported in Table 1 for the base odds and odds ratios. On average, base odds of obesity, defined as the odds of obesity for the aforementioned reference group, are 0.32 (SD = 0.03). Asian Americans in California who are female (95% CI = [0.61, 0.79]) or are 60 years or older (95% CI = [0.54, 0.77]) have lower odds of obesity, controlling for all other variables. Conversely, Asian Americans in California who have a high school education or less (95% CI = [1.01, 1.53]) or some college education (95% CI = [1.07, 1.49]) have higher odds of obesity, all else equal. Odds of obesity does not significantly depend on income, marital status, or smoking status, adjusting for all other variables.

Prior and posterior densities are plotted for all regression coefficients and the intercept on the log odds scale (Figure 1). Overall, all posteriors are much more peaked and are far narrower than the corresponding priors, meaning that posterior mean estimates are much more precisely estimated. This indicates that posterior standard deviation is closer to that of the likelihood, so priors weakly influence posterior shape, if at all. Posterior means are lower than the matching prior means for the female sex, senior age, high income, and married effects but higher for the some college education and middle age effects as well as the intercept. The other effects have approximately equal posterior and prior means.

Sensitivity Analyses

Sensitivity analyses were performed to determine if different priors for regression coefficients have more influence on posterior distributions. For the first sensitivity analysis, all prior variances were first decreased by a factor of three then by a factor of nine to determine if narrower priors are more influential. Table 2 shows the results of this analysis for the odds ratios. Overall, while other marital status becomes a positive predictor of obesity when prior variances are decreased by a factor of 9 (95% CI = [1.03, 1.53]), the other posterior estimates do not significantly change with prior variance, so it can be concluded that narrower priors are generally not more influential. Since normal priors have minimal influence on posterior distributions, the second sensitivity analysis explored specifying t-distributed priors for all regression coefficients and the intercept to see if changing prior distribution type influences results (Table 3). Ultimately, this analysis also concludes that posterior estimates did not significantly change, so the final conclusion is that posteriors for regression coefficients are almost completely driven by the data and not by priors.

References

About CHIS. UCLA Center for Health Policy Research. (n.d.). Retrieved March 15, 2023, from https://healthpolicy.ucla.edu/chis/about/Pages/about.aspx

Chen, Y., Peng, Q., Yang, Y., Zheng, S., Wang, Y., & Lu, W. (2019). The prevalence and increasing trends of overweight, general obesity, and abdominal obesity among Chinese adults: a repeated cross-sectional study. *BMC public health*, *19*(1), 1293. https://doi.org/10.1186/s12889-019-7633-0

Liu, B., Du, Y., Wu, Y., Snetselaar, L. G., Wallace, R. B., & Bao, W. (2021). Trends in obesity and adiposity measures by race or ethnicity among adults in the United States 2011-18: population based study. *BMJ (Clinical research ed.)*, *372*, n365. https://doi.org/10.1136/bmj.n365

Ogden, C. L., Fakhouri, T. H., Carroll, M. D., Hales, C. M., Fryar, C. D., Li, X., & Freedman, D. S. (2017). Prevalence of Obesity Among Adults, by Household Income and Education - United States, 2011-2014. *MMWR. Morbidity and mortality weekly report*, *66*(50), 1369–1373. https://doi.org/10.15585/mmwr.mm6650a1

WHO Expert Consultation (2004). Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet (London, England)*, *363*(9403), 157–163. https://doi.org/10.1016/S0140-6736(03)15268-3

Appendix

Table 1. Posterior distribution estimates of base odds (Base) and odds ratios (OR[*characteristic]*) of a logistic regression model modeling obesity among Asian Americans in California. Posterior means, standard deviations, and 95% credible intervals are presented for each parameter. Posterior probabilities that each parameter is greater than and less than 1 are also shown. Abbreviations: CI = credible interval; OR = odds ratio; SD = standard deviation.

