BDA Project

Anonymous

Contents

1. Project Introduction

Studying the relationship between income and education has been the focus of many studies. The studies have concluded that a strong connection exists between higher education and income (Card, 1999). In general, individuals with stronger education are more likely to be employed and earn a big salary compared to less educated people (Card, 1999). For that reason, education is described as an investment in human capital (Wolla & Sullivan, 2017).

This study examines this phenomenon from the perspective of people that have acquired their education from colleges in the United States. As the connection between education and income has been shown in the existing literature, this study strives to further examine the associations between college related features and income level years after graduation. This project is not limited to only considering educational aspects but expands it to family backgrounds.

In this study, a Bayesian approach is taken to observe the bond between the educational and family related features and earnings in a multivariate linear regression setting. It is in our interest to find out how accurately the selected features can predict future income for the students. Aim of our project is to find a way to predict average income after studying at a university for any non-specialized university in the United States. We build three different multivariate models – separate, pooled and hierarchical – to predict income after school.

As this study is conducted in a university environment by university students and presented mainly to other students and faculty, the findings of this study could be especially meaningful for the members of the group and peers on the course.

2. Data Description and Analysis Problem

The used dataset is the most recent institutional-level college scorecard data from the US Department of Education[1]. The institutional-level dataset contains aggregate data for each educational institution and includes data on institutional characteristics, enrollment, student aid, costs and student outcomes. The dataset has 6681 observations on 2989 variables.

This dataset was chosen because it seemed to provide information that could answer interesting education-related questions in the project, and also because the dataset has a large number of observations and a large number of variables. The large amount of variables permits for a lot of flexibility when it comes to modelling with the data and also having enough data is important in order to be able to make valid inferences.

After investigating the dataset, the most intriguing analysis problem was to study how college related factors affect the median earnings of alumni 10 years after entry, which is our dependent variable:

Name	Data type	Description
MD_EARN_WNE_P10	double	Median earnings of students 10 years after entry

To answer the analysis problem, a subset of college related factors is extracted and generated from the dataset and after that weigths for each factor are calculated based on three models that are introduced later in this document.

In the next section, we outline how the independent variables, hereafter features, were selected

3. Feature Selection

The initial data set has almost 3000 features, which is a large number compared to the total number of observations of 6681. Moreover, there are many missing values for certain variables in the dataset and not all variables are interesting when trying to predict future earnings. For these reasons, there was a clear need to prune down features.

[1] Available at: https://collegescorecard.ed.gov/data

We conducted feature selection in three phases: in the first phase, a subset of features was selected based on features commonly used in previous studies that have examined the relationship between educational factors and earnings. In addition, we tried to come up with additional interesting features that could predict future earnings. In the second phase, we assessed the relationship of the features with our dependent variable, and accounted for any multicollinearity arising from mutually correlated features. Furthermore, we visualized categorical variables and engineered new features to be added to the model. In the last phase, we utilized stepwise regression in streamlining our model to include only the most significant features in terms of the Akaike Information Criterion (AIC).

Phase 1 - Feature selection based on literature

The initial set of features chosen based on past literature and general intuition is outlined in the table below. (Dominic et. Al, 1996), (David, Krueger, 1996)

	Name	Data type	Description
1	SATVRMID	integer	Midpoint of SAT scores at the institution (critical reading)
2	SATMTMID	integer	Midpoint of SAT scores at the institution (math)
3	SATWRMID	integer	Midpoint of SAT scores at the institution (writing)
4	MD_FAMINC	double	Median family income
5	AGE_ENTRY	double	Average age of entry
6	FEMALE	double	Share of female students
7	FIRST_GEN	double	Share of first-generation students
8	PCT_WHITE	double	Percent of the population from students' zip codes that is White
9	DEBT_MDN_SUPP	integer	Median debt, suppressed for n=30
10	C150_4	double	Completion rate for first-time, full-tim students
11	COSTT4_A	integer	Average cost of attendance (academic year institutions)
12	POVERTY_RATE	double	Poverty rate
13	UNEMP_RATE	double	Unemployment rate
14	MARRIED	double	Share of married students
15	VETERAN	double	Share of veteran students
16	LOCALE	categorical	Locale of institution
17	CCBASIC	categorical	Carnegie Classification – basic
18	CONTROL	categorical	Control of institution

SAT scores were aggregated into one composite variable called SAT_ALL, because SAT components were highly correlated and could easily be viewed as one entity. However, writing SAT scores had too few observations due to discontinued tracking. Therefore, SAT_ALL was formed by summing the math and critical thinking SAT scores. After this, we ended up with a total of 793 observations for the feature subset.

Descriptive statistics for the resulting numerical variables can be found from the table below. Min is the minimum, 1st Qu. is the 1st quartile, Median is the median, Mean is the arithmetic mean, 3rd Qu. is the 3rd quartile, Max is the maximum and St.dev is the sample standard deviation.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	St.dev
MD_EARN_WNE_P10	24209.00	43093.00	49240.50	51246.87	56516.75	103246.00	11587.58
MD_FAMINC	10702.00	35386.12	44858.50	48915.83	61402.38	121852.50	18732.71
AGE_ENTRY	19.55	20.92	22.04	22.53	23.51	33.82	2.18
FEMALE	0.12	0.54	0.59	0.59	0.64	0.98	0.10
FIRST_GEN	0.10	0.27	0.34	0.33	0.39	0.62	0.09
PCT_WHITE	24.24	71.45	80.24	77.62	87.62	95.96	13.09
DEBT_MDN_SUPP	3688.00	13971.00	16000.00	16190.37	19000.00	26800.00	3595.32
C150_4	0.09	0.46	0.56	0.57	0.68	0.98	0.16
COSTT4_A	11704.00	23250.25	31904.00	35719.42	45932.00	79750.00	15267.87
POVERTY_RATE	3.58	6.44	7.74	8.68	9.87	45.73	3.81
UNEMP_RATE	2.12	3.01	3.34	3.48	3.78	7.92	0.74
MARRIED	0.01	0.03	0.06	0.08	0.10	0.38	0.07
SAT_ALL	395.50	513.25	545.00	553.57	584.62	760.00	59.93

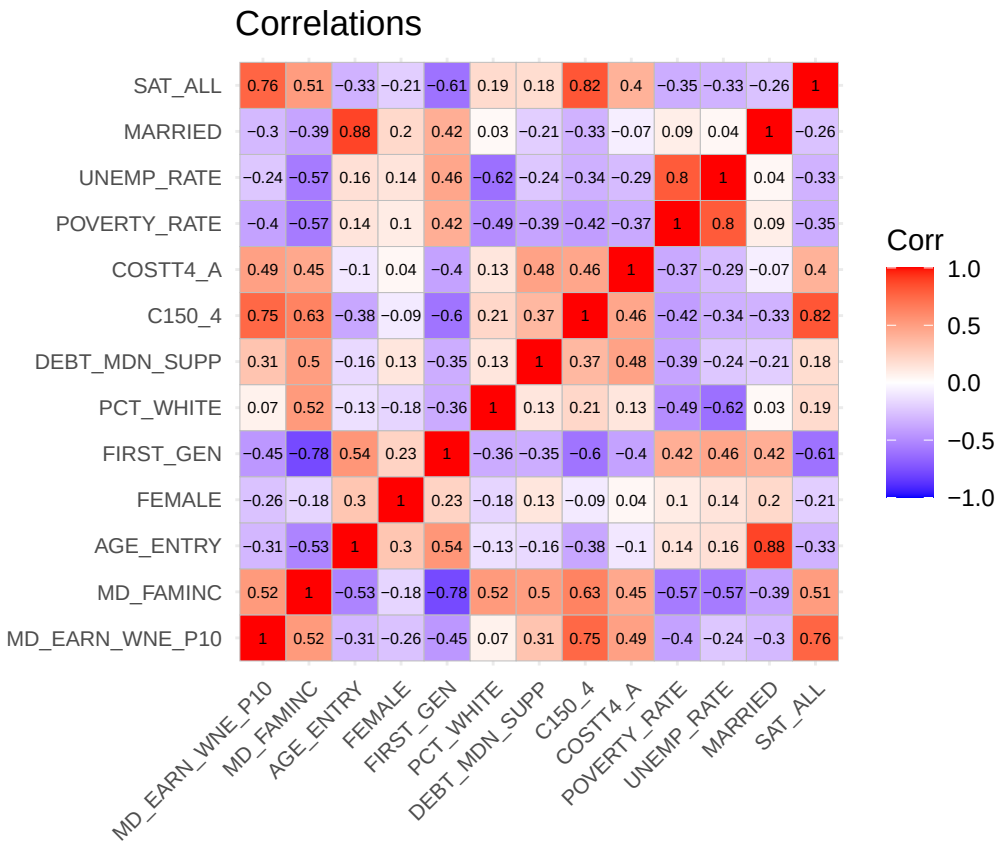
The table shows that our dependent variable - median earnings 10 years after entry - range from \$24.2k to \$103.2k. Overall one can see that there is great variation in almost all features as well. For example, age of entry varies from 19.6 to 33.8, average composite SAT scores range from 395 to 760, and cost of attendance varies from \$11.7k to \$79.8k, to name a few interesting details.

Phase 2 - Feature Selection with Correlation and Linear Association

In the second phase of feature selection, a subset of features was selected from the 18 variables resulting from the first phase. We examined the correlations between the dependent variable and the features as well as between-features correlations. Moreover, we examined the linear relationships between the features and the dependent variable. Lastly, we visually examined categorical variables and engineered new features that we believed could have predictive ability over the dependent variable.

