Causal Inference Demo

Alex van Vorstenbosch

20 April, 2022

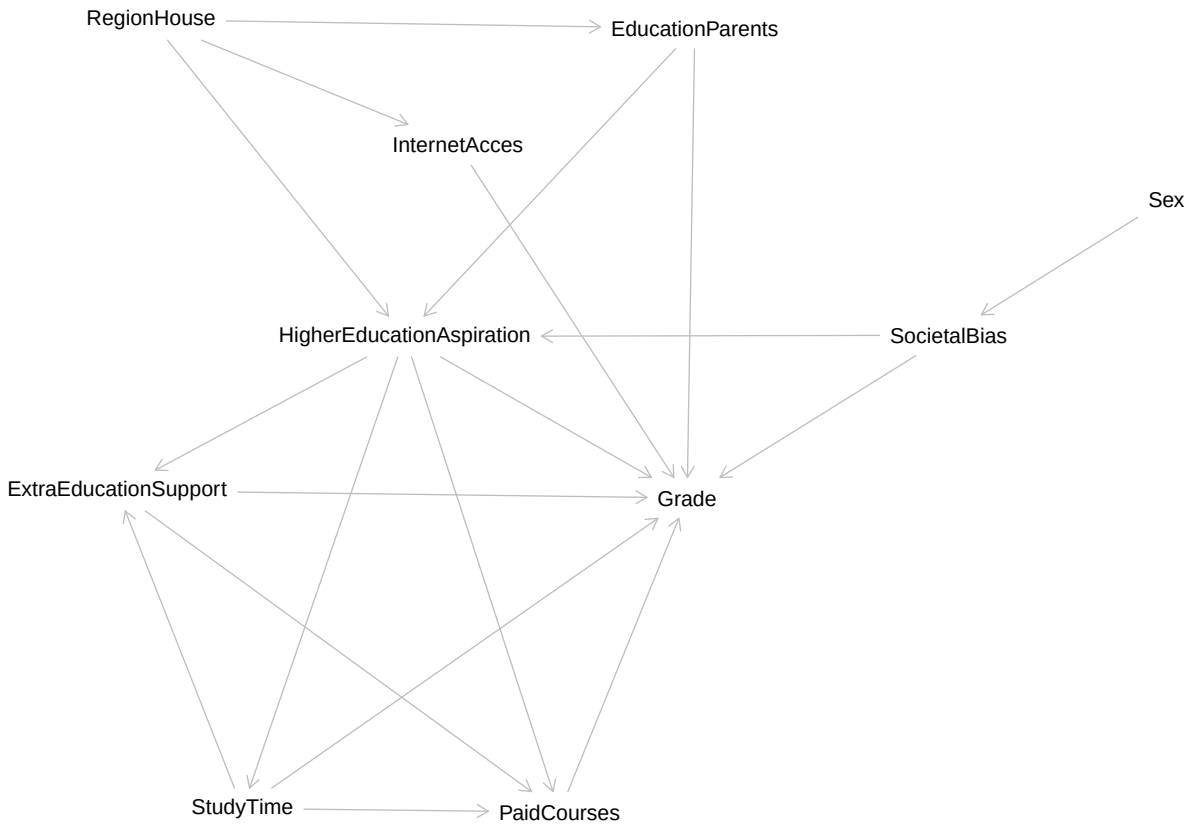
What are we going to see

- Defining the problem
- DAG
- Simulating the data
- Can we use BRMS to retreive the parameters
- **Real World Data...**
- Is our dag valid?
- Results

Defining the Problem

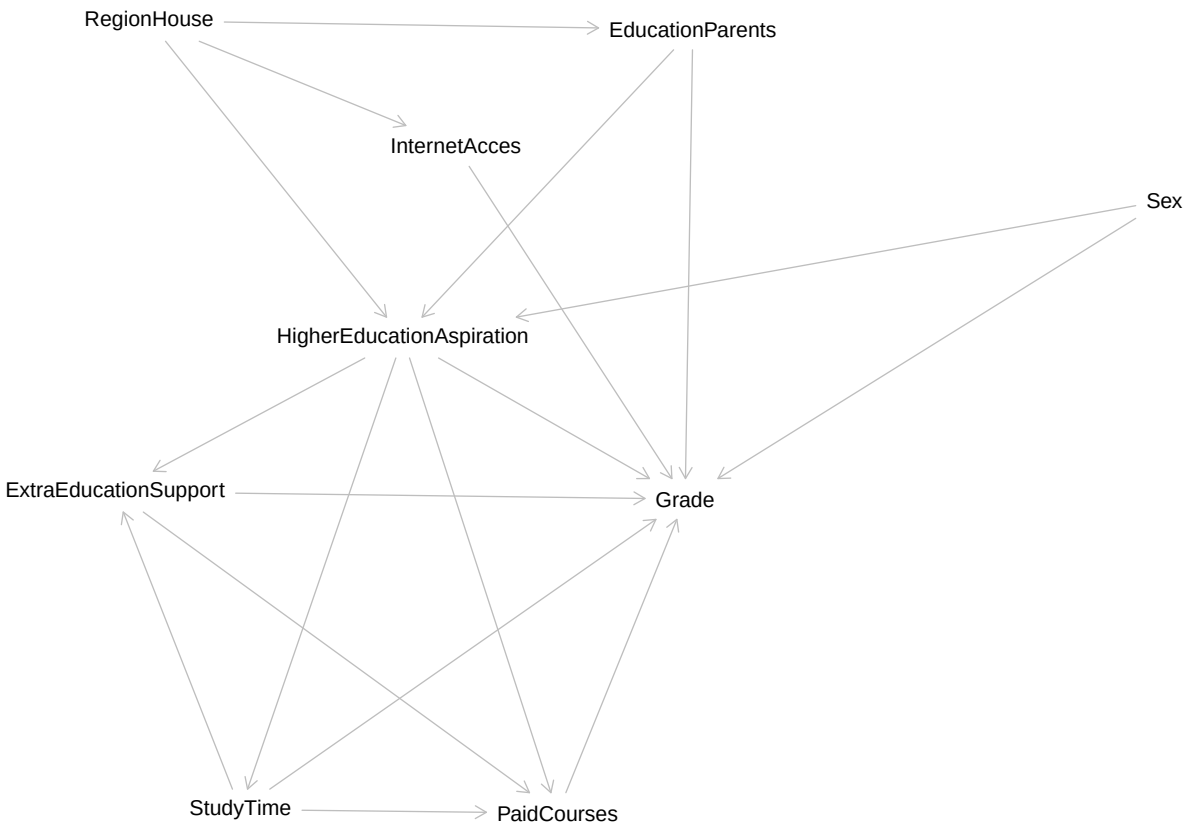
- Mathematics is on of THE fundamental school subject...
- ...yet, it is stereotypically seen as boring and difficult.
- In this demo we explore what influences school performance in Math
- ... and possible intervention paths we could use.
- As outcome variable we take the end of year grade in Mathematics for high school students.