Parameter	Mean	SD	[95% CI]	P < 1	P > 1
Base	0.32	0.03	[0.26, 0.38]	1	0
OR[female]	0.69	0.05	[0.61, 0.79]	1	0
OR[middleage]	1.01	0.08	[0.85, 1.18]	0.49	0.51
OR[senior]	0.65	0.06	[0.54, 0.77]	1	0
OR[poor]	0.98	0.12	[0.77, 1.23]	0.58	0.42
OR[rich]	0.88	0.07	[0.75, 1.02]	0.95	0.05
OR[highorless]	1.25	0.13	[1.01, 1.53]	0.02	0.98
OR[somecollege]	1.27	0.11	[1.07, 1.49]	0	1
OR[married]	0.94	0.08	[0.79, 1.11]	0.78	0.22
OR[othemarital]	1.21	0.13	[0.98, 1.48]	0.04	0.96
OR[smoking]	0.92	0.13	[0.68, 1.20]	0.74	0.26

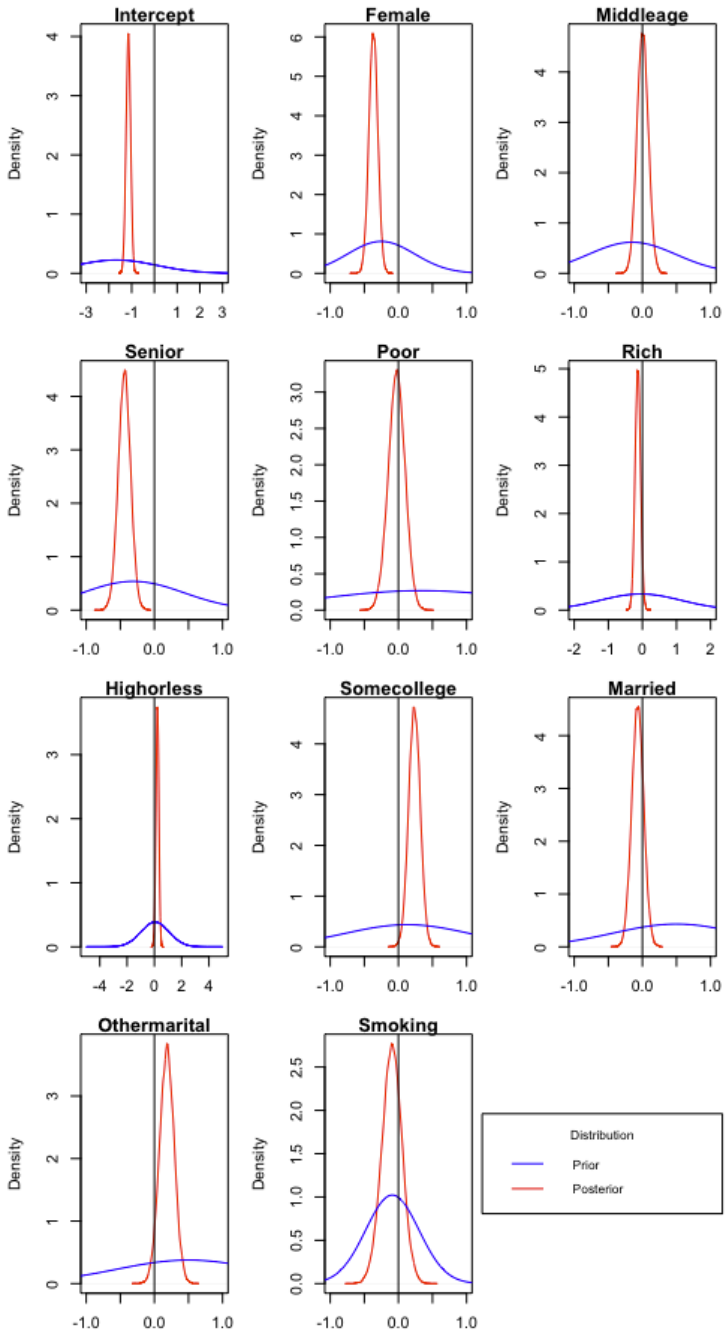

Figure 1. Prior and posterior densities of intercept and regression coefficients on the log odds scale. Prior densities are blue; posterior densities are red. Vertical black line in all plots is at the zero point.