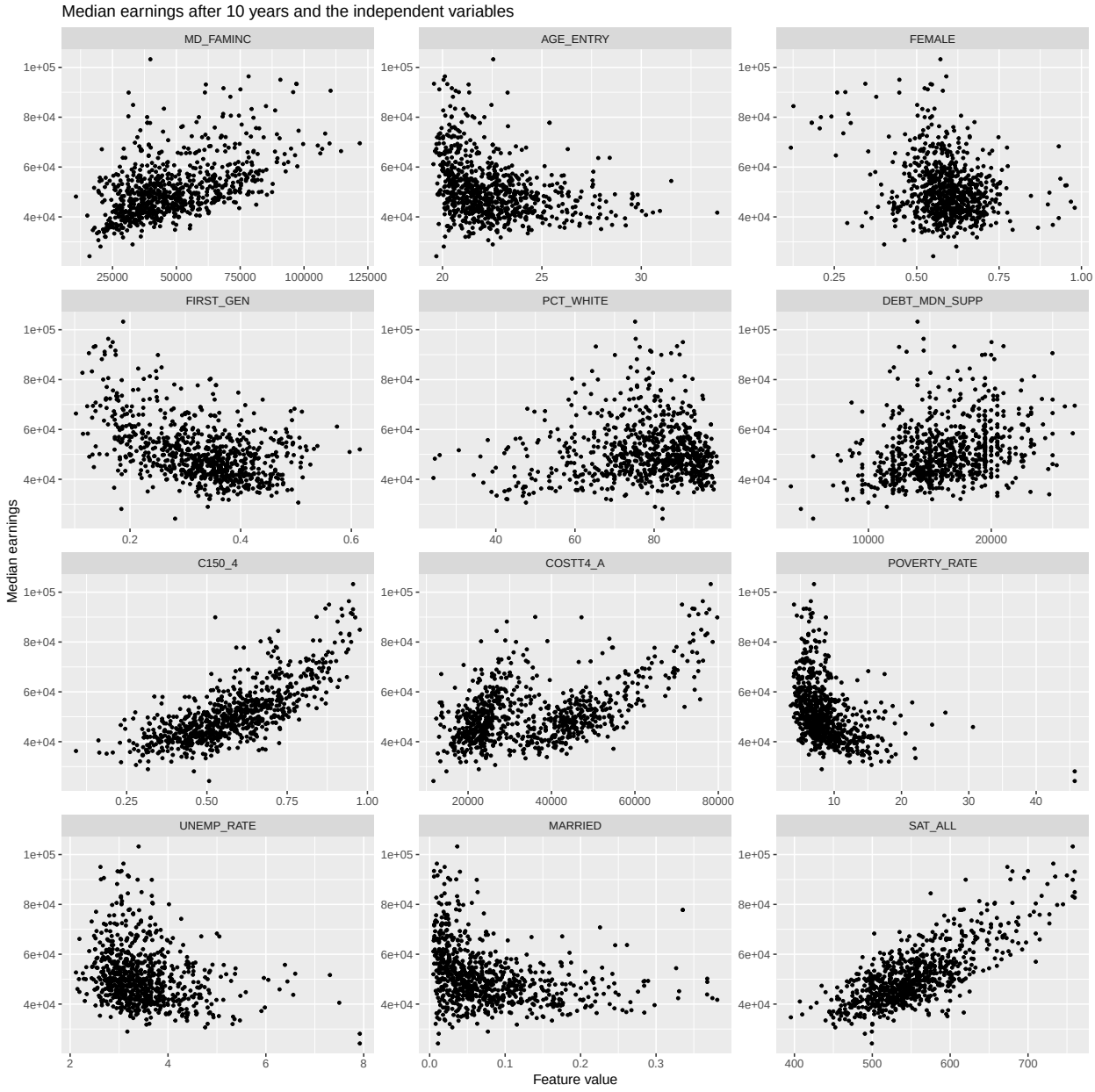
Numerical Variables

Correlation Analysis The figure below shows the (Pearson's) correlation matrix for the dependent variable and the features.



The correlation matrix shows on the bottom row that SAT scores, cost of attendance, and family income are the most positively correlated features with our dependent variable. The respective correlations are 0.76, 0.75, and 0.52. Most negatively correlated features are the percentage of first generation students, poverty rate, and average age of entry with respective correlations of -0.45, -0.4, and -0.31. These are interesting observations and make intuitive sense as well.

Bivariate Plotting The panel of figures below show scatter plots between the dependent variable and all numerical features. The x-axis represents the value of the feature and the y-axis represents the value of the dependent variable.



The figures show that many of the features exhibit a linear like relationship with the dependent variable. Especially, SAT scores (SAT_ALL) show a clear linear association. Other features that show a linear trend are e.g., family income (MD_FAMINC), completion rate (C140_4), and the percentage of first generation students (FIRST_GEN). There are hints of some non-linear relationships as well, e.g., poverty rate appears to have a convex non linear relationship with the dependent variable. Cost of attendance (COSTT4_A), on the other hand, suggests that there could be two groups, perhaps cheaper public school and private school.

Summary of Numerical Feature Selection Based on the analysis presented thus far, we selected the following numerical variables: SAT_ALL (median sum of math and critical thinking SAT scores), MD_FAMINIC (median family income of the student), AGE_ENTRY (median age of starting at the college), COSTT4_A (median cost of college), and POVERTY_RATE (poverty rate in the area the college is located).

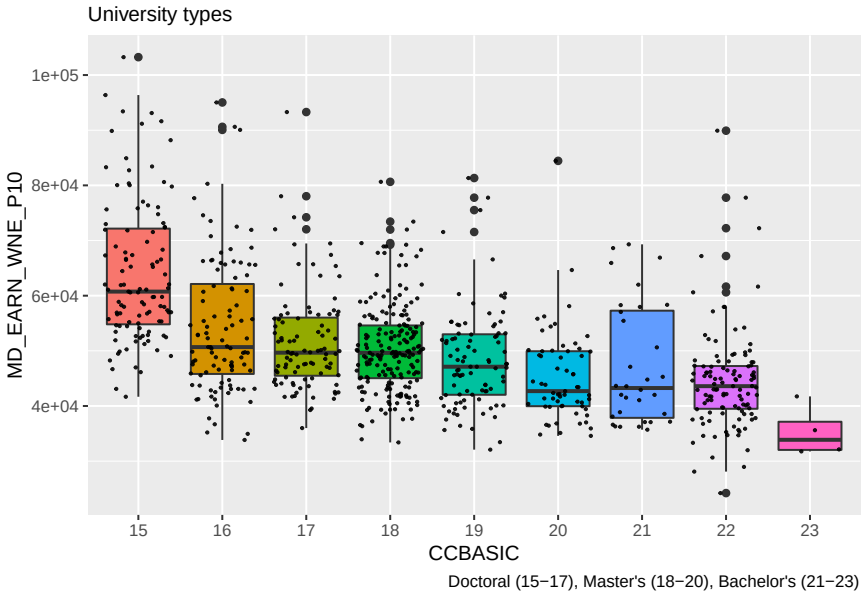
After including the obvious features, such as SAT scores, we included features with the following reasoning: if the correlation between a feature and the dependent variable was low and there was no observable dependency

between the two in the scatter plot, the feature was excluded. Furthermore, to avoid multicollinearity, we removed features that were highly correlated with other independent variables, especially if they were not clearly correlated with the dependent variable and we couldn't form a believable hypothesis for the mechanism through which that variable affected income after college.

Categorical Variables

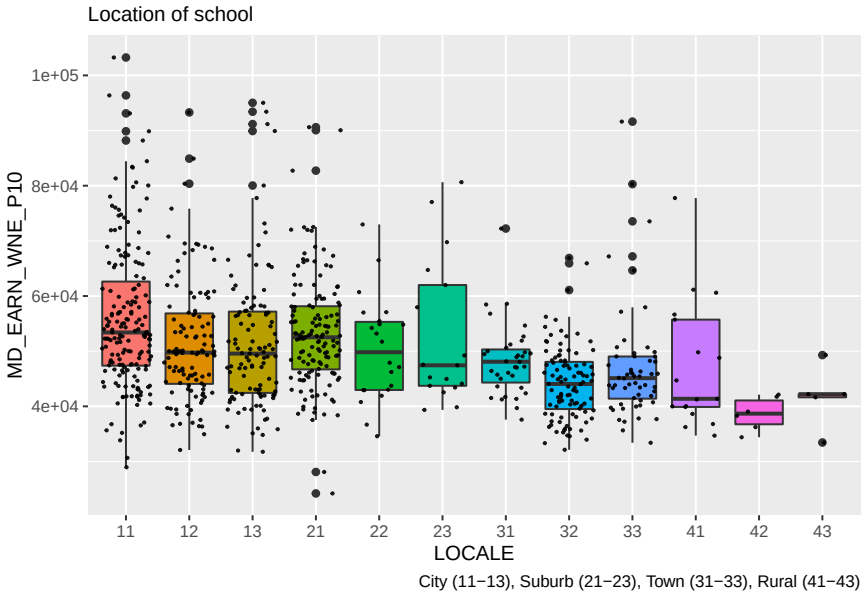
Box Plots The categorical variables were visualized with box plots to assess if income after college differs among the categories.

The figure below shows the median earnings for different university types. CCBASIC stands for the Carnegie Classification for the university.



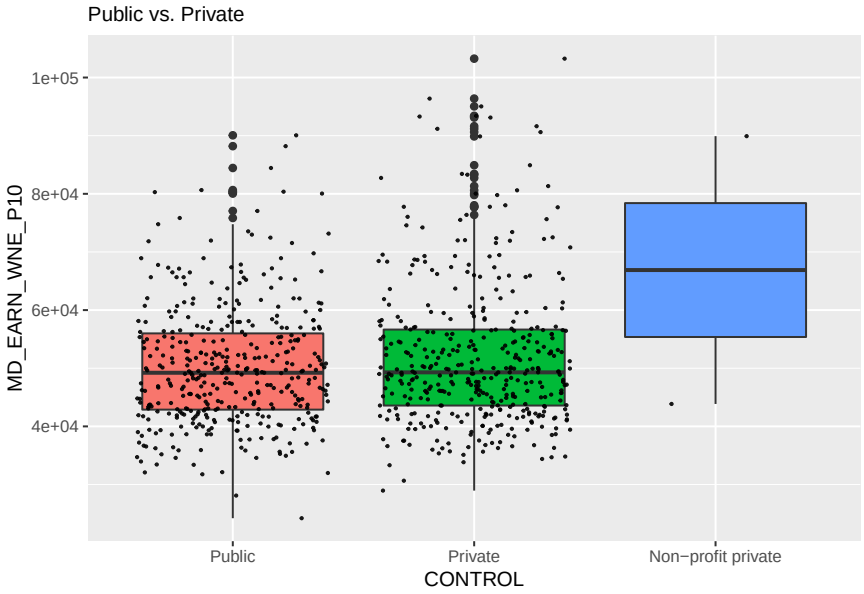
Based on the plot, CCBASIC was divided into two dummy variables: MASTER and DOCTORAL. These variables represent if the college is classified as Master's college and university or Doctoral's college and university. If a college does not belong to either of those, it is a Bachelor's college and university. Other special focus colleges and universities were discarded from the dataset to keep the model simpler and to avoid unnecessary outliers due to unconventional nature of some very specialized colleges. Our model does not attempt to provide accurate predictions for specialized colleges.

The figure below shows the median earnings for different university locations. LOCALE specifies whether the university locates in a metropolis, city, suburb, town, or a rural area.



Based on figure, we engineered a new dummy variable URBAN, which takes the value 1 if the university LOCALE is not classified as Rural.

The categorical variable CONTROL had three classes: public, for-profit private and nonprofit private. The figure below shows the median earnings depending on the control status.



Based on the figure, CONTROL was modified into a dummy variable PRIVATE, which takes the value 1 if the school is privately controlled (either for profit or non profit).

The selected categorical variables were: URBAN, DOCTORAL, MASTER, PRIVATE.

In the third and final phase, all remaining variables were used with stepwise regression to test which subset of features perform the best, whether the received coefficients are reasonable, and if there are signs of overfitting.

The stepwise regression suggested using all of the variables, except AGE ENTRY and DOCTORAL. The output of the stepwise regression can be found in the Appendix. Because stepwise regression suggested leaving out DOCTORAL, for the sake of consistency we decided to also leave out MASTER as it was a the middle

level in our institutional classifications and wouldn't have made much sense on its own. To test whether AGE ENTRY would be included by stepwise after removing MASTER and DOCTORAL, stepwise was ran again. Even then, AGE ENTRY was discarded from the model and the other variables remained the same.