DAG - The Full Causal Pentagram



The unmeasured variable SocietalBias explains how the student's Sex effects Grade and HigherEducationAspiration. However, since SocietalBias is only a 'Mechanism', we can remove it from the DAG and simplify it.

DAG - The (slightly simplified) Causal Pentagram



This DAG is slightly simpler, but the 'Causal Pentagram' looks like it might cause problems. Time to find out!

Simulate data - 1

```
#Starting in the upper half of the DAG
N=500
RegionHouse = sample(x = c("R", "U"), size=N, replace=TRUE, prob=c(0.3,0.7))

EducationParents = ifelse(RegionHouse=="U",
                          sample(x = 0:1, size=N, replace=TRUE, prob=c(0.4,0.6)),
                          sample(x = 0:1, size=N, replace=TRUE, prob=c(0.6,0.4)))

InternetAcces = ifelse(RegionHouse=="U",
                          sample(x = c(TRUE,FALSE), size=N, replace=TRUE, prob=c(0.85,0.15)),
                          sample(x = c(TRUE,FALSE), size=N, replace=TRUE, prob=c(0.95,0.05)))

Sex = sample(x = c("M", "F"), size=N, replace=TRUE, prob=c(0.5,0.5))
```

Simulate data - 2

```
CalcHigherEducation <- function(RegionHouse, EducationParents, Sex){
  # Samples HigherEducationAspiration based on 3 inputs.
  # Assumption: Growing up in a rural region you may be less motivated to go to college.
  # Assumption: Due to societal pressure, women might be less likely to persue college.
  HEA_base = 0.6
  Sex_effect = ifelse(Sex=="M", 0.05, -0.05)
  EducationParents_effect = ifelse(Sex==1,-0.1,0.1)
  RegionHouse_effect = ifelse(RegionHouse=="R", -0.05, 0.05)
  HEA_p = HEA_base + Sex_effect + EducationParents_effect + RegionHouse_effect
  HEA = rbernoulli(n=length(HEA_p) ,p=HEA_p)
  return(HEA)
}
HigherEducationAspiration = CalcHigherEducation(RegionHouse, EducationParents, Sex)

StudyTime = ifelse(HigherEducationAspiration==1,
                   sample(x = 0:3, size=N, replace=TRUE, prob=c(0.15,0.15,0.35,0.35)),
                   sample(x = 0:3, size=N, replace=TRUE, prob=c(0.3,0.35,0.25,0.1))
)
```