Table 2. Sensitivity analysis for regression coefficient prior variances. Posterior means, standard deviations, 95% credible intervals and probabilities for base odds (Base) and odd ratios (OR[*characteristic*]) are recorded for low, medium, and high values of prior variances. Low variances are defined as those of the primary analysis divided by 3; very low variances are defined as those of the primary analysis divided by 9. Normal variances matches that of the primary analysis. Abbreviation: CI = credible interval; OR = odds ratio; SD = standard deviation.

Parameter	**Variance**	**Mean**	**SD**	**[95% CI]**	**P < 1**	**P > 1**
Base	Normal	0.32	0.03	[0.26, 0.38]	1	0
	Low	0.32	0.03	[0.26, 1.38]	1	0
	Very Low	0.31	0.03	[0.25, 0.37]	1	0
OR[female]	Normal	0.69	0.05	[0.61, 0.79]	1	0
	Low	0.7	0.04	[0.61, 0.79]	1	0
	Very Low	0.7	0.04	[0.62, 0.79]	1	0
OR[middleage]	Normal	1.01	0.08	[0.85, 1.18]	0.49	0.51
	Low	1	0.08	[0.85, 1.17]	0.52	0.48
	Very Low	0.98	0.07	[0.84, 1.13]	0.62	0.38
OR[senior]	Normal	0.65	0.06	[0.54, 0.77]	1	0
	Low	0.64	0.06	[0.54, 0.76]	1	0
	Very Low	0.64	0.05	[0.54, 0.75]	1	0
OR[poor]	Normal	0.98	0.12	[0.77, 1.23]	0.58	0.42
	Low	0.99	0.12	[0.77, 1.24]	0.56	0.44
	Very Low	1	0.12	[0.79, 1.26]	0.51	0.49
OR[rich]	Normal	0.88	0.07	[0.75, 1.02]	0.95	0.05
	Low	0.88	0.07	[0.75, 1.02]	0.96	0.05
	Very Low	0.88	0.07	[0.75, 1.02]	0.96	0.04
OR[highorless]	Normal	1.25	0.13	[1.01, 1.53]	0.02	0.98
	Low	1.25	0.13	[1.01, 1.52]	0.02	0.98
	Very Low	1.23	0.13	[1.00, 1.49]	0.02	0.98
OR[somecollege]	Normal	1.27	0.11	[1.07, 1.49]	0	1
	Low	1.27	0.11	[1.07, 1.49]	0	1
	Very Low	1.26	0.1	[1.07, 1.47]	0	1
OR[married]	Normal	0.94	0.08	[0.79, 1.11]	0.78	0.22
	Low	0.95	0.08	[0.80, 1.12]	0.73	0.27
	Very Low	0.98	0.08	[0.84, 1.15]	0.59	0.41
OR[othermarital]	Normal	1.21	0.13	[0.98, 1.48]	0.04	0.96
	Low	1.22	0.13	[0.99, 1.49]	0.03	0.97
	Very Low	1.26	0.13	[1.03, 1.53]	0.01	0.99
OR[smoking]	Normal	0.92	0.13	[0.68, 1.20]	0.74	0.26
	Low	0.92	0.12	[0.71, 1.17]	0.76	0.24
	Very Low	0.92	0.09	[0.75, 1.17]	0.81	0.19

Table 3. Posterior distributions estimates of base odds (Base) and odds ratios (OR[*characteristic*]) of the logistic model using t-distributed priors with 5 degrees of freedom. Means and variances of these priors match those of the primary analysis. Posterior means, standard deviations, and 95% credible intervals are presented for each parameter. Posterior probabilities that each parameter is greater than and less than 1 are also shown. Abbreviations: CI = credible interval; OR = odds ratio; SD = standard deviation.