The final features are listed in the table below:

	Name	Data type	Description
1	SAT_ALL	float	Midpoint of SAT scores at the institution (critical reading ,math)
2	MD_FAMINC	float	Median family income
3	COSTT4_A	float	Average cost of attendance (academic year institutions)
4	POVERTY_RATE	float	Poverty rate
5	URBAN	binary	Urban or rural area
6	PRIVATE	binary	Private or public school

3. Model Descriptions

Description of the separate model

In the separate model, posteriors for the parameters are constructed. In the context of the project, the separate model considers all regions independent from each other, meaning that each region have individual parameters (mu, sigma).

Mathematical description:

$y_{ij}|\mu_j, \sigma_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$

where $\mu_j = \alpha_j + \beta_\mathbf{j}\mathbf{X}$

The parameters of the parameter vector are given in the section 4 with their priors.

Description of the pooled model

As in the separate model, the pooled model constructs posteriors for the parameters. However, the regions are considered as one entity meaning that all regions share the same distribution and parameter values.

Mathematical description:

$y_i|\mu, \sigma \sim \mathcal{N}(\mu, \sigma^2)$

where $\mu = \alpha + \beta\mathbf{X}$

The parameters of the parameter vector are given in the section 4 with their priors.

Description of the hierarchical model

Contrary to the other two models, in the hierarchical model posteriors are constructed for the prior parameters. With the hierarchical model, the regions are considered as individual but similar. In the context of the project, the regions share same sigma and the parameters forming mu are similarly distributed (sharing the hyperparameters).

Mathematical description:

$y_{ij}|\mu_j \sim \mathcal{N}(\mu_j, \sigma^2)$

where $\mu_j|\mu_P, \sigma_P^2 \sim \mathcal{N}(\mu_P, \sigma_P^2)$

The hyperparameters and parameters are given in the section 4 with their priors.

4. Choice of Priors

We use weakly informative priors, as we do not posses enough information about the dependency between the independent variables and the dependent variable. When selecting the priors, proper distribution type was considered for each variable. The prior standard deviations were selected based on the magnitude of absolute values of the variables and what kind of impact they could have for income, with loose enough estimates to avoid limiting the values too much. Alternative approach could have been to standardize all variables, which would have reduced the need to think about magnitudes of absolute values.

SAT_ALL ~ Normal(43, 500)

Justification: We agreed that it was somewhat reasonable to expect scores in the SAT exam to be associated with higher income. Due to our approach of using weakly informative priors, we set a high standard deviation, but set a positive mean due to expected positive association.

The mean is calculated with the following formula: Median individual income in the United States / Average SAT Score (math and writing).

The median individual income in the US was approximately 31 000 in 2020 (Data Commons, 2020). The average SAT score in the US in 2020 were 523 in Math and 582 in Evidence-Based Reading and Writing (Number2, 2020).

mu = 31 000 / (523 + 528/2) = 43.2

MD_FAMINIC ~ Normal(0, 100)

Justification: Weakly informative prior is again selected as we don't posses enough information on the dependency. The absolute values are in general high, (for example compared to AGE_OF_ENTRY) and thus the standard deviation is set lower for this variable. However, there could be cases where MD_FAMINIC is low, due to for example unemployment, so the standard deviation is still set considerably high.

COSTT4_A ~ Normal(0, 500)

Justification: Weakly informative prior is selected as we don't posses enough information on the dependency. The average cost per academic year is likely to take lower values than median family income, but higher values than average age of entry and thus the standard deviation is set between the standard deviations of those.

POVERTY_RATE ~ Normal(0, 2500)

Justification: The possible values are between 0 and 100. There could be situations were poverty rate in an area is really low, for instance 0.5%. For that reason, the standard deviation in the prior is set high to enable possibly high weight for a small value.

URBAN ~ Normal(0, 2500)

Justification: Weakly informative prior was selected for URBAN. Although, urban areas could potentially have higher wages on average, in this project, we are considering earnings ten years after entry, which means that there might have been a lot of people moving from rural areas to urban areas after graduation and vice versa. For that reason, the prior is not set to be too informative. Because the values are always either 0 and 1, the standard deviation is set high.

PRIVATE ~ Normal(0, 2500)

Justification: For MASTER and PRIVATE weakly informative prior, is selected as we don't posses enough information on the dependency with the dependent variable. Because the values are always either 0 and 1, the standard deviation is set high.

Alpha ~ Normal(0, 10000) Sigma ~ Normal(10000, 10000)

Justification: For Alpha and Sigma, weekly informative priors were selected as we did not posses enough information.

Hyperpriors:

$Alpha_p$ ~ Normal(100, 100)

SAT_p ~ Normal(100, 100)

$MD\_FAMINIC$ ~ Normal(10, 100)

$COSTT4\_A_p$ ~ Normal(50, 500)

$POVERTY\_RATE_p$ ~ Normal(500, 500)

$URBAN_p$ ~ Normal(500, 500)

$PRIVATE_p$ ~ Normal(500, 500)

$Sigma$~ Normal(500, 500)

Justifications for the hyperpriors: When selecting hyperpriors it was considered that the constructions would be consistent in a sense that there would not be conflicting levels of informativeness.

5. Stan Code

Before fitting the models in Stan, we split our data set of 793 observations into a training set of 763 observations and a test set of 30 observations. The test set is used in assessing the predictive ability of our models in Section 9.

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(categorical.vars.model)` instead of `categorical.vars.model` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

We used cmdstanr for our models. Source code for all three models can be found in in the Appendix.

Separate model

Summary of model fit for main parameters:

```
## Running MCMC with 4 sequential chains...
##
## Chain 1 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 1 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 1 finished in 1172.2 seconds.
## Chain 2 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 2 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 2 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 2 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 2 finished in 1022.9 seconds.
## Chain 3 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 3 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 3 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 3 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 3 finished in 1044.3 seconds.
## Chain 4 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 4 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 4 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 4 Iteration: 2000 / 2000 [100%]  (Sampling)
```

```
## Chain 4 finished in 924.0 seconds.
##
## All 4 chains finished successfully.
## Mean chain execution time: 1040.9 seconds.
## Total execution time: 4164.2 seconds.
```

```
## # A tibble: 1,605 x 10
##    variable    mean  median      sd    mad      q5     q95  rhat ess_b~1 ess_t~2
##    <chr>      <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl> <dbl>   <dbl>   <dbl>
##  1 lp__     -62791. -62791.  6.01e0 5.78e0 -62802. -62782.  1.00   505.    1219.
##  2 alpha[1]  -9937.  -9832.  3.85e3 3.73e3 -16474.  -3522.  1.05    73.6    313.
##  3 alpha[2] -23616. -23612.  1.50e3 1.44e3 -26196. -21187.  1.02   191.     411.
##  4 alpha[3] -18438. -18458.  2.66e3 2.73e3 -22896. -14244.  1.02   150.     326.
##  5 alpha[4]   4801.   4765.  3.03e3 3.20e3   -218.   9780.  1.02   201.     346.
##  6 alpha[5]  -6095.  -6131.  1.59e3 1.55e3  -8628.  -3480.  1.02   126.     292.
##  7 alpha[6] -19032. -19017.  3.21e3 3.20e3 -24300. -13809.  1.02   137.     306.
##  8 alpha[7] -16051. -16082.  4.91e3 4.94e3 -24066.  -8070.  1.02   144.     315.
##  9 alpha[8]  -7640.  -7784.  2.94e3 3.02e3 -12287.  -2521.  1.03   134.     288.
## 10 alpha[9]  -1341.  -1226.  1.02e4 9.81e3 -18385.  15924.  1.01   282.     553.
## # ... with 1,595 more rows, and abbreviated variable names 1: ess_bulk,
## #   2: ess_tail
## # i Use `print(n = ...)` to see more rows
```

Pooled model

Summary of model fit for main parameters:

```
## Running MCMC with 4 sequential chains...
##
## Chain 1 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 1 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 1 finished in 29.4 seconds.
## Chain 2 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 2 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 2 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 2 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 2 finished in 33.8 seconds.
## Chain 3 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 3 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 3 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 3 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 3 finished in 30.7 seconds.
## Chain 4 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 4 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 4 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 4 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 4 finished in 30.9 seconds.
##
## All 4 chains finished successfully.
## Mean chain execution time: 31.2 seconds.
## Total execution time: 125.4 seconds.
```

```
## # A tibble: 1,541 x 10
##    variable        mean   median      sd     mad       q5      q95  rhat ess_b~1
```

```
##    <chr>          <dbl>    <dbl>   <dbl>   <dbl>    <dbl>    <dbl> <dbl>   <dbl>
##  1 lp__        -7.10e+3 -7.10e+3 2.01e+0 1.87e+0 -7.10e+3 -7.09e+3  1.00   1737.
##  2 alpha       -1.04e+4 -1.04e+4 2.71e+3 2.69e+3 -1.48e+4 -5.84e+3  1.00   1726.
##  3 beta_SAT_A~  9.14e+1  9.14e+1 5.31e+0 5.35e+0  8.28e+1  1.00e+2  1.00   1711.
##  4 beta_MD_FA~  4.05e-2  4.02e-2 1.71e-2 1.73e-2  1.26e-2  6.94e-2  1.00   2483.
##  5 beta_COSTT~  3.68e-1  3.67e-1 3.30e-2 3.21e-2  3.15e-1  4.23e-1  1.00   1708.
##  6 beta_POVER~ -2.38e+2 -2.39e+2 7.55e+1 7.56e+1 -3.59e+2 -1.15e+2  1.00   2001.
##  7 beta_URBAN   2.23e+3  2.23e+3 5.34e+2 5.23e+2  1.34e+3  3.13e+3  1.00   3705.
##  8 beta_PRIVA~ -7.42e+3 -7.40e+3 8.88e+2 8.80e+2 -8.88e+3 -5.98e+3  1.00   1734.
##  9 sigma        6.44e+3  6.44e+3 1.70e+2 1.69e+2  6.17e+3  6.74e+3  1.00   3494.
## 10 log_lik[1]  -1.02e+4 -1.02e+4 1.48e+2 1.47e+2 -1.05e+4 -9.98e+3  1.00   3930.
## # ... with 1,531 more rows, 1 more variable: ess_tail <dbl>, and abbreviated
## #   variable name 1: ess_bulk
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Hierarchical model

Summary of model fit for main parameters:

```
## Running MCMC with 4 sequential chains...
##
## Chain 1 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 1 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 1 finished in 94.5 seconds.
## Chain 2 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 2 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 2 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 2 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 2 finished in 92.4 seconds.
## Chain 3 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 3 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 3 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 3 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 3 finished in 84.1 seconds.
## Chain 4 Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 4 Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 4 Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 4 Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 4 finished in 94.0 seconds.
##
## All 4 chains finished successfully.
## Mean chain execution time: 91.3 seconds.
## Total execution time: 365.8 seconds.
```