Simulate data - 3

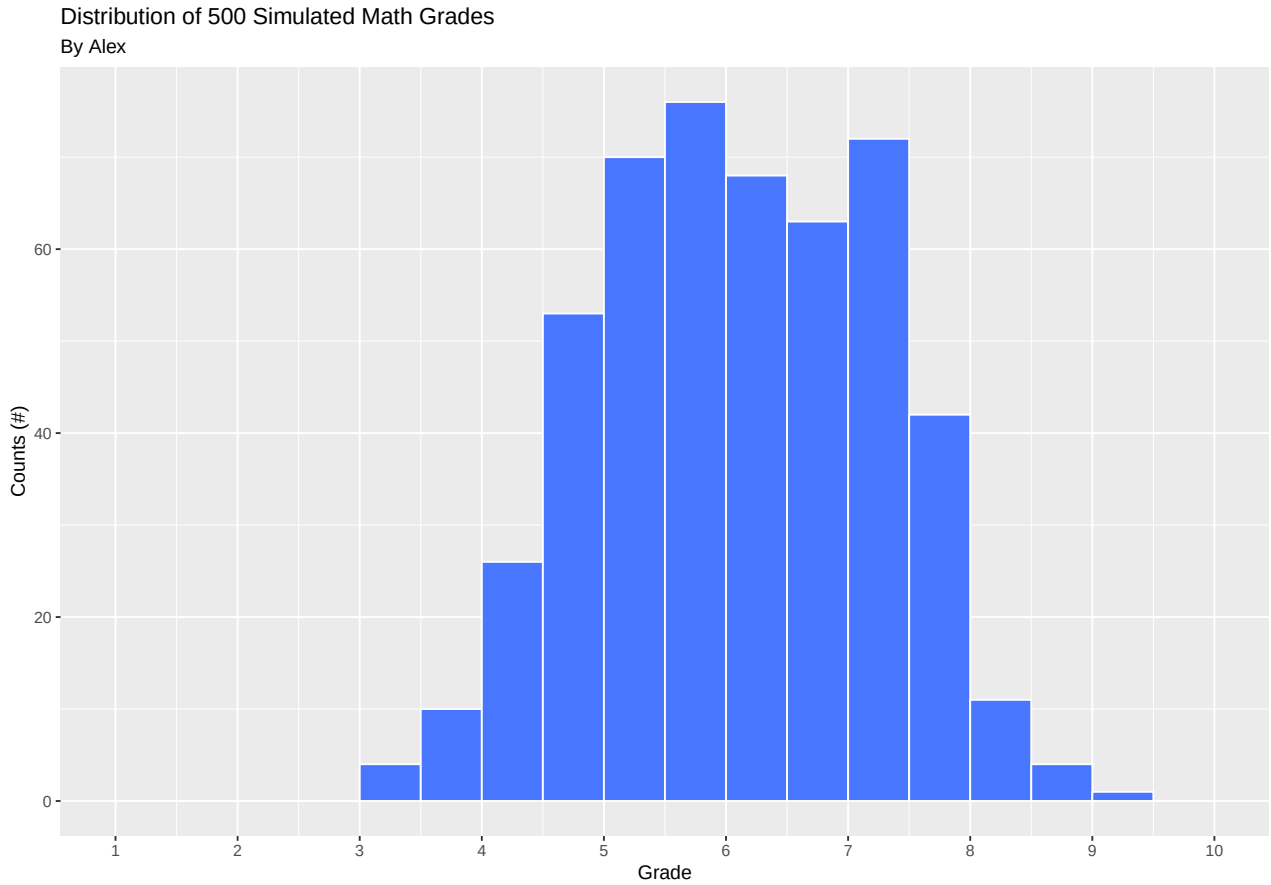
```
CalcExtraEducationSupport <- function(HigherEducationAspiration, StudyTime){
  # Samples ExtraEducationSupport based on 2 inputs
  # Assumption: If you are good at doing homework yourself (more hours), you get less support
  # Assumption: If you are motivated to go to college, you might seek out more support even if you study a lot
  EAS_base = 0.2
  HEA_effect = ifelse(HigherEducationAspiration==1, 0.075, -0.05)
  StudyTime_effect = -0.025*StudyTime
  EAS_p = EAS_base + HEA_effect + StudyTime_effect
  EAS = rbernoulli(n=length(EAS_p) ,p=EAS_p)
  return(EAS)
}
ExtraEducationalSupport = CalcExtraEducationSupport(HigherEducationAspiration, StudyTime)

CalcPaidCourses <- function(StudyTime, HigherEducationAspiration, ExtraEducationalSupport){
  # Samples PaidCourses based on 3 inputs
  # Assumption: If you study little, and get no extra support, but want to go to college,
  #             you might try to catch up using paid courses
  # Assumption: The opposite might also be true, college + many hours studying + extra support,
  #             may mean the student is more likely to also pay for extra courses
  Paid_base = 0.05
  effect = rep(0, length(StudyTime))
  for (i in length(StudyTime)){
    if(StudyTime[i]<2 & HigherEducationAspiration[i]==1 & ExtraEducationalSupport[i]==0){
      effect[i] = 0.1
    } else if(StudyTime[i]>3 & HigherEducationAspiration[i]==1 & ExtraEducationalSupport[i]==1){
      effect[i] = 0.1
    }
  }
  Paid_p = Paid_base + effect
  Paid = rbernoulli(n=length(Paid_p) ,p=Paid_p)
}
PaidCourses = CalcPaidCourses(StudyTime, HigherEducationAspiration, ExtraEducationalSupport)
```

True weights for calculating grades

```
# Finally, calculate grade
set.seed(925)
Grade = rnorm(n=500, mean=5.0, sd=0.3) +
  1*PaidCourses +
  0.5*StudyTime +
  1*ExtraEducationalSupport +
  0.7*HigherEducationAspiration +
  0.6*InternetAcces -0.6 + #convenient way to include nega
  ifelse(EducationParents==1, 0.5, -1) +
  ifelse(Sex=="F", -0.25, 0.25)
```