Parameter	Mean	SD	[95% CI]	P < 1	P > 1
Base	0.32	0.03	[0.26, 0.38]	1	0
OR[female]	0.70	0.05	[0.61, 0.79]	1	0
OR[middleage]	1.01	0.08	[0.86, 1.17]	0.48	0.52
OR[senior]	0.65	0.06	[0.54, 0.77]	1	0
OR[poor]	0.99	0.12	[0.78, 1.24]	0.58	0.42
OR[rich]	0.88	0.07	[0.75, 1.02]	0.96	0.04
OR[highorless]	1.25	0.13	[1.02, 1.52]	0.02	0.98
OR[somecollege]	1.27	0.11	[1.07, 1.49]	0	1
OR[married]	0.94	0.08	[0.79, 1.12]	0.77	0.23
OR[othermarital]	1.21	0.12	[0.98, 1.47]	0.04	0.96
OR[smoking]	0.92	0.13	[0.68, 1.21]	0.73	0.27

a)

```
model {
  for(i in 1:N)
  {
    y[i] ~ dbern(p[i])
    logit(p[i]) <- beta0 + inprod(x[i,],beta[])
  }
  beta0 ~ dnorm(mbeta0, precbeta0)
  baseodds <- exp(beta0)
  for(j in 1:10)
  {
    beta[j] ~ dnorm(m[j], prec[j])
    OR[j] <- exp(beta[j])
  }
}
```

b)

```
# Declare priors and data
fdapdata <- list(y = y,
  x = x,
  m=c(-0.253,-0.15, -0.31, 0.34, -0.051, 0.051, 0.13, 0.50, 0.50, -0.089),
  prec = c( 4.13, 2.40, 1.82, 0.45, 0.70, 0.94, 1.20, 1.17, 0.89, 6.56),
  mbeta0 = -1.68, precbeta0 = 0.318, N = 6818)
# Initial values and parameters to track
fdapinits <- rep(list(list(beta0 = 0, beta = c(0,0,0,0,0,0,0,0,0,0))), 5)
fdapparameters <- c("beta0", "beta","OR", "baseodds")
# Run the model
run1 <- jags(fdapdata, fdapinits, fdapparameters,
             "FDAP_model.txt", n.chains=5, n.iter=10000,
             n.burnin=1000, n.thin=1)
```

Figure 2. Primary analysis **a)** BUGS model and **b)** JAGS code.

```
## Very Low Variance
fdapdata2 <- list(y = y,
  x = x,
  m=c(-0.253,-0.15, -0.31, 0.34, -0.051, 0.051, 0.13, 0.50, 0.50, -0.089),
  prec = c( 4.13*9, 2.40*9, 1.82*9, 0.45*9, 0.70*9,
    0.94*9, 1.20*9, 1.17*9, 0.89*9, 6.56*9),
  mbeta0 = -1.68, precbeta0 = 0.318, N = 6818)
run2 <- jags(fdapdata2, fdapinits, fdapparameters,
             "FDAP_model.txt", n.chains=5, n.iter=10000,
             n.burnin=1000, n.thin=1)
## Low Variance
fdapdata3 <- list(y = y,
  x = x,
  m=c(-0.253,-0.15, -0.31, 0.34, -0.051, 0.051, 0.13, 0.50, 0.50, -0.089),
  prec = c( 4.13*3, 2.40*3, 1.82*3, 0.45*3, 0.70*3, 0.94*3,
    1.20*3, 1.17*3, 0.89*3, 6.56*3),
mbeta0 = -1.68, precbeta0 = 0.318, N = 6818)
run3 <- jags(fdapdata3, fdapinits, fdapparameters,
             "FDAP_model.txt", n.chains=5, n.iter=10000,
             n.burnin=1000, n.thin=1)
```

Figure 3. JAGS code to run sensitivity analysis for regression coefficient prior variances.

```
model {
  for(i in 1:N)
  {
    y[i] ~ dbern(p[i])
    logit(p[i]) <- beta0 + inprod(x[i,],beta[])
  }
  beta0 ~ dt(mbeta0, precbeta0, 5)
  baseodds <- exp(beta0)
  for(j in 1:10)
  {
    beta[j] ~ dt(m[j], prec[j], 5)
    OR[j] <- exp(beta[j])
  }
}
```

Figure 4. BUGS model with t-distributed intercept and regression coefficients.