```
## # A tibble: 1,613 x 10
##    variable    mean  median     sd    mad      q5     q95  rhat ess_bulk ess_t~1
##    <chr>      <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl> <dbl>    <dbl>   <dbl>
##  1 lp__      -7281.  -7282.   12.4   12.4  -7301.  -7258.  1.02     173.    136.
##  2 alpha[1]  -2777.  -2965. 5907.  5757.  -12438.   6972.  1.01    2471.   2412.
##  3 alpha[2] -18072. -17848. 4242.  4226.  -25284. -11254.  1.00    1423.   2375.
##  4 alpha[3] -12979. -12935. 5290.  5086.  -21784.  -4511.  1.00    2172.   1422.
##  5 alpha[4]  -2644.  -2742. 5827.  5711.  -11873.   7033.  1.00    2485.   2461.
##  6 alpha[5]  -7961.  -8187. 4294.  4166.  -14964.   -811.  1.00    1845.   2490.
```

```
##  7 alpha[6]  -9686.  -9568. 5616.  5728.  -19238.  -1035.  1.01     568.   1920.
##  8 alpha[7]  -3154.  -2766. 7199.  6892.  -15620.   8276.  1.00    1699.   2128.
##  9 alpha[8]  -9118.  -9143. 5474.  5488.  -17776    -186.  1.01     483.    883.
## 10 alpha[9]  -3445.  -3754. 7281.  6879.  -15121.   8559.  1.00    3255.   2726.
## # ... with 1,603 more rows, and abbreviated variable name 1: ess_tail
## # i Use `print(n = ...)` to see more rows
```

6. Stan Specifications

We used the following Stan options:

- Interface: cmdstanr
- Seed: 1234
- Chains: 4 (default)
- Iterations per chain: 2000 (default)

7. Convergence Diagnostics

In this section, we are performing the diagnostics for the separate and hierarchical model based on the first region (N = 37). This region was selected at random, and it is not the region with the lowest or highest number of samples. Although, considering separately all regions would have given a better overall picture, for simplicity reasons, one region is selected. For the project group, this seemed to be more interesting way to do analysis than blending all the regions together with for example averages.

Rhat

We used the rhat() function for all models to get the rhat convergence diagnostic. The function compares the between chain and within chain estimates for model parameters and other univariate quantities of interest. If the between chain and within chain estimates agree, it is said that the chains have mixed well. This happens when R-hat is less than 1.05 or less than 1.01 depending on the standard. We decided not be strict with our multivariate model and chose the less strict threshold of 1.05.

Rhat statistics revealed that the separate model converged worse than the hierarchical or pooled model, both of which got good rhat convergence values for every beta in every chain. The fact that the separate model has problems with convergence while the pooled model and the hierarchical model converge very well speaks clearly against the separate model while suggesting the other two models perform well on this front.

Separate model Rhat

```
## # A tibble: 7 x 5
##   Parameter            `Chain 1` `Chain 2` `Chain 3` `Chain 4`
##   <chr>                    <dbl>     <dbl>     <dbl>     <dbl>
## 1 alpha[1]                  1.01      1.04      1.00      1.01
## 2 beta_SAT_ALL[1]           1.10      1.00      1.00      1.03
## 3 beta_MD_FAMINIC[1]        1.05      1.06      1.01      1.00
## 4 beta_COSTT4_A[1]          1.08      1.00      1.00      1.05
## 5 beta_POVERTY_RATE[1]      1.00      1.06      1.00      1.00
## 6 beta_URBAN[1]             1.00      1.01      1.01      1.00
## 7 beta_PRIVATE[1]           1.09      1.00      1.00      1.04
```

Pooled model Rhat

```
## # A tibble: 8 x 5
```

```
##   Parameter          `Chain 1` `Chain 2` `Chain 3` `Chain 4`
##   <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
## 1 alpha                  0.999      1.01     1.00      1.00
## 2 beta_SAT_ALL           0.999      1.01     1.00      1.00
## 3 beta_MD_FAMINIC        1.00       1.00     0.999     1.00
## 4 beta_COSTT4_A          1.01       1.01     1.00      1.00
## 5 beta_POVERTY_RATE      1.00       1.00     1.00      0.999
## 6 beta_URBAN             1.00       1.00     1.00      0.999
## 7 beta_PRIVATE           1.00       1.01     1.00      1.00
## 8 sigma                  0.999      1.00     1.00      1.01
```

Hierarchical model Rhat

```
## # A tibble: 7 x 5
##   Parameter             `Chain 1` `Chain 2` `Chain 3` `Chain 4`
##   <chr>                     <dbl>     <dbl>     <dbl>     <dbl>
## 1 alpha[1]                   1.00      1.00      0.999      1.03
## 2 beta_SAT_ALL[1]            1.00      1.00      1.00       1.02
## 3 beta_MD_FAMINIC[1]         1.00      1.00      1.01       1.02
## 4 beta_COSTT4_A[1]           1.02      1.00      1.00       1.00
## 5 beta_POVERTY_RATE[1]       1.00      0.999     1.00       1.04
## 6 beta_URBAN[1]              1.00      1.00      1.00       1.01
## 7 beta_PRIVATE[1]            1.00      1.00      1.00       1.01
```

The codes for the plots are provided in the appendix.

Effective sample size (ESS)

Effective sample size (ESS) evaluates the uncertainty in estimates that is caused by autocorrelation of the chains. The value of effective sample size represents the number of independent samples that poses the same predictive power as all the autocorrelated samples. In other words, if the number of effective sample size is high, then the number of independent samples is high. (Stan reference manual, n.d)

For the separate model the ESS values range from 54 to 96 (N = 37). For the pooled model the ESS values range from 408 to 996 (N = 766). For the hierarchical model the ESS values range from 340 to 1225 (N = 766).

For pooled and hierarchical models, for many but not all parameters there are auto correlated samples that are reduced from the effective sample size. There are also in some cases a lot of independent samples when ESS is close to N. However, for all the three models, especially for the separate modelthere are ESS values that are greater than N.

According to Stan reference manual, ESS can be higher than the number of samples due to anthitehetic Markov Chains which have negative auto correlations on odd lags. Based on the same source, another reason for this could be the NUTS algorithm used in Stan, which might give ESS that is higher than the sample size for parameters that have close to Gaussian posterior and low dependency on the rest of the parameters.

Separate Model ESS

```
## # A tibble: 7 x 2
##   Parameter              ESS
##   <chr>                <dbl>
## 1 alpha[1]              40.7
## 2 beta_COSTT4_A[1]      46.5
## 3 beta_MD_FAMINIC[1]    65.0
## 4 beta_POVERTY_RATE[1]  74.1
## 5 beta_PRIVATE[1]       52.8
```

```
## 6 beta_SAT_ALL[1]       48.9
## 7 beta_URBAN[1]         84.0
```

Pooled Model ESS

```
## # A tibble: 8 x 2
##   Parameter           ESS
##   <chr>             <dbl>
## 1 alpha              411.
## 2 beta_COSTT4_A      417.
## 3 beta_MD_FAMINIC    640.
## 4 beta_POVERTY_RATE  535.
## 5 beta_PRIVATE       429.
## 6 beta_SAT_ALL       412.
## 7 beta_URBAN         939.
## 8 sigma              850.
```

Hierarchical Model ESS

```
## # A tibble: 7 x 2
##   Parameter              ESS
##   <chr>                <dbl>
## 1 alpha[1]              636.
## 2 beta_COSTT4_A[1]      266.
## 3 beta_MD_FAMINIC[1]    493.
## 4 beta_POVERTY_RATE[1]  710.
## 5 beta_PRIVATE[1]       407.
## 6 beta_SAT_ALL[1]       482.
## 7 beta_URBAN[1]        1091.
```

HMC specific convergence diagnostics for all models (divergences, tree depth)

HMC specific convergence diagnostics were analysed for all models. The R code and output for diagnostics can be found from the appendix. Based on the HMC specific convergence diagnostics, the separate model reached the initial maximum tree depth of 10 in 100% of the transitions. The diagnostics state that this leads to premature termination of trajectories and slow exploration. The proposed action of increasing the limit was tested, but it resulted to too long running time for the already time consuming fitting of the separate model. The rest of the diagnostics consisting of divergences, E-BMFI, effective sample size and split R-hat satisfactory.

For the pooled model, tree depth, divergences, E-BFMI, effective sample size and split R-hat were satisfactory.

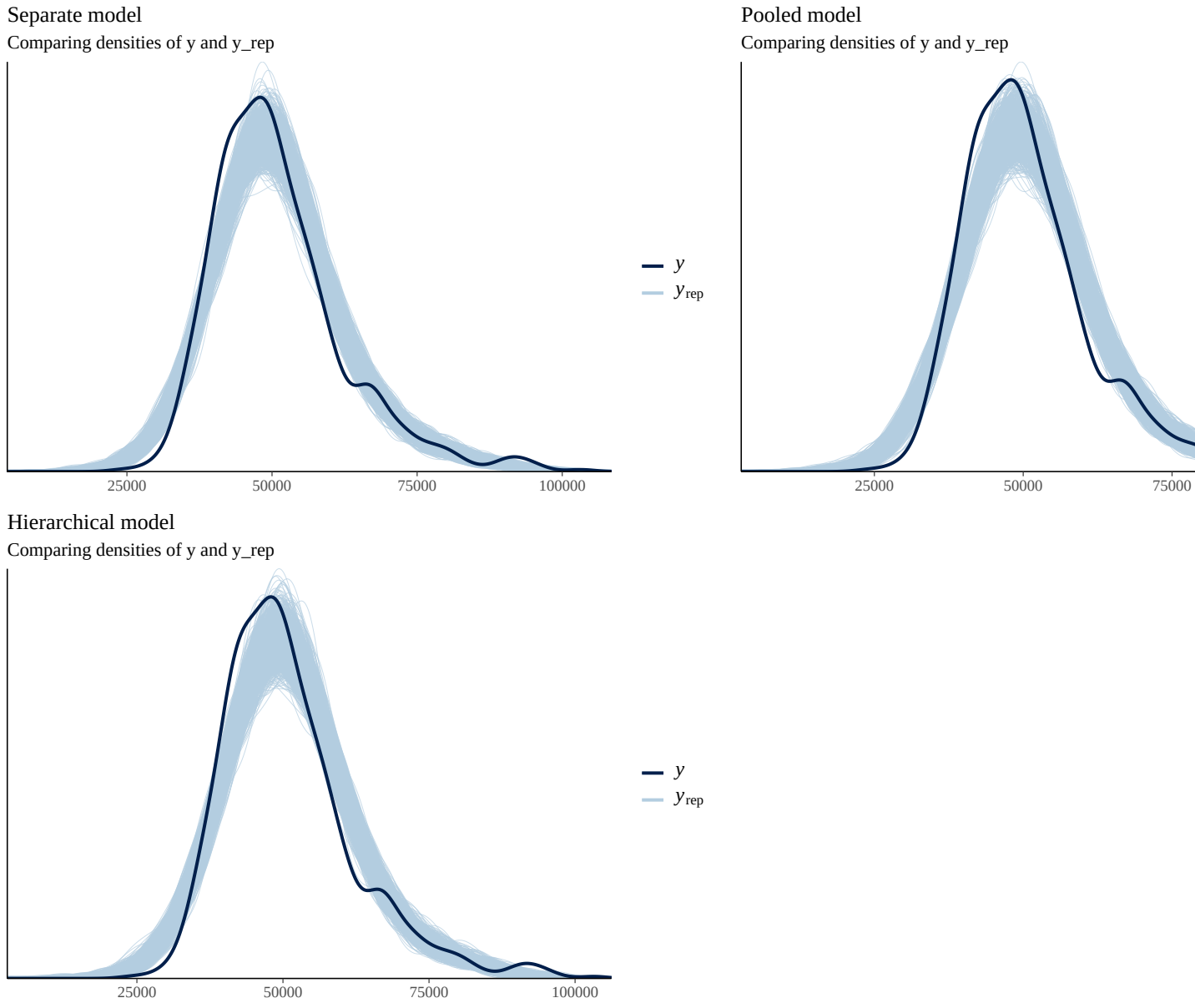
For the hierarchical model, only 4 out of 4000 hit the maximum tree depth and 23 out of 4000 transitions did not converge. Due to relatively low numbers (0.1% and 0.57%) no actions were taken for the hierarchical model. The rest of the diagnostics were satisfactory.

8. Posterior Predictive Checks and Model Comparison

Posterior Predictive Checks

When doing posterior predictive checking we look for systematic discrepancies between the real observations and the data we get from simulating replicated data under the fitted model (Gelman and Hill, 2006). Posterior predictive checking is a form of internal validation that is a helpful phase of model building and checking in assessing whether our fitted model makes sense.

When comparing the densities of y (real sample values) and y_rep (Stan normal_rng generated values) in the graphs below, it looks like that y and y_rep are aligning quite well, with all three models, although it seems that the y_rep values are on average somewhat more on right (higher estimates for earnings). There are also intervals where y is suddenly changing direction, which are not visible in y_rep values.



Model comparison (LOO, K hat)

LOO elpd_loo is an estimate of the expected log pointwise pridictive dentisty (ELPD). elpd_loo sums individual pointwise log predictive densities (Stan Reference Manual, n.d).

Based on elpd_loo the model with the highest value should be selected (highest ELPD). From the tables below we can observe that separate model has the highest elpd_loo value, hierarchical model the second highest value, and the pooled the lowest value. For that reason, separate model is suggested by elpd_loo although the value of hierachical model comes quite close.