Simulated grades distribution

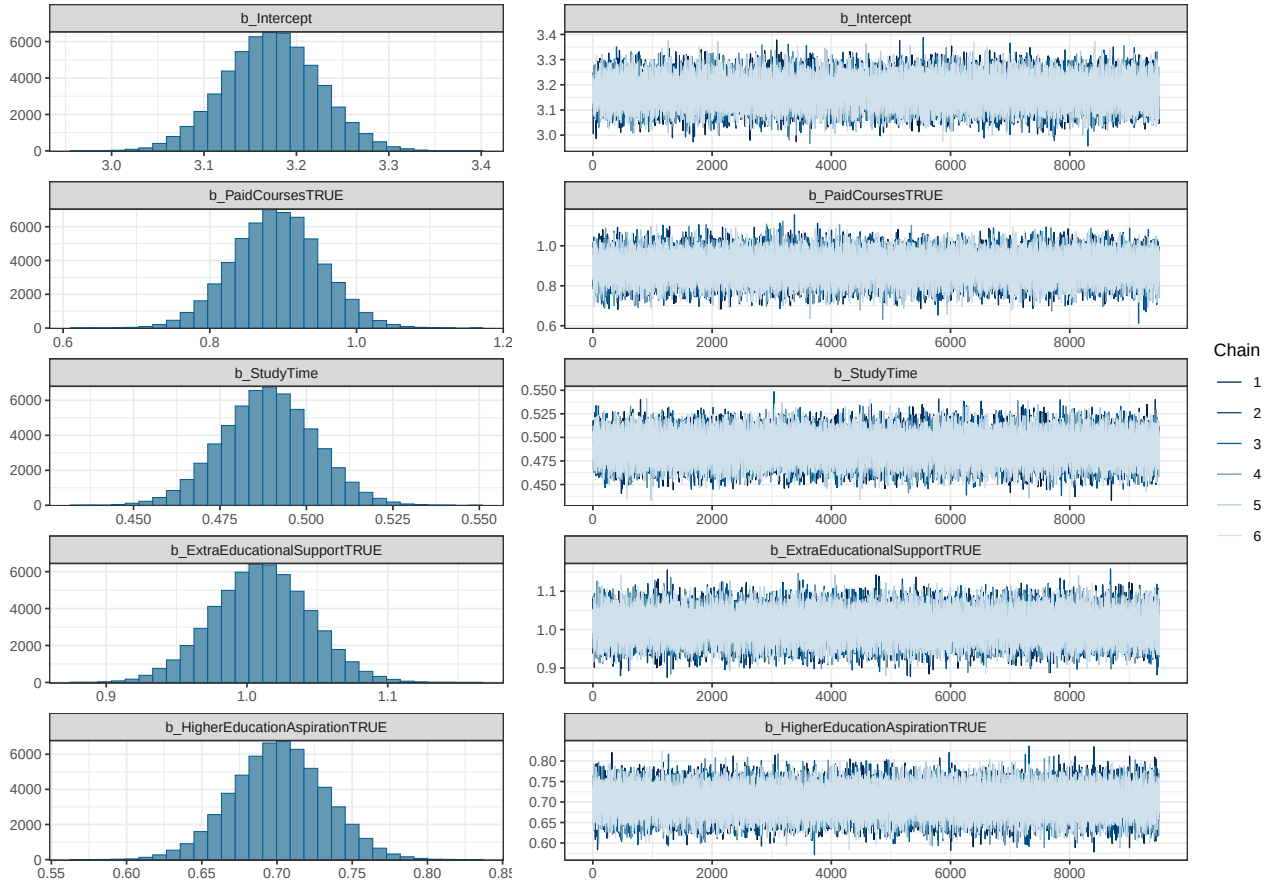


In our simulation, Most students score between a 5.5 and a 6. This seems reasonable. Also, very high scores are rare, as well as extremely low scores. The distribution is roughly normal, but not quite.

