Modelling obesity rates in Europe based on level of education attained and gender

Ali Shana'a - r0857263

August 2021

1 Introduction

The purpose of this project is to create a Bayesian model that predicts the probability that a citizen of a European country is obese based on the level of education attained, as well as the individual's gender. From here on out, it wil lbe assumed that the population of humans being discussed are from European countries.

The creation of the final model was carried out gradually where several more basic models being created, slowly building up to the eventual predictive model that incorporates both variables. More specifically the steps taken were:

1. Generic prediction of P(obesity)
2. Prediction of P(obesity | education level)
3. Prediction of P(obesity | education level, gender)
4. Prediction of P(obesity | education level, gender) through a linear regression

The datasets utilized are from Eurostat[1], the same dataset that worldobesity.org utilizes for all its data regarding Europe. They consist of information regarding 32 European countries and their respective %'s of people that are underweight, normal weight, and obese, with more specific breakdowns available according to:

- Gender
- Level of education
- Age class

Before handling the data, some pre-processing is carried out where any country with NA values are removed from the datasets, as well as the separation of obese, normal weight, and underweight into their own distinct datasets (since we only want to deal with obesity data). Additionally, the % values were converted to the number of obese/underweight/normal people per 10,000 people.

Finally, it is worth noting, that while the age class variable was not explored, it would be interesting to explore it as a follow up to the work done here.

2 Generic obesity prediction

The model created to carry out this prediction is the following:

```
model {
    for (i in 1:n){
        obese[i] ~ dbin(theta,num_people[i]) # likelihood
    }
    theta ~ dbeta( 1 , 1 ) # prior
}
```

[1] https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_ehis_bm1e&lang=en

where we have a binomial likelihood function for the number of obese people per 10,000 and a non-informative uniformly distributed prior for theta.

The inputs that this model takes are:

- n: number of observations, i.e. number of countries that we have data on
- obese: number of obese people per 10,000
- num_people: number of people, which is fixed to be 10,000

Afterwards, the model was run through JAGS, providing the following posterior distribution for theta.

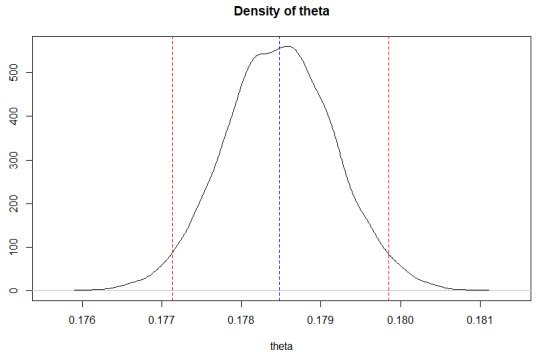

Figure 1: Density of theta

As can be seen above, the distribution is quite a small one as is summarized below:

	mean	sd	CI - LL	CI - UL
theta	17.85%	0.0692%	17.71%	17.99%

Table 1: Model 1 summary

3 Predicting obesity based on level of education

In this model we have added the dimension of level of education which are split into 3 groups:

1. Levels 0-2: Less than primary, primary and lower secondary education
2. Levels 3-4: Upper secondary and post-secondary non-tertiary education
3. Levels 5-8: Tertiary education

The model created to carry out these predictions is the following:

```
model {
    for (i in 1:n){
      obese[i] ~ dbin(theta[eid[i]],num_people[i]) # likelihood
    }
    for (g in 1:eg){
      theta[g] ~ dbeta( 1 , 1 ) # prior
    }
}
```

where we have a binomial likelihood function for the number of obese people per 10,000 and eg non-informative uniformly distributed prior for theta, where eg is the number of levels of education.

The inputs that the model takes are the same as model 1 (generic obesity) as well as:

- eid: a vector identifying which level of education theta[] belongs to for each obese[i]

- eg: the number of levels of education

Afterwards the model is run through JAGS, providing the following posterior distributions:

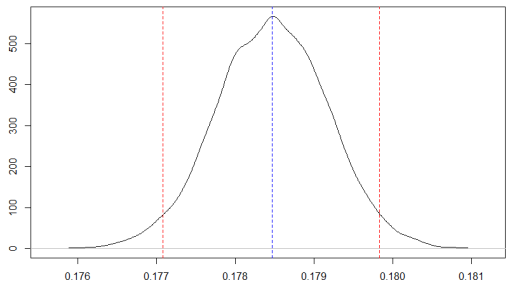

Figure 2: Theta[1] - Average person

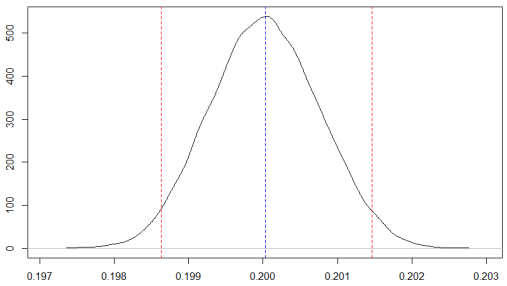

Figure 3: Theta[2] - levels 0-2 of education

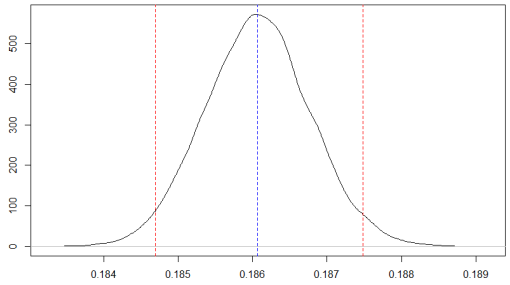

Figure 4: Theta[3] - levels 3-4 of education

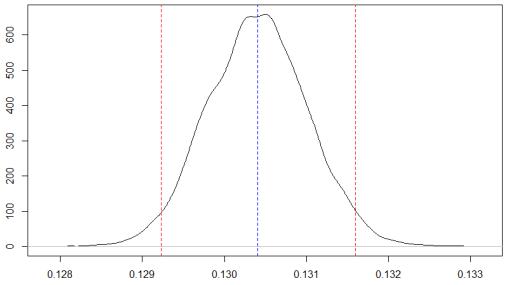

Figure 5: Theta[4] - levels 5-8 of education

The above distributions can be summarized in the following table:

	mean	sd	CI - LL	CI - UL
theta[1]	17.85%	0.0699%	17.71%	17.98%
theta[2]	20.00%	0.0731%	19.86%	20.15%
theta[3]	18.61%	0.0706%	18.47%	18.75%
theta[4]	13.01%	0.0605%	12.92%	13.16%

Table 2: Model 2 summary

Here, we can see that theta[1] is approximately identical to what we saw in model 1. However, the new information comes in from theta[2] - theta[4], all of which have quite similar standard deviations, but show a clear pattern that the more educated a person is, the lower the probability that a person is obese. This will be explored even further in the following models.

4 Predicting obesity based on level of education and gender

In this model we have added the dimension of gender, split into male and female. The model created to carry out these predictions is the following:

```
model {
    for (i in 1:n){
        obese[i] ~ dbin(theta[ed_id[i],gen_id[i]],num_people[i]) # likelihood
    }
    for (x in 1:ed){
```

```
            for(y in 1:gd){
                theta[x,y] ~ dbeta( 1 , 1 ) # priors
            }
        }
    }
```

where we have a binomial likelihood function for the number of obese people per 10,000 and ed*gd non-informative uniformly distributed priors, where eg is the number of levels of education, and gd is the number of levels of gender.

The inputs that the model takes are the same as model 2 (education level predictions) as well as:

- gen_id: a vector identifying which level of gender theta[,] belongs to for each obese[i]
- gd: number of levels of gender

Afterwards the model is run through JAGS, providing us with posterior distributions for each theta, as can be seen in the table below[2]:

	mean	sd	CI - LL	CI - UL
theta[1,1]	17.85%	0.0696%	17.71%	17.98%
theta[2,1]	20.00%	0.0732%	19.86%	20.15%
theta[3,1]	18.61%	0.0705%	18.47%	18.75%
theta[4,1]	13.04%	0.0617%	12.92%	13.16%
theta[1,2]	17.61%	0.0698%	17.47%	17.74%
theta[2,2]	21.34%	0.0739%	21.20%	21.49%
theta[3,2]	17.87%	0.0695%	17.74%	18.01%
theta[4,2]	11.71%	0.0582%	11.60%	11.82%
theta[1,3]	18.08%	0.0712%	17.95%	18.22%
theta[2,3]	18.12%	0.0697%	17.99%	18.26%
theta[3,3]	19.22%	0.0720%	19.08%	19.37%
theta[4,3]	14.47%	0.0638%	14.34%	14.59%

Table 3: Model 3 summary

Before looking at the results, we provide a key for what each theta represents. We have that for every theta[x,y] x represents the level of education and y represents the gender. Moreover, x is such that:

- 1 represents the average person
- 2 represents levels 0-2 of education
- 3 represents levels 3-4 of education
- 4 represents levels 5-8 of education

and y is such that:

- 1 represents the average person
- 2 represents females
- 3 represents males

With that out of the way, we can see that the first rows show virtually identical estimates as we saw from model 2. However, a quick look at the table shows us a difference between the relationship that exists with education between genders. We can see this difference more clearly by comparing the thetas for each level of education for each gender with the average of the gender as can be seen in tables 3 and 4:

[2]Graphs are quite similar to the ones seen before and so were not provided for the sake of space

	Mean	CI - LL	CI - UL
lvls 0-2 - Average	0.04%	-0.16%	0.23%
lvls 3-4 - Average	1.14%	0.94%	1.34%
lvls 5-8 - Average	-3.62%	-3.80%	-3.43%

Table 4: Male

	Mean	CI - LL	CI - UL
lvls 0-2 - Average	3.73%	3.53%	3.92%
lvls 3-4 - Average	0.26%	0.07%	0.45%
lvls 5-8 - Average	-5.90%	-6.08%	-5.72%

Table 5: Female

Looking at table 3 it is quickly evident that the linear relationship of more education implying lower probability of obesity does not hold so strongly for males. However, it does still show a significant difference between the college+ educated males and the other two levels. Furthermore, looking at table 4 we see that the probability of females being obese is significantly more affected by the level of education they have attained.