```
##        elpd_diff  se_diff
```

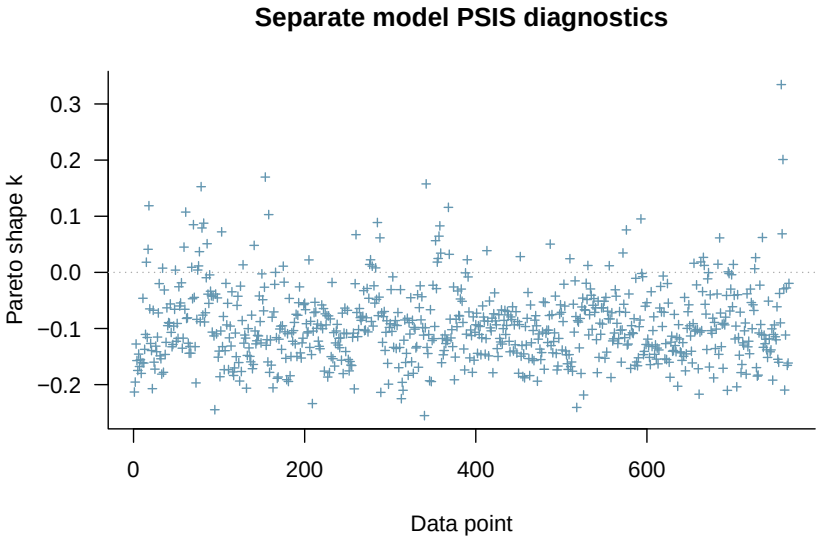
```
## model1        0.0        0.0
## model3      -53.7        7.2
## model2 -7504698.6    81720.7
```

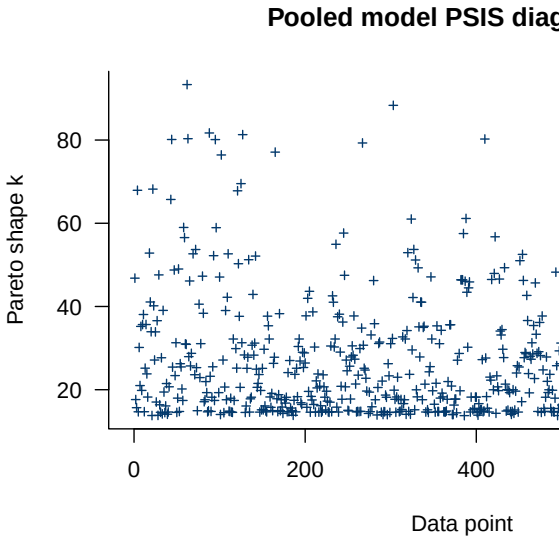
K hat

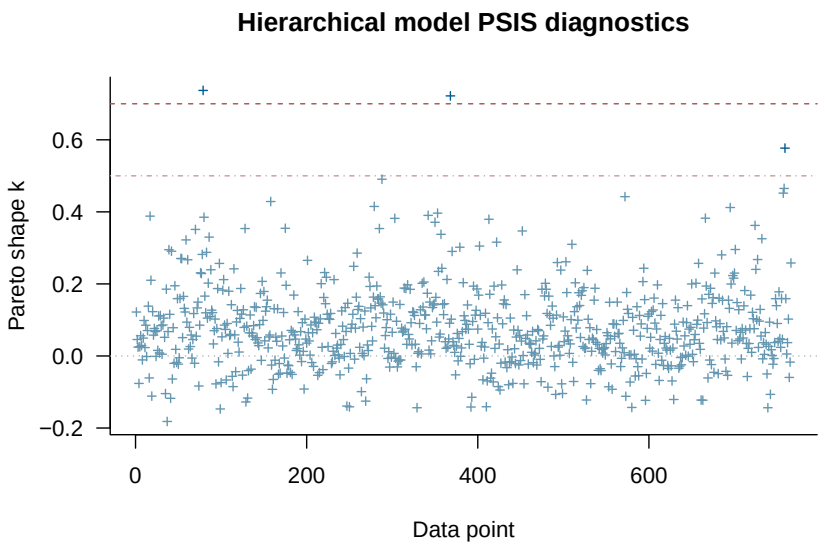
In general, if $\hat{k}$ values are greater than 0.7 the PSIS-LOO estimates might be too optimistic (biased), and on the contrary, if the $\hat{k}$ values are smaller than 0.7 the estimated can be considered reliable. (Gabry, Gelman, Vehtari, 2016)

As can be observed from the graphs "Separate mode PSIS diagnostics", "Pooled model PSIS diagnostics" and "Hierarchical model PSIS diagnostics" for separate model all $\hat{k}$ values are smaller than 0.7, for hierarchical only one value $\hat{k}$ is greater than 0.7. This means that PSIS-LOO estimates for these models can be considered reliable. However, for pooled model there are many $\hat{k}$ values that are a lot greater than 0.7. This indicates that this model may be biased and the estimates might not be reliable.

The more precise diagnostics for $\hat{k}$ are available in appendix.







9. Predictive Performance Assessment

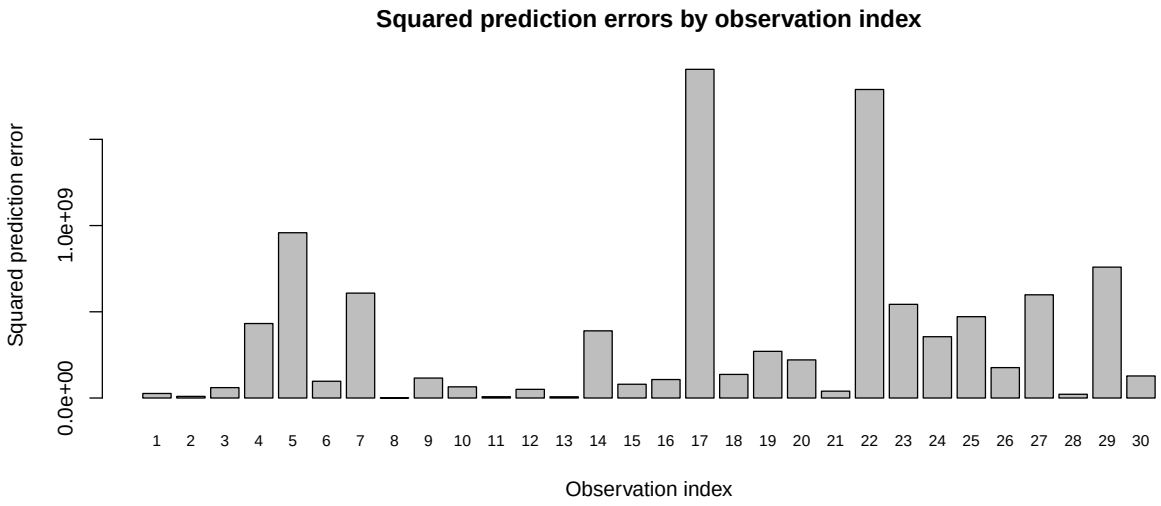
Root Mean Squared Error (RMSE)

We use RMSE for assessing the predictive ability of our model. The assessment was conducted as follows: we used the median of the posterior distribution for each parameter in our model fitted with the training set to predict the dependent variable based on the values of the independent variables in the test set. Lastly, after obtaining the predicted values, we calculated the RMSE with respect to the actual observed values of the dependent variable in the test set. The complete test set and the RMSE generating code can be found in the Appendix.

Due to limited space, we assess only the pooled model. The pooled model is the simplest model and provides a straightforward way to get a general feel of the predictive ability of our linear model.

Pooled model

The plot below shows the computed squared prediction errors for all 30 observations in the test set.



The resulting RMSE is 12805.23, which means that on average, our prediction for the median earnings 10 years post entry of the alumni from 30 universities in the test set was off by $12.8k. To get some feel for how good the accuracy is, we can look at the standard deviation of the dependent variable in the training set, which is $11.5k. This means that we were on average about one standard deviation off in our predictions. We consider this accuracy satisfactory. After all, our model is quite simple and there are several other variables, which affect earnings and which we were unable to include in the model.

10. Prior Sensitivity Analysis

To test whether our results are sensitive to prior choices, and to see how changing the priors affects our results, we ran all the models with both wider and narrower priors. The wide priors were obtained by multiplying each standard deviation by three, and the narrow priors were obtained by dividing the original priors by two.

For the pooled model, when scaling sigma values with 3, there were only minor changes in beta values. Also, when scaling sigma values with 0.5 the results stayed about the same except for beta_PRIVATE whose mean increased from around -7000 to around -5000. Based on these results it seems that pooled model is not that sensitive for prior changes that are this magnitude.

In conclusion, separate and hierarchical models are more sensitive for changes in priors compared to pooled model. This robustness could be explained by the fact that in pooled model the regions are considered as one and for that reason, there are more observations for that one entity compared to individual regions.

As the codes of the sensitivity analysis are long, they are included in the appendix.

Separate model sensitivity analysis

11. Discussion of Issues and Potential Improvements.

During the project, the large amount of variables also posed challenges, as it was arguably rather slow and burdensome to find the most relevant variables to use in our analysis. It was somewhat surprising how fast the observation count started to shrink in data cleaning process, so in hindsight more attention could have been paid to cleanliness, as this dataset had for example a lot of missing values.

It was a shame we couldn't use a larger testing dataset because all data used for testing would be away from training the model. We tried to reduce the need to have a very long dataset by working very hard on removing independent variables. That work also ended up expanding our observation count from roughly 200 to over 700 observations, as we found variables with less null values and removed variables with too

much missing data. For the feature selection, with more time we could have experimented with alternative methods, like the LASSO or Ridge regression.

As we were building a predictive model, we had to take certain ethical questions into account, especially to avoid building a discriminatory model. For example, if our model was used to by an employer to assess how well alumni from a certain school would perform in their career, it would be discriminatory to have a model that predicts lower performance for a university with say, high female or black student population as the real reason for differences in our data might be something else entirely than skin color or gender. In the end, this wasn't a big issue as for example the correlations between both gender vs income and share of white population vs income were very low and they would have been dropped from our model no matter what.

Another clear improvement would be to do more extensive predictive assessment of the models. We only assessed the pooled model, so both the separate model and hierarchical model could be assessed if we had more time and space. Furthermore, we only computed a simple RMSE metric. Perhaps it would be interesting to look at the probabilistic certainty of the predictions by plotting confidence regions, for example. The test set could also have been a bit bigger, as it is under 5% of the total number of available observations.

12. Conclusions

Like many previous studies, we found a strong link between education and income after school. The main aim of our project was to find a way to predict average income after studying at a university for any non-specialized university in the United States. We built three different multivariate models – separate, pooled and hierarchical – to predict income after school. All the models used the same set of numerical and categorical university-level independent variables.

As our initial dataset had almost 3000 features, feature selection was a very important and influential part of our project. We conducted feature selection in three phases: 1) We selected a subset of features based on insights from previous studies and also used common sense to come up with additional features that could predict future earnings. 2) We assessed the relationship of the features with our dependent variable and accounted for any multicollinearity arising from mutually correlated features. We also visualized categorical variables and engineered new features to be added to the model. 3) We utilized stepwise regression in streamlining our model to include only the most significant features.

All our priors strived to be weakly informative but not too vague. We experimented with three different prior specifications: our conservative estimate for a good set of priors as well as two sets of priors set to be either wider or narrower than the conservative estimate. Separate and hierarchical models are more sensitive to changes in priors compared to pooled model. This robustness could be explained by the fact that in pooled model the regions are considered as one and for that reason, there are more observations for that one entity compared to individual regions.

We diagnosed convergence by calculating rhat, effective sample size and HMC specific convergence diagnostics for all models. None of the models converged perfectly according to all statistics.

Rhat statistics revealed that the separate model converged worse than the hierarchical or pooled model, both of which got good rhat convergence values for every beta in every chain. ESS diagnostics revealed that both the pooled model and the hierarchical model had autocorrelation issues. For the HMC specific convergence diagnostics, the separate model failed to produce results in reasonable running time. The separate model got satisfactory results from divergence diagnostics, E-BMFI, ESS and split rhat. The pooled model was overall satisfactory regarding these tests and the hierarchical model was also satisfactory save for a few hiccups when trying to obtain the HMC specific convergence diagnostics.

We ran posterior predictive checks to compare our data to simulated results. The visual comparison did not give any reason for concern and our models performed well on these checks. Additionally, we used LOO to perform model comparison. LOO suggested the separate model followed closely by the hierarchical model.

We got good k hat values for the separate model and the hierarchical model, which means the PSIS-LOO estimates for these models can be considered reliable. However, for pooled model there were many k hat

values that were much greater than 0.7 . This indicates that the pooled model may be biased and the estimates might not be reliable.

As we have been building a predictive model, we though it would be important to perform predictive performance assessment. We use RMSE to assess the predictive ability of our model. Our data had been initially split into a training set and a test set. Due to limited space, we assessed only the pooled model. We consider our RMSE results for model accuracy to be satisfactory. Of course, our model is quite simple and there are several other variables which affect earnings and which we have been unable to include in the model.

13. Self-reflection

In hindsight taking on a task of creating three multivariate models was very ambitious, because multivariable models had not been covered very much during the course. That ambition paid off, however, and we think we have now a much stronger grasp of how the different models - pooled, separate and hierarchical - work and also know how to build multivariate bayesian linear models in general. We also learned a lot about feature selection, because the 3000 variables in the dataset forced us to create sensible procedures for selection. The combination of using logic and feature engireering tools like stepwise regression is a very powerful skill we developed while making this project.

I think this project made us think really hard about what do we want to show the reader to make our work as understandable as possible and we learned to visualize data and results in interesting ways and apply the visualization techniques covered on the course. Making the visualizations also aided our own thinking and certain visualizations gave valuable insight on our results and their quality.

Working with Stan is something none of us had done before this course. Stan is clearly a very powerful tool for statistical and data science use and we feel this project really ingrained basic Stan workflows and made working with Stan feel more of an efficient routine than an obstacle.

14. References

Card, D. (1999). THE CAUSAL EFFECT OF EDUCATION ON EARNINGS. Wolla, S. A., & Sullivan, J. (2017). Education, Income, and Wealth. https://fred.stlouisfed.org/graph/?g=7yKu.

Brewer, Dominic J., Eric R. Eide, and Ronald G. Ehrenberg. "Does it pay to attend an elite private college? Cross cohort evidence on the effects of college quality on earnings." (1996).

Card, David, and Alan B. Krueger. "Does school quality matter? Returns to education and the characteristics of public schools in the United States." Journal of political Economy 100.1 (1992): 1-40.

Data Commons. (2020). Gross domestic product per capita in United States of America [Graph]. Retrieved from https://datacommons.org/place/country/USA?utm_medium=explore&mprop=income&popt=Person&cpv=age%2CYears15Onwards&hl=en

Gelman, Andrew, and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge university press, 2006.

Gabry, Gelman, Vehtari. (2016) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.

Number2. (2020). Average SAT Score [Blog Post]. Retrieved from: https://www.number2.com/average-sat-score/

Stan. (n.d) LOO package glossary [Website]. Retrieved from: ttps://mc-stan.org/loo/reference/loo-glossary.html

Stan. (n.d) Effective Sample Size [Website]. Retrieved from: https://mc-stan.org/docs/2_19/reference-manual/effective-sample-size-section.html

15. Appendix

Stepwise regression