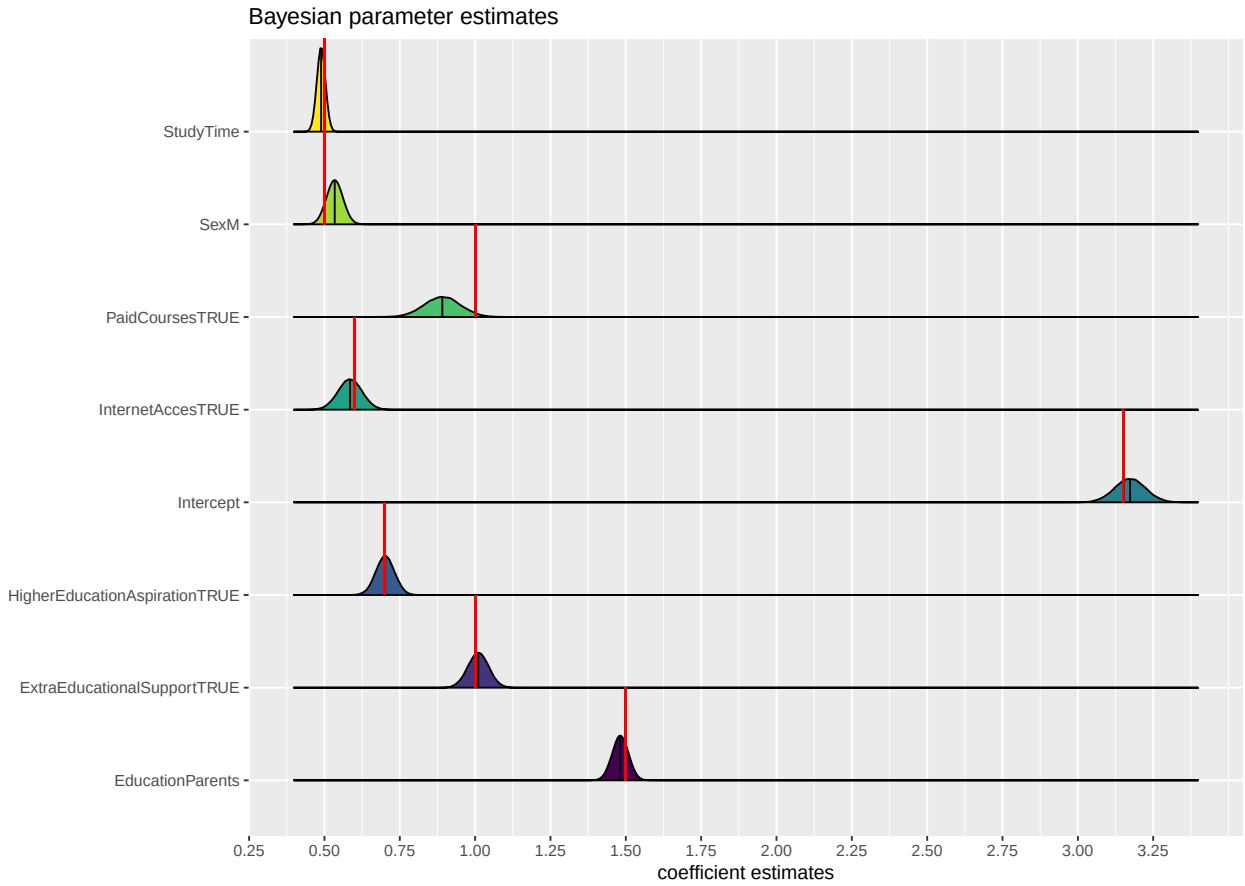
Retrieve these direct effects using BRMS

```
## Running MCMC with 6 parallel chains...
##
## Chain 1 finished in 4.0 seconds.
## Chain 3 finished in 3.7 seconds.
## Chain 4 finished in 4.9 seconds.
## Chain 2 finished in 5.2 seconds.
## Chain 6 finished in 6.7 seconds.
## Chain 5 finished in 7.8 seconds.
##
## All 6 chains finished successfully.
## Mean chain execution time: 5.4 seconds.
## Total execution time: 8.4 seconds.
```

Checking our posterior and chains!

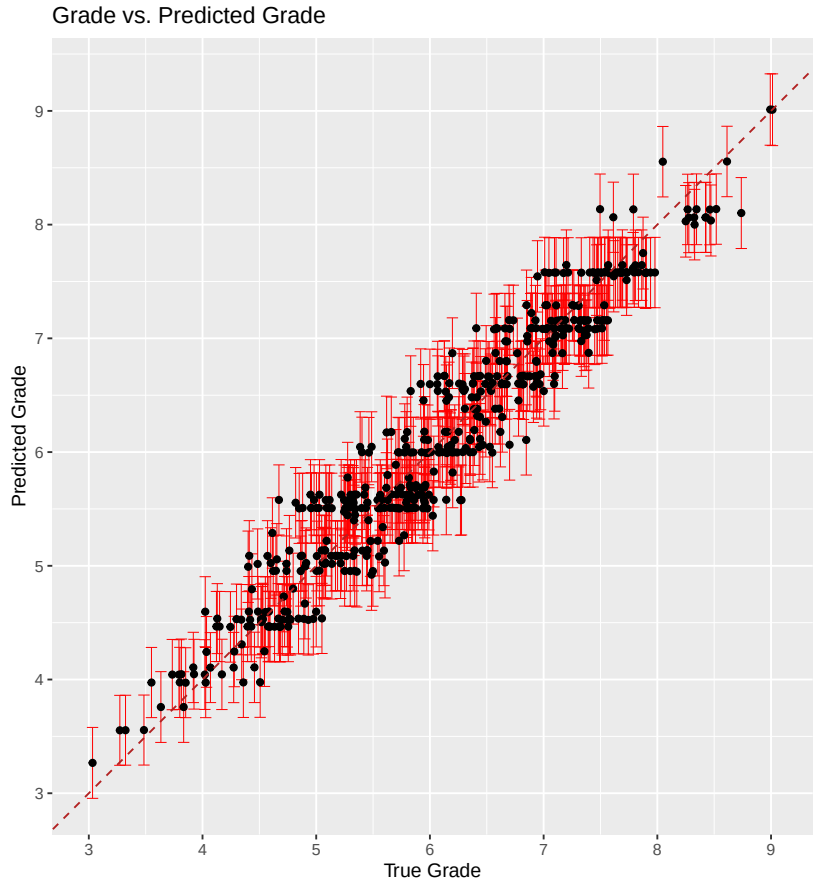


Bayesian Parameter Estimates



Almost all Parameter values are perfectly retrieved! PaidCourses is slightly underestimated, And Sex is slighty overestimated. But these results are very promising