5 Predicting obesity rates based on level of education and gender through linear regression

The model that will be created here will utilize a linear regression, allowing us to gain a better understanding of the effect of each variable of the two variables, but not of the relationship between the two variables. The model that is utilized is the following:

```
    model {
        for(i in 1:n){
          obese[i] ~ dbin(theta[ed_id[i],gen_id[i]],num_people[i]) # likelihood
        }
    # priors
    for(x in 1:4){
        for(y in 1:3){
            theta[x,y] ~ dnorm(mu[x,y],tau)
            mu[x,y] <- B0 + B1 * edu1[x] + B2 * edu2[x] +
                       B3 * edu3[x] + B4 * gender1[y] + B5 * gender2[y]
        }
    }
    # priors for regression
    B0 ~ dunif(0,1)
    B1 ~ dnorm(0,0.01)
    B2 ~ dnorm(0,0.01)
    B3 ~ dnorm(0,0.01)
    B4 ~ dnorm(0,0.01)
    B5 ~ dnorm(0,0.01)
    sigma ~ dunif(0,100)
    tau <- 1/(sigma*sigma) # precision, i.e. 1/variance
}
```

where similar to previously, we have a binomial likelihood for the number of obese people per 10,000 people. However, here theta[,] is taken to follow a normal distribution with the mean being a function of the education and gender variables as its prior, and the mean being a linear function of the gender and education variables. Furthermore, the priors for the regression are taken to be non-informative priors with the following logic:

- B0 - uniform between 0 and 1, since as the intercept it can't be out of the range of possible values for theta
- B1 to B5 - normal around 0 with a large variance in order to be able to have all sorts of positive or negative effects
- sigma - uniform between 0 and 100 in order for it to have the freedom to be extremely large

Now moving forward, the inputs that the model takes are quite similar with the addition of:

- edu1/edu2/edu3 as vectors with binary values depending on the level of education. Mainly:

  - edu1 = edu2 = edu3 = 0 means we are looking at the average person
  - edu1 = 1, edu2 = edu3 = 0 means levels 0-2
  - edu1 = edu2 = 1, edu3 = 0 means levels 3-4
  - edu1 = edu2 = edu3 = 1 means levels 5-8
- gender1/gender2 as vectors with binary values depending on a person's gender. Mainly:
  - gender1 = gender2 = 0 means we are looking at the average person
  - gender1 = 1, gender2 = 0 means we are looking at a female
  - gender1 = gender2 = 1 means we are looking at a male

Afterwards the model was is through JAGS, providing us with posterior distributions for numerous variables. However, given that the posterior distributions of thetas are virtually the same as we have already seen, we will simply look at the most interesting additions, the beta values:

	mean	sd	CI - LL	CI - UL
B0	17.85%	1.18%	15.44%	20.20%
B1	2.01%	1.44%	-0.80%	4.91%
B2	-1.26%	1.49%	-4.28%	1.69%
B3	-5.50%	1.47%	-8.45%	-2.56%
B4	-0.21%	1.22%	-2.62%	2.22%
B5	0.34%	1.25%	-2.14%	2.85%

Table 6: Model 4 summary

Understanding output

- B0 represents the intercept, i.e. the value of mu[x,y] when B1 = B2 = B3 = B4 = B5 = 0 (the value of mu[x,y] for the average human)
- B1 represents the average increase in the rate of obesity when someone is not educated (levels 0-2)
- B2 represents the average increase in the rate of obesity when someone is part of the second group of education (levels 3-4), relative to the non-educated person
- B3 represents the average increase in the rate of obesity when someone is part of the third group of education (levels 5-8), relative to the someone in the second group
- B1 + B2 and B1 + B2 + B3 represent the average increase in the rate of obesity when someone is part of the second and third education groups respectively, relative to the average person
- B2 + B3 represents the average increase in the rate of obesity when someone is part of the third group relative to someone who is not educated
- B4 and B4 + B5 represent the average increase in the rate of obesity when someone is a female and a male respectively, relative to the average person
- B5 represents the average increase in the rate of obesity when someone is a male, relative to a female

Interpreting output

- Nothing concrete can be interpreted regarding the effect of gender since the confidence intervals of B4 and B5 include both negative and positive values
- B1, B2, and B3 show a consistent pattern in more education resulting in a lower probability of obesity, especially when considering college+ (B3) educated people, where it concretely shows that there is a decrease in the rate of obesity when going from levels 3-4 to levels 5-8.

6 Conclusion and next steps

The models that have been created have allowed us to conclude that there is certainly a relation between college+ education and a lower rate of obesity in humans. Furthermore, we were able to see that women tend to have a stronger relation between the rate of obesity and the level of education attained.

With regards to next steps, it would be interesting to:

1. Explore the reason behind the college+ educated people having substantially lower rates of obesity, and seeing if it is simply a correlation or if there is causation at play
2. Explore why the rates of obesity in women tend to be more greatly affected by the level of education than men
3. Explore the age class variable as well