```
# baseline model
model <- lm(MD_EARN_WNE_P10 ~ ., data = data.joined.model)

# step wise regression implied "best" model in terms of AIC
step(model, direction = "backward")

# step wise second iteration without MASTER and DOCTORAL

data.stepwise <- select(data.joined.model, -c(MASTER, DOCTORAL))

model <- lm(MD_EARN_WNE_P10 ~ ., data = data.stepwise)

# step wise regression implied "best" model in terms of AIC
step(model, direction = "backward")
```

Separate model Stan code (separate.stan)

```
print(separate.model)
```

```
## data {
##
##   // number of observations
##   int<lower=0> N;
##
##   // number of regions
##   int<lower=0> K;
##
##   // region indicators
##   int<lower=1, upper=K> x[N];
##
##   // data vectors
##   vector[N] SAT_ALL; // composite SAT score
##   vector[N] MD_FAMINIC; // median family income
##   vector[N] COSTT4_A; // cost of education
##   vector[N] POVERTY_RATE; // povery rate
##   vector[N] URBAN; // urban dummy
##   vector[N] PRIVATE; // private institution dummy
##   vector[N] y; // dependent variable (median earnings 10 years post-graduation)
##
##   // prior distribution parameters
##   real pm_alpha; // prior mean of intercept term
##   real ps_alpha; // prior sd of intercept term
##   real pm_SAT_ALL;
##   real ps_SAT_ALL;
##   real pm_MD_FAMINC;
##   real ps_MD_FAMINC;
##   real pm_COSTT4_A;
##   real ps_COSTT4_A;
##   real pm_POVERTY_RATE;
##   real ps_POVERTY_RATE;
```

```
##   real pm_URBAN;
##   real ps_URBAN;
##   real pm_PRIVATE;
##   real ps_PRIVATE;
##   real pm_sigma;
##   real ps_sigma;
##
## }
##
## parameters {
##
##   // regression model parameters
##   vector[K] alpha;
##   vector[K] beta_SAT_ALL;
##   vector[K] beta_MD_FAMINIC;
##   vector[K] beta_COSTT4_A;
##   vector[K] beta_POVERTY_RATE;
##   vector[K] beta_URBAN;
##   vector[K] beta_PRIVATE;
##   vector<lower=0>[K] sigma;
##
## }
##
## model {
##
##   // priors
##   for (j in 1:K) {
##     alpha[j] ~ normal(pm_alpha, ps_alpha);
##     beta_SAT_ALL[j] ~ normal(pm_SAT_ALL, ps_SAT_ALL);
##     beta_MD_FAMINIC[j] ~ normal(pm_MD_FAMINC, ps_MD_FAMINC);
##     beta_COSTT4_A[j] ~ normal(pm_COSTT4_A, ps_COSTT4_A);
##     beta_POVERTY_RATE[j] ~ normal(pm_POVERTY_RATE, ps_POVERTY_RATE);
##     beta_URBAN[j] ~ normal(pm_URBAN, ps_URBAN);
##     beta_PRIVATE[j] ~ normal(pm_PRIVATE, ps_PRIVATE);
##     sigma[j] ~ normal(pm_sigma, ps_sigma);
##   }
##
##
##   // likelihoods
##   for (i in 1:N) {
##     y[i] ~ normal(alpha[x[i]] + beta_SAT_ALL[x[i]] * SAT_ALL[i] + beta_MD_FAMINIC[x[i]] *
##     MD_FAMINIC[i] + beta_COSTT4_A[x[i]] * COSTT4_A[i] +
##     beta_POVERTY_RATE[x[i]] * POVERTY_RATE[i] +
##     beta_URBAN[x[i]] * URBAN[i] + beta_PRIVATE[x[i]] * PRIVATE[i], sigma);
##   }
##
## }
##
## generated quantities {
##
##   // log-likelihoods
##   vector[N] log_lik;
##   vector[N] y_rep;
##
```

```
##   for (i in 1:N) {
##     log_lik[i] = normal_lpdf(y[i] | alpha[x[i]] + beta_SAT_ALL[x[i]]
##     * SAT_ALL[i] + beta_MD_FAMINIC[x[i]] * MD_FAMINIC[i] + beta_COSTT4_A[x[i]] * COSTT4_A[i] + beta_
##     * POVERTY_RATE[i] + beta_URBAN[x[i]] * URBAN[i] +
##     beta_PRIVATE[x[i]] * PRIVATE[i], sigma[x[i]]);
##
##     y_rep[i] = normal_rng(alpha[x[i]] + beta_SAT_ALL[x[i]]
##     * SAT_ALL[i] + beta_MD_FAMINIC[x[i]] * MD_FAMINIC[i] + beta_COSTT4_A[x[i]] * COSTT4_A[i] + beta_
##     * POVERTY_RATE[i] + beta_URBAN[x[i]] * URBAN[i] +
##     beta_PRIVATE[x[i]] * PRIVATE[i], sigma[x[i]]);
##   }
##
## }
```

Pooled model Stan code (pooled.stan)