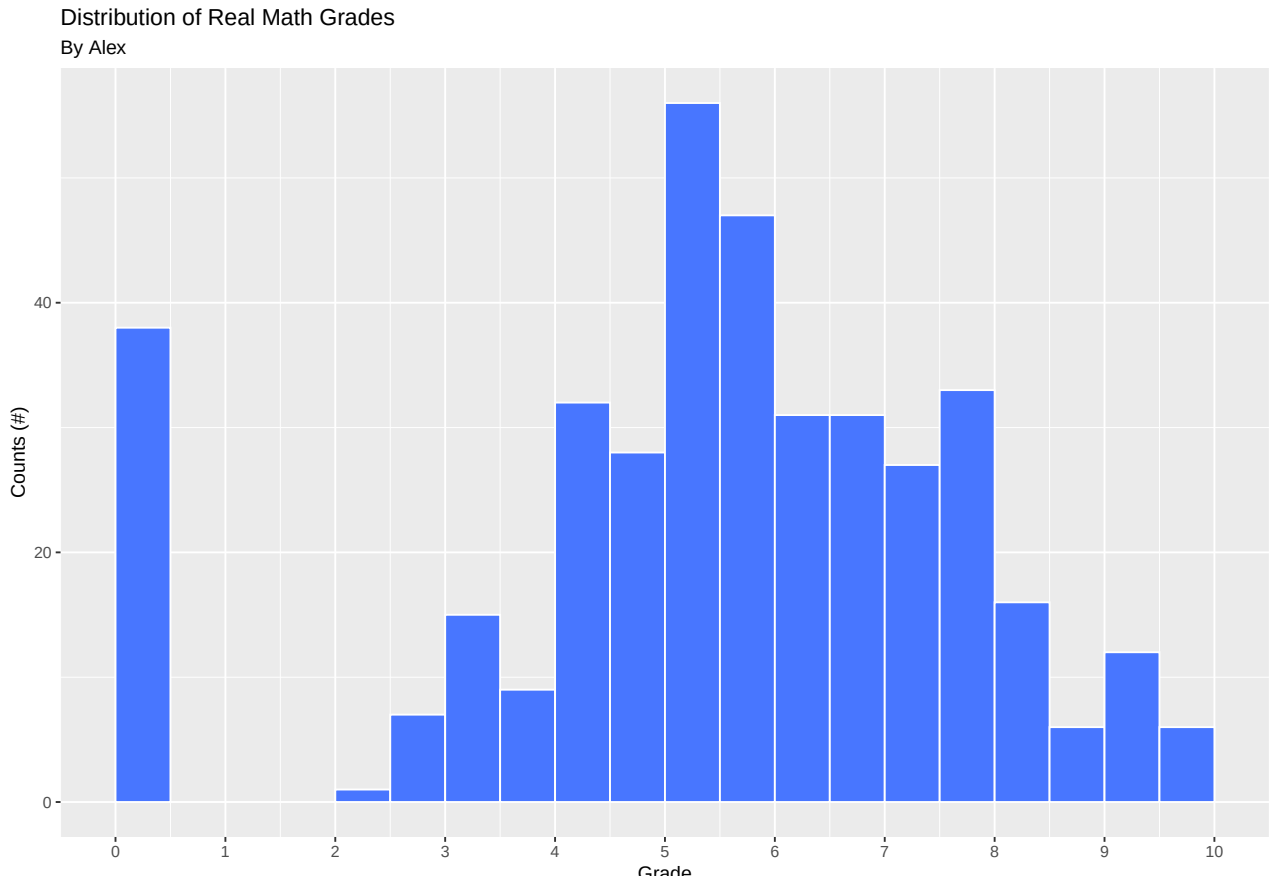
How does the model perform in terms of predictions?



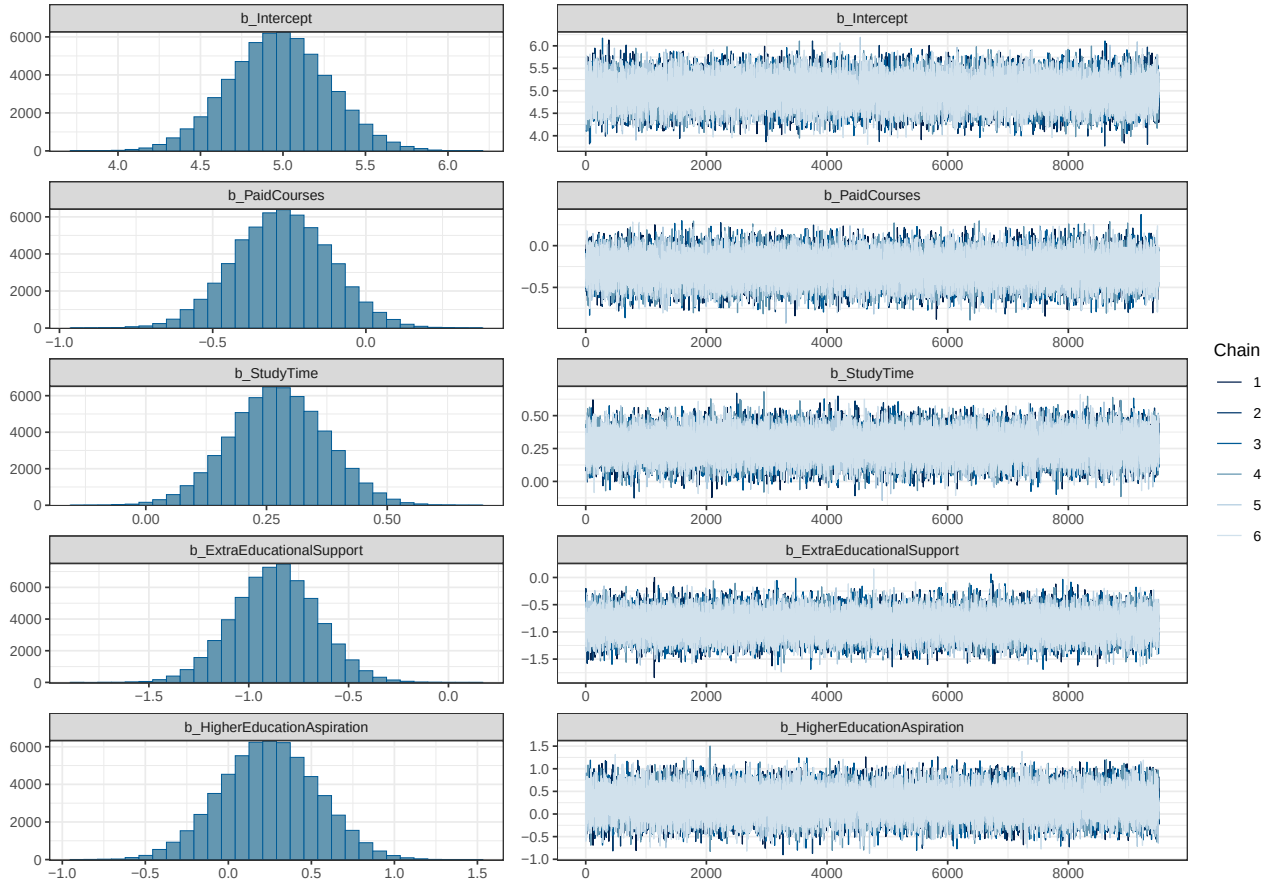
- This looks good! Most estimates overlap the true value within 1 sigma deviation.
- Estimates seem consistent across the whole domain

Time for real data

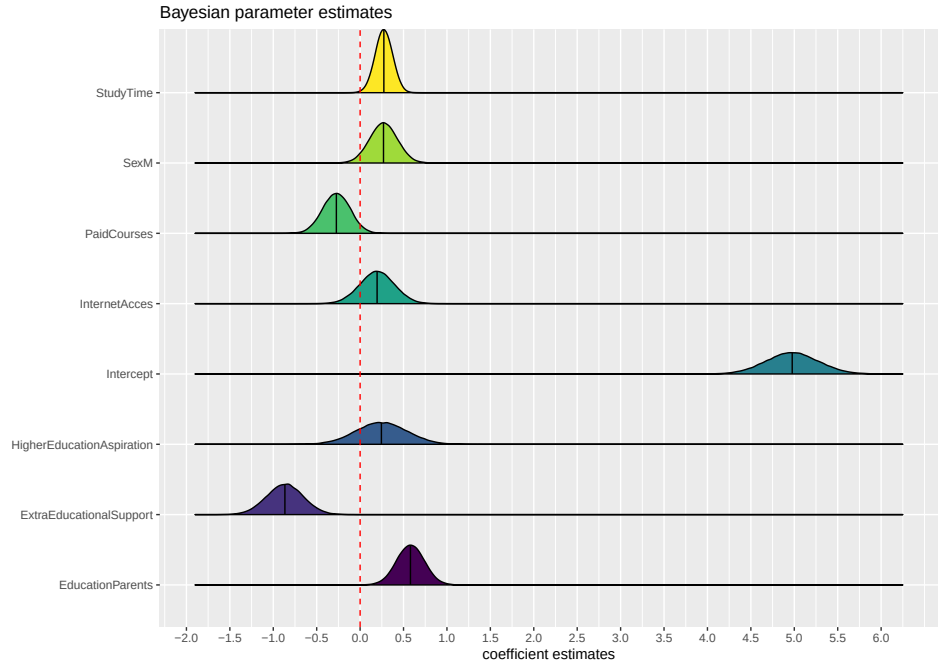
- We'll use https://www.kaggle.com/datasets/dipam7/student-grade-prediction
- This dataset contains portugese data on Math grades of 395 students
- Collected from 2 schools using Questionnaires and the schools grading Administration



Checking our posterior and chains!

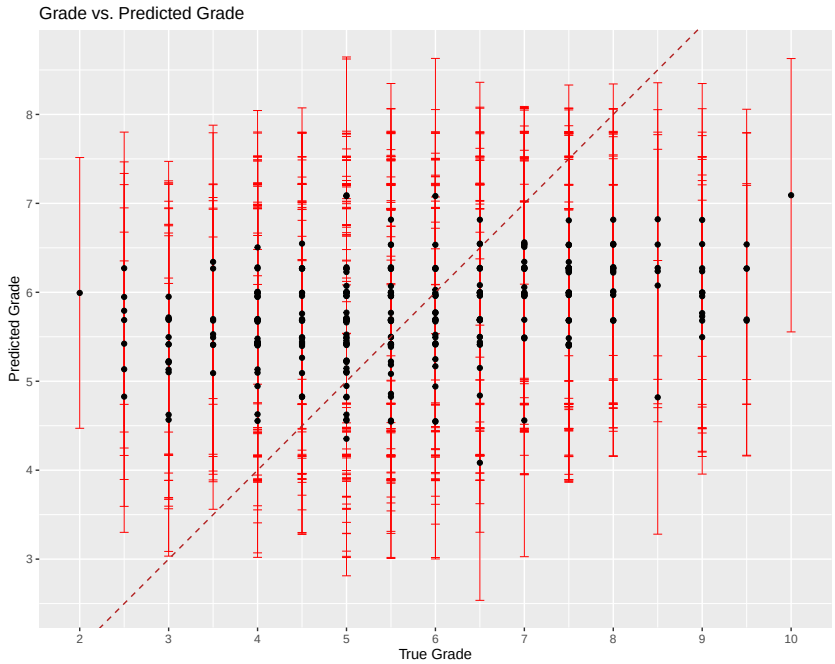


Bayesian Parameter Estimates on the real data



- Most parameters appear to have some effect
- InternetAcces, HigherEducationAspiration and PaidCourses are questionable
- The effect of ExtraEducationalSupport and PaidCourses is negative Hypothesis: Only ‘problematic’ students receive this support, Biasing the inference

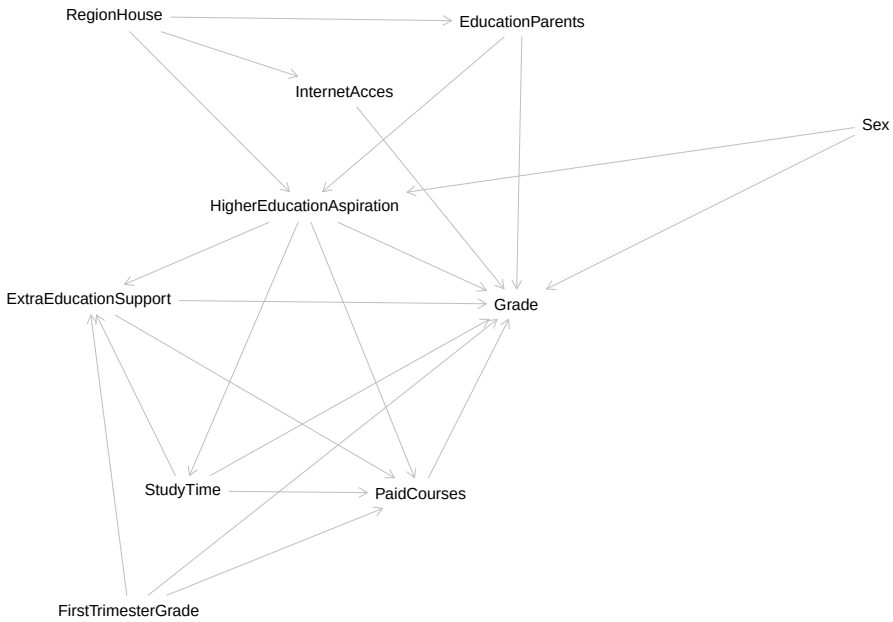
How does the model do, predicting on the real data?



- Hmm, our model has very little predictive power
- As is often the case, real life is more complicated than our toy model

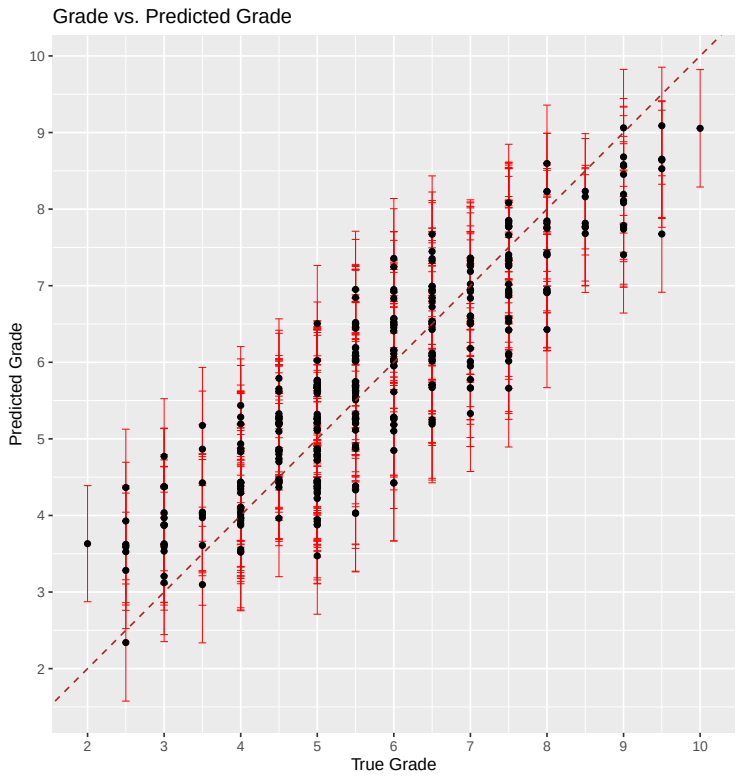
Appendix

Updated DAG



- What happens when we include the start of the year grade? Will it fix our problems?
- When we don't include FirstTrimesterGrade, it's influence flows via ExtraEducationSupport and PaidCourses.

Better, but not good enough



- As expected, previous grades are a good indication of the final grade.
- Other than that, the model is clearly broken. More work is needed!