```
print(pooled.model)
```

```
## data {
##
##   int<lower=0> N; // number of observations
##
##   // data vectors
##   vector[N] SAT_ALL;
##   vector[N] MD_FAMINIC;
##   vector[N] COSTT4_A;
##   vector[N] POVERTY_RATE;
##   vector[N] URBAN;
##   vector[N] PRIVATE;
##   vector[N] y;
##
##   // prior distribution parameters
##   real pm_alpha; // prior mean of intercept term
##   real ps_alpha; // prior sd of intercept term
##   real pm_SAT_ALL;
##   real ps_SAT_ALL;
##   real pm_MD_FAMINC;
##   real ps_MD_FAMINC;
##   real pm_COSTT4_A;
##   real ps_COSTT4_A;
##   real pm_POVERTY_RATE;
##   real ps_POVERTY_RATE;
##   real pm_URBAN;
##   real ps_URBAN;
##   real pm_PRIVATE;
##   real ps_PRIVATE;
##   real pm_sigma;
##   real ps_sigma;
##
## }
##
## parameters {
##   real alpha;
##   real beta_SAT_ALL;
```

```
##   real beta_MD_FAMINIC;
##   real beta_COSTT4_A;
##   real beta_POVERTY_RATE;
##   real beta_URBAN;
##   real beta_PRIVATE;
##   real<lower=0> sigma;
##
## }
##
## model {
##
##   // weakly informative priors
##   alpha ~ normal(pm_alpha, ps_alpha);
##   beta_SAT_ALL ~ normal(pm_SAT_ALL, ps_SAT_ALL);
##   beta_MD_FAMINIC ~ normal(pm_MD_FAMINC, ps_MD_FAMINC);
##   beta_COSTT4_A ~ normal(pm_COSTT4_A, ps_COSTT4_A);
##   beta_POVERTY_RATE ~ normal(pm_POVERTY_RATE, ps_POVERTY_RATE);
##   beta_URBAN ~ normal(pm_URBAN, ps_URBAN);
##   beta_PRIVATE ~ normal(pm_PRIVATE, ps_PRIVATE);
##   sigma ~ normal(pm_sigma, ps_sigma);
##
##   y ~ normal(alpha + beta_SAT_ALL * SAT_ALL + beta_MD_FAMINIC * MD_FAMINIC +
##    beta_COSTT4_A * COSTT4_A +
##   beta_POVERTY_RATE * POVERTY_RATE + beta_URBAN * URBAN +
##   beta_PRIVATE * PRIVATE, sigma);
## }
##
## generated quantities {
##   vector[N] log_lik;
##   vector[N] y_rep;
##
##   for (i in 1:N) {
##     log_lik[i] = normal_lpdf(y[i] | alpha + beta_SAT_ALL * SAT_ALL
##     + beta_MD_FAMINIC * MD_FAMINIC +
##     beta_COSTT4_A * COSTT4_A + beta_POVERTY_RATE * POVERTY_RATE +
##     beta_URBAN * URBAN + beta_PRIVATE * PRIVATE, sigma);
##
##     y_rep[i] = normal_rng(alpha + beta_SAT_ALL * SAT_ALL[i]
##     + beta_MD_FAMINIC * MD_FAMINIC[i] +
##     beta_COSTT4_A * COSTT4_A[i] + beta_POVERTY_RATE * POVERTY_RATE[i]
##     + beta_URBAN * URBAN[i] + beta_PRIVATE * PRIVATE[i], sigma);
##   }
##
## }
```

Hierarchical model Stan code (hierarchical.stan)

```
print(hierarchical.model)
```

```
## data {
##
##   int<lower=0> N; // number of observations
##   int<lower=0> K; // number of regions
##   int<lower=1, upper=K> x[N]; // discrete group indicators
```

```
##
##   // data vectors
##   vector[N] SAT_ALL; // SAT score
##   vector[N] MD_FAMINIC; // medain family income
##   vector[N] COSTT4_A; // average cost of education
##   vector[N] POVERTY_RATE; // poverty rate in school area
##   vector[N] URBAN; // urban dummy
##   vector[N] PRIVATE; // private innstitution dummy
##   vector[N] y; // median earnings
##
##   // params for prior distributions
##   real pm_alpha; // prior mean of intercept term
##   real ps_alpha; // prior sd of intercept term
##   real pm_s_alpha; // sd of prior mean of intercept term
##   real ps_s_alpha; // sd of prior sd of intercept term
##
##   real pm_SAT_ALL;
##   real ps_SAT_ALL;
##   real pm_s_SAT_ALL;
##   real ps_s_SAT_ALL;
##
##   real pm_MD_FAMINC;
##   real ps_MD_FAMINC;
##   real pm_s_MD_FAMINC;
##   real ps_s_MD_FAMINC;
##
##   real pm_COSTT4_A;
##   real ps_COSTT4_A;
##   real pm_s_COSTT4_A;
##   real ps_s_COSTT4_A;
##
##   real pm_POVERTY_RATE;
##   real ps_POVERTY_RATE;
##   real pm_s_POVERTY_RATE;
##   real ps_s_POVERTY_RATE;
##
##   real pm_URBAN;
##   real ps_URBAN;
##   real pm_s_URBAN;
##   real ps_s_URBAN;
##
##   real pm_PRIVATE;
##   real ps_PRIVATE;
##   real pm_s_PRIVATE;
##   real ps_s_PRIVATE;
##
##   real pm_sigma;
##   real ps_sigma;
##   real pm_s_sigma;
##   real ps_s_sigma;
##
## }
##
## parameters {
```

```
##
##   vector[K] alpha;
##   vector[K] beta_SAT_ALL;
##   vector[K] beta_MD_FAMINIC;
##   vector[K] beta_COSTT4_A;
##   vector[K] beta_POVERTY_RATE;
##   vector[K] beta_URBAN;
##   vector[K] beta_PRIVATE;
##   real<lower=0> sigma; // common standard deviation
##
##   real mu_alpha;
##   real<lower=0> sigma_alpha;
##
##   real mu_SAT_ALL;
##   real<lower=0> sigma_SAT_ALL;
##
##   real mu_MD_FAMINIC;
##   real<lower=0> sigma_MD_FAMINIC;
##
##   real mu_COSTT4_A;
##   real<lower=0> sigma_COSTT4_A;
##
##   real mu_POVERTY_RATE;
##   real<lower=0> sigma_POVERTY_RATE;
##
##   real mu_URBAN;
##   real<lower=0> sigma_URBAN;
##
##   real mu_PRIVATE;
##   real<lower=0> sigma_PRIVATE;
##
##   real mu_sigma;
##   real<lower=0> sigma_sigma;
##
## }
##
## model {
##
##   mu_alpha ~ normal(pm_alpha, pm_s_alpha);
##   sigma_alpha ~ normal(ps_alpha, ps_s_alpha);
##   alpha ~ normal(mu_alpha, sigma_alpha);
##
##   mu_SAT_ALL ~  normal(pm_SAT_ALL, pm_s_SAT_ALL);
##   sigma_SAT_ALL ~ normal(ps_SAT_ALL, ps_s_SAT_ALL);
##   beta_SAT_ALL ~ normal(mu_SAT_ALL, sigma_SAT_ALL);
##
##   mu_MD_FAMINIC ~  normal(pm_MD_FAMINC, pm_s_MD_FAMINC);
##   sigma_MD_FAMINIC ~ normal(ps_MD_FAMINC, ps_s_MD_FAMINC);
##   beta_MD_FAMINIC ~ normal(mu_MD_FAMINIC, sigma_MD_FAMINIC);
##
##   mu_COSTT4_A ~  normal(pm_COSTT4_A, pm_s_COSTT4_A);
##   sigma_COSTT4_A ~ normal(ps_COSTT4_A, ps_s_COSTT4_A);
##   beta_COSTT4_A ~  normal(mu_COSTT4_A, sigma_COSTT4_A);
##
```

```
##   mu_POVERTY_RATE ~  normal(pm_POVERTY_RATE, pm_s_POVERTY_RATE);
##   sigma_POVERTY_RATE ~ normal(ps_POVERTY_RATE, ps_s_POVERTY_RATE);
##   beta_POVERTY_RATE ~ normal(mu_POVERTY_RATE, sigma_POVERTY_RATE);
##
##   mu_URBAN ~  normal(pm_URBAN, pm_s_URBAN);
##   sigma_URBAN ~ normal(ps_URBAN, ps_s_URBAN);
##   beta_URBAN ~ normal(mu_URBAN, sigma_URBAN);
##
##   mu_PRIVATE ~  normal(pm_PRIVATE, pm_s_PRIVATE);
##   sigma_PRIVATE ~ normal(ps_PRIVATE, ps_s_PRIVATE);
##   beta_PRIVATE ~ normal(mu_PRIVATE, sigma_PRIVATE);
##
##   mu_sigma ~ normal(pm_sigma, pm_s_sigma);
##   sigma_sigma ~ normal(ps_sigma, ps_s_sigma);
##   sigma ~ normal(mu_sigma, sigma_sigma);
##
##
##   // likelihoods
##   for (i in 1:N) {
##     y[i] ~ normal(alpha[x[i]] + beta_SAT_ALL[x[i]] * SAT_ALL[i] + beta_MD_FAMINIC[x[i]] * MD_FAMINIC
##     + beta_POVERTY_RATE[x[i]] * POVERTY_RATE[i] + beta_URBAN[x[i]] * URBAN[i] + beta_PRIVATE[x[i]] *
##   }
##
## }
##
## generated quantities {
##
##   vector[N] log_lik;
##   vector[N] y_rep;
##
##   for (i in 1:N) {
##
##     log_lik[i] = normal_lpdf(y[i] | alpha[x[i]] + beta_SAT_ALL[x[i]] *
##     SAT_ALL[i] + beta_MD_FAMINIC[x[i]] * MD_FAMINIC[i] + beta_COSTT4_A[x[i]] * COSTT4_A[i] + beta_PO
##     POVERTY_RATE[i] + beta_URBAN[x[i]] * URBAN[i] +
##     beta_PRIVATE[x[i]] * PRIVATE[i], sigma);
##
##     y_rep[i] = normal_rng(alpha[x[i]] + beta_SAT_ALL[x[i]] * SAT_ALL[i]
##     + beta_MD_FAMINIC[x[i]] * MD_FAMINIC[i]  + beta_COSTT4_A[x[i]] * COSTT4_A[i] + beta_POVERTY_RATE
##     POVERTY_RATE[i] + beta_URBAN[x[i]] * URBAN[i] +
##     beta_PRIVATE[x[i]] * PRIVATE[i], sigma);
##  }
##
##
## }
```

Model fitting

Separate model fit

Summary of model fit for main parameters:

```
separate.model <- cmdstan_model(stan_file = "./Stan/separate.stan")

separate.model.data <- list(N = nrow(data.joined.stan),
```

```
                            K = length(unique(data.joined.stan$REGION)),
                            x = data.joined.stan$REGION,

                            SAT_ALL = data.joined.stan$SAT_ALL,
                            MD_FAMINIC = data.joined.stan$MD_FAMINC,
                            COSTT4_A = data.joined.stan$COSTT4_A,
                            POVERTY_RATE = data.joined.stan$POVERTY_RATE,
                            URBAN = data.joined.stan$URBAN,
                            PRIVATE = data.joined.stan$PRIVATE,
                            y = data.joined.stan$MD_EARN_WNE_P10,

                            pm_alpha = 0,
                            ps_alpha = 10000,
                            pm_SAT_ALL = 50,
                            ps_SAT_ALL = 500,
                            pm_MD_FAMINC = 0,
                            ps_MD_FAMINC = 100,
                            pm_COSTT4_A = 0,
                            ps_COSTT4_A = 500,
                            pm_POVERTY_RATE = 0,
                            ps_POVERTY_RATE = 2500,
                            pm_URBAN = 0,
                            ps_URBAN = 2500,
                            pm_PRIVATE = 0,
                            ps_PRIVATE = 2500,
                            pm_sigma = 10000,
                            ps_sigma = 10000

)


separate.fit <- separate.model$sample(data = separate.model.data, seed = 1234, refresh = 1e3)

separate.fit$summary()
```

Pooled model fit

Summary of model fit for main parameters:

```
pooled.model <- cmdstan_model(stan_file = "./Stan/pooled.stan")

pooled.model.data <- list(N = nrow(data.joined.stan),

                          SAT_ALL = data.joined.stan$SAT_ALL,
                          MD_FAMINIC = data.joined.stan$MD_FAMINC,
                          COSTT4_A = data.joined.stan$COSTT4_A,
                          POVERTY_RATE = data.joined.stan$POVERTY_RATE,
                          URBAN = data.joined.stan$URBAN,
                          PRIVATE = data.joined.stan$PRIVATE,
                          y = data.joined.stan$MD_EARN_WNE_P10,

                          pm_alpha = 0,
                          ps_alpha = 10000,
                          pm_SAT_ALL = 50,
```

```
                          ps_SAT_ALL = 500,
                          pm_MD_FAMINC = 0,
                          ps_MD_FAMINC = 100,
                          pm_COSTT4_A = 0,
                          ps_COSTT4_A = 500,
                          pm_POVERTY_RATE = 0,
                          ps_POVERTY_RATE = 2500,
                          pm_URBAN = 0,
                          ps_URBAN = 2500,
                          pm_PRIVATE = 0,
                          ps_PRIVATE = 2500,
                          pm_sigma = 10000,
                          ps_sigma = 10000

                          )


pooled.fit <- pooled.model$sample(data = pooled.model.data, seed = 1234, refresh = 1e3)

pooled.fit$summary()
```

Hierarchical model fit

Summary of model fit for main parameters:

Separate model Rhat

```
rhat.df <- tibble()

params <- c("alpha[1]", "beta_SAT_ALL[1]", "beta_MD_FAMINIC[1]", "beta_COSTT4_A[1]",
            "beta_POVERTY_RATE[1]", "beta_URBAN[1]", "beta_PRIVATE[1]")

for (param in params) {
  rhats <- extract_variable_matrix(separate.fit$draws(), variable = param) %>% apply(2, rhat)
  row <- tibble("Parameter" = param,
                "Chain 1" = rhats[1],
                "Chain 2" = rhats[2],
                "Chain 3" = rhats[3],
                "Chain 4" = rhats[4])
  rhat.df <- rbind(rhat.df, row)
}

rhat.df
```

Pooled model Rhat

```
params <- pooled.model$variables()$parameters %>% names()
rhat.df <- tibble()
for (param in params) {
  rhats <- extract_variable_matrix(pooled.fit$draws(), variable = param) %>% apply(2, rhat)
  row <- tibble("Parameter" = param,
                "Chain 1" = rhats[1],
                "Chain 2" = rhats[2],
```

```
                "Chain 3" = rhats[3],
                "Chain 4" = rhats[4])
  rhat.df <- rbind(rhat.df, row)
}
rhat.df
```

Hierarchical model Rhat

```
rhat.df <- tibble()

params <- c("alpha[1]", "beta_SAT_ALL[1]", "beta_MD_FAMINIC[1]", "beta_COSTT4_A[1]",
            "beta_POVERTY_RATE[1]", "beta_URBAN[1]", "beta_PRIVATE[1]")

for (param in params) {
  rhats <- extract_variable_matrix(hierarchical.fit$draws(), variable = param) %>% apply(2, rhat)
  row <- tibble("Parameter" = param,
                "Chain 1" = rhats[1],
                "Chain 2" = rhats[2],
                "Chain 3" = rhats[3],
                "Chain 4" = rhats[4])
  rhat.df <- rbind(rhat.df, row)
}

rhat.df
```

HMC diagnostics output

```
separate.fit$cmdstan_diagnose()
```

```
## Processing csv files: /tmp/Rtmp47bWiK/separate-202212100842-1-51736d.csv, /tmp/Rtmp47bWiK/separate-2
##
## Checking sampler transitions treedepth.
## 4000 of 4000 (100.00%) transitions hit the maximum treedepth limit of 10, or 2^10 leapfrog steps.
## Trajectories that are prematurely terminated due to this limit will result in slow exploration.
## For optimal performance, increase this limit.
##
## Checking sampler transitions for divergences.
## No divergent transitions found.
##
## Checking E-BFMI - sampler transitions HMC potential energy.
## E-BFMI satisfactory.
##
## Effective sample size satisfactory.
##
## The following parameters had split R-hat greater than 1.05:
##   beta_SAT_ALL[1], beta_SAT_ALL[9], beta_MD_FAMINIC[9], beta_COSTT4_A[1], beta_COSTT4_A[9], beta_POV
## Such high values indicate incomplete mixing and biased estimation.
## You should consider regularizating your model with additional prior information or a more effective
##
## Processing complete.
```

```
pooled.fit$cmdstan_diagnose()
```

```
## Processing csv files: /tmp/Rtmp47bWiK/pooled-202212100953-1-5d44dc.csv, /tmp/Rtmp47bWiK/pooled-20221
```

```
##
## Checking sampler transitions treedepth.
## Treedepth satisfactory for all transitions.
##
## Checking sampler transitions for divergences.
## No divergent transitions found.
##
## Checking E-BFMI - sampler transitions HMC potential energy.
## E-BFMI satisfactory.
##
## Effective sample size satisfactory.
##
## Split R-hat values satisfactory all parameters.
##
## Processing complete, no problems detected.
```

```
hierarchical.fit$cmdstan_diagnose()
```

```
## Processing csv files: /tmp/Rtmp47bWiK/hierarchical-202212100956-1-318ba6.csv, /tmp/Rtmp47bWiK/hierarc
##
## Checking sampler transitions treedepth.
## Treedepth satisfactory for all transitions.
##
## Checking sampler transitions for divergences.
## 182 of 4000 (4.55%) transitions ended with a divergence.
## These divergent transitions indicate that HMC is not fully able to explore the posterior distributio
## Try increasing adapt delta closer to 1.
## If this doesn't remove all divergences, try to reparameterize the model.
##
## Checking E-BFMI - sampler transitions HMC potential energy.
## E-BFMI satisfactory.
##
## Effective sample size satisfactory.
##
## Split R-hat values satisfactory all parameters.
##
## Processing complete.
```

K hat

```
pareto_k_table(separate.fit$loo())
```

```
##
## All Pareto k estimates are good (k < 0.5).
```

```
pareto_k_table(pooled.fit$loo())
```

```
## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.
```

```
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)       0     0.0%  <NA>
##  (0.5, 0.7]   (ok)         0     0.0%  <NA>
##    (0.7, 1]   (bad)        0     0.0%  <NA>
##    (1, Inf)   (very bad) 766   100.0%  0
```

```
pareto_k_table(hierarchical.fit$loo())
```

```
## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.
```

```
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     763   99.6%   133
##  (0.5, 0.7]   (ok)         1    0.1%   309
##    (0.7, 1]   (bad)        2    0.3%   111
##    (1, Inf)   (very bad)   0    0.0%   <NA>
```

RMSE

```
pooled.fit.coeff.df <- pooled.fit$summary() %>% slice(2:8)

alpha.hat <- pooled.fit.coeff.df %>% head(1) %>% select(median)
beta.hat <- pooled.fit.coeff.df %>% tail(-1) %>% select(median) %>% as.matrix()

test.X <- data.joined.stan.test %>%
  select(SAT_ALL, MD_FAMINC, AGE_ENTRY, COSTT4_A, POVERTY_RATE, PRIVATE) %>%
  as.matrix()

N <- nrow(data.joined.stan.test)
y.hat <- numeric(N)
y <- data.joined.stan.test$MD_EARN_WNE_P10
se <- numeric(N)
for (i in 1:N) {

  y.hat[i] <- alpha.hat+beta.hat%*%test.X[i,]
  se[i] <- (y.hat[[i]]-y[i])^2

}

rmse <- sqrt(mean(se))
```

Test set

```
data.joined.stan.test
```

```
## # A tibble: 30 x 11
##    REGION MD_EAR~1 SAT_ALL MD_FA~2 AGE_E~3 COSTT~4 POVER~5 URBAN PRIVATE DOCTO~6
##     <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>   <dbl>
##  1      5    53449    642.   43729    20.6   22278    8.90     1       0       1
##  2      5    41792    538.   37260    26.3   35949    9.30     1       1       0
##  3      5    45428    526    53444    21.0   46034    9.22     0       1       0
##  4      2    72980    685    61401    20.2   28010    8.78     1       0       1
##  5      8    80643    658.   74513    20.1   28927    7.16     1       0       0
##  6      6    55205    610    43043    21.3   27461    9.29     1       0       1
##  7      2    74578    660    97877    20.4   71388    9.08     1       1       1
##  8      5    40101    542.   49230    23.3   31522    8.02     1       1       0
##  9      5    54501    592.   23924    23.5   19384   11.6      1       0       1
## 10      5    52252    598.   44089    22.0   23478    6.77     1       0       1
## # ... with 20 more rows, 1 more variable: MASTER <dbl>, and abbreviated
## #   variable names 1: MD_EARN_WNE_P10, 2: MD_FAMINC, 3: AGE_ENTRY, 4: COSTT4_A,
```

```
## #   5: POVERTY_RATE, 6: DOCTORAL
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Separate model sensitivity analysis

Summary of model fit for main parameters with wide priors:

Summary of model fit for main parameters with narrow priors:

Pooled model sensitivity analysis

Summary of model fit for main parameters with wide priors:

Summary of model fit for main parameters with narrow priors:

Hierarchical model sensitivity analysis

Summary of model fit for main parameters with wide priors:

Summary of model fit for main parameters with narrow priors: