BDA Project

Niko Miller, Akseli Manninen and Santeri Löppönen

Contents

1. Project Description and Motivation

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. Furthermore, finding well-predictive features among the vast number of variables is pursed, and then evaluating predictive performance using these features with selected statistical models.

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

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 t 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

In the next section, we outline how 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. Moreover, we used common sense when trying 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 and common sense

The initial set of features chosen based on past literature and common sense is outlined in the table below:

	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 combined as one variable, because they were highly correlated and could easily be viewed as one entity. However, writing SAT scores had too few observations due to discontinued tracking. Therefore, a variable SAT_ALL was formed by summing the math and critical thinking SAT scores. After this, we have a total of 793 observations for these features.

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

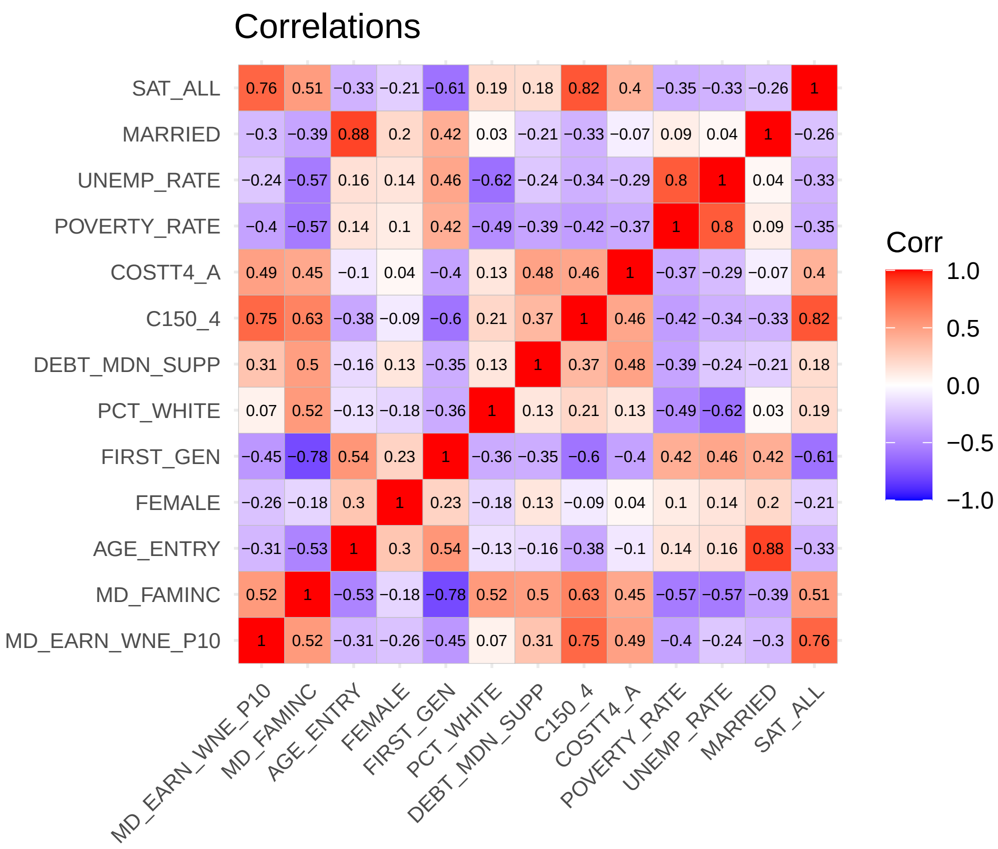
The table shows that our dependent variable - median earnings 10 years after entry - range from $24.2k to $103.2k. Overall one can see that there is great variation in almost all features as well. For example, age of entry varies from 19.6 to 33.8, average composite SAT scores range from 395 to 760, 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 whether there were any linear relationships between the features and the dependent variable. Lastly, we visually examined categorical variables and engineered new features that we believed to have predictive ability over future earnings.

Numerical Variables

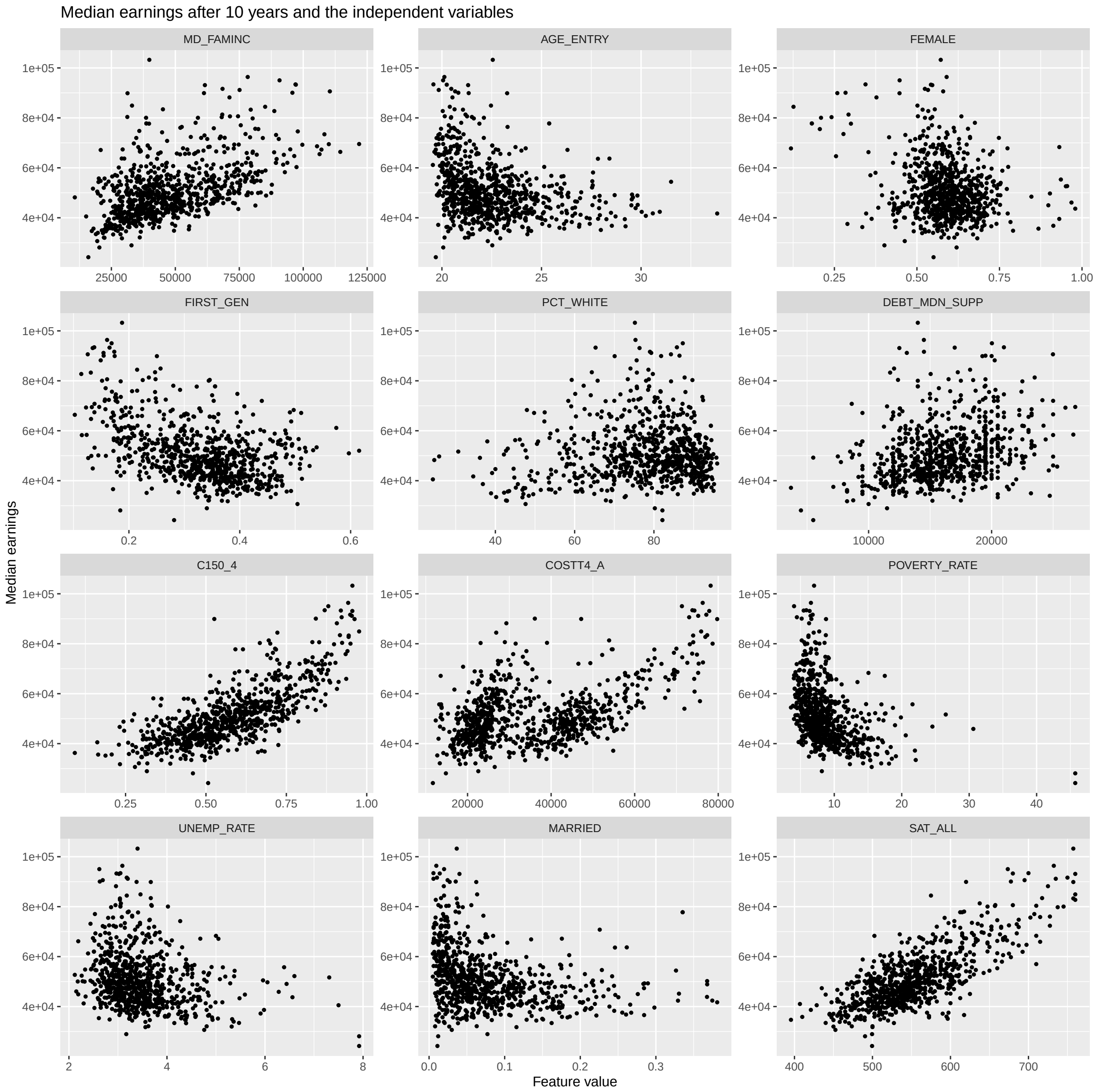
Correlation Analysis The figure below shows the (Pearson's) correlation matrix for the dependent variable and the features.



	MD_EARN_WNE_P10	MD_FAMINC	AGE_ENTRY	FEMALE	FIRST_GEN	PCT_WHITE	DEBT_MDN_SUPP	C150_4	COSTT4_A	POVERTY_RATE	UNEMP_RATE	MARRIED	SAT_ALL
SAT_ALL	0.76	0.51	−0.33	−0.21	−0.61	0.19	0.18	0.82	0.4	−0.35	−0.33	−0.26	1
MARRIED	−0.3	−0.39	0.88	0.2	0.42	0.03	−0.21	−0.33	−0.07	0.09	0.04	1	−0.26
UNEMP_RATE	−0.24	−0.57	0.16	0.14	0.46	−0.62	−0.24	−0.34	−0.29	0.8	1	0.04	−0.33
POVERTY_RATE	−0.4	−0.57	0.14	0.1	0.42	−0.49	−0.39	−0.42	−0.37	1	0.8	0.09	−0.35
COSTT4_A	0.49	0.45	−0.1	0.04	−0.4	0.13	0.48	0.46	1	−0.37	−0.29	−0.07	0.4
C150_4	0.75	0.63	−0.38	−0.09	−0.6	0.21	0.37	1	0.46	−0.42	−0.34	−0.33	0.82
DEBT_MDN_SUPP	0.31	0.5	−0.16	0.13	−0.35	0.13	1	0.37	0.48	−0.39	−0.24	−0.21	0.18
PCT_WHITE	0.07	0.52	−0.13	−0.18	−0.36	1	0.13	0.21	0.13	−0.49	−0.62	0.03	0.19
FIRST_GEN	−0.45	−0.78	0.54	0.23	1	−0.36	−0.35	−0.6	−0.4	0.42	0.46	0.42	−0.61
FEMALE	−0.26	−0.18	0.3	1	0.23	−0.18	0.13	−0.09	0.04	0.1	0.14	0.2	−0.21
AGE_ENTRY	−0.31	−0.53	1	0.3	0.54	−0.13	−0.16	−0.38	−0.1	0.14	0.16	0.88	−0.33
MD_FAMINC	0.52	1	−0.53	−0.18	−0.78	0.52	0.5	0.63	0.45	−0.57	−0.57	−0.39	0.51
MD_EARN_WNE_P10	1	0.52	−0.31	−0.26	−0.45	0.07	0.31	0.75	0.49	−0.4	−0.24	−0.3	0.76

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.

Bivariate plotting The panel of figures below show scatter plots between the dependent variable and all numerical features.



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.

Concluding remarks 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).

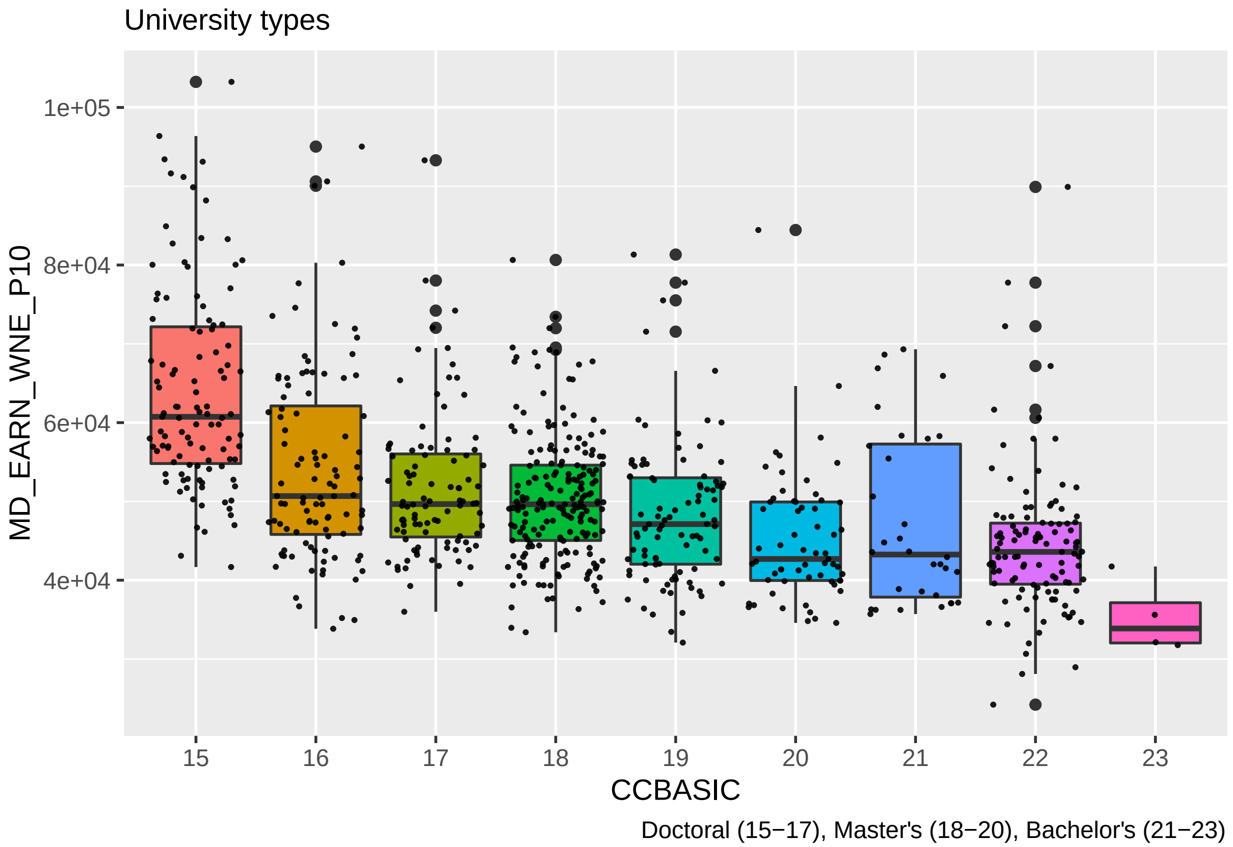
After including the obvious features, such as SAT scores, we included features with the following reasoning: if the correlation between a feature and the dependent variable was low and there was no observable dependency between the two in the scatter plot, the feature was excluded. Furthermore, to avoid 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 see if income after college differed among the categories.

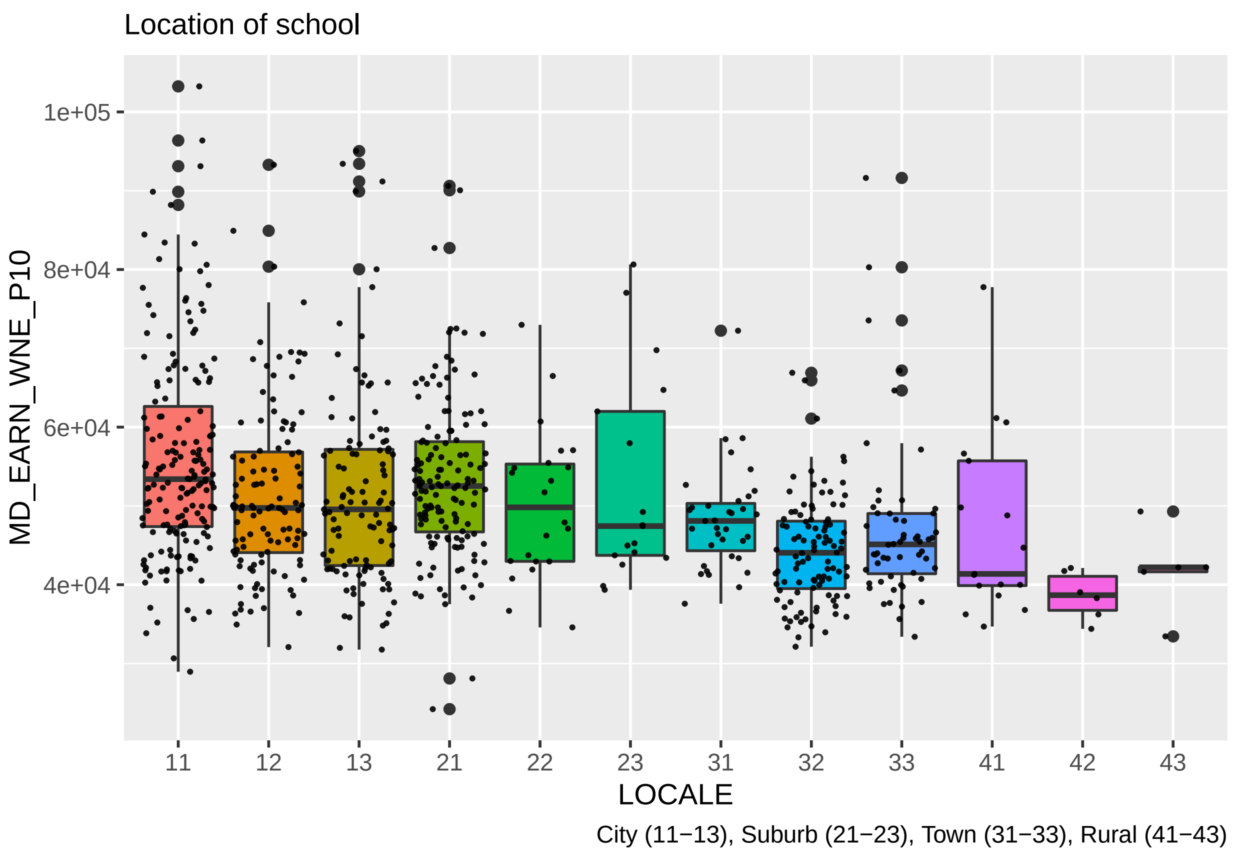
The categorical variable CCBASIC which represented the Carnegie Classification was divided into two binary MASTER and DOCTORAL. These variables represent if the college is classified as Master's college and university or Doctoral's college and university. If a college does not belong to either of those, it is a Bachelor's college and university. Other special focus colleges and universities were discarded from the dataset to keep the model more simple and avoid unnecessary outliers due to unconventional nature of some very specialized colleges. Our model does not attempt to provide accurate predictions for specialized colleges.

```r
ggplot(melted.categorial.ccbasic, aes(x=value, y=MD_EARN_WNE_P10, fill=value)) +
  geom_boxplot() +
  geom_jitter(color="black", size=0.4, alpha=0.9) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("University types") +
  xlab("CCBASIC") +
  labs(caption = "Doctoral (15-17), Master's (18-20), Bachelor's (21-23)")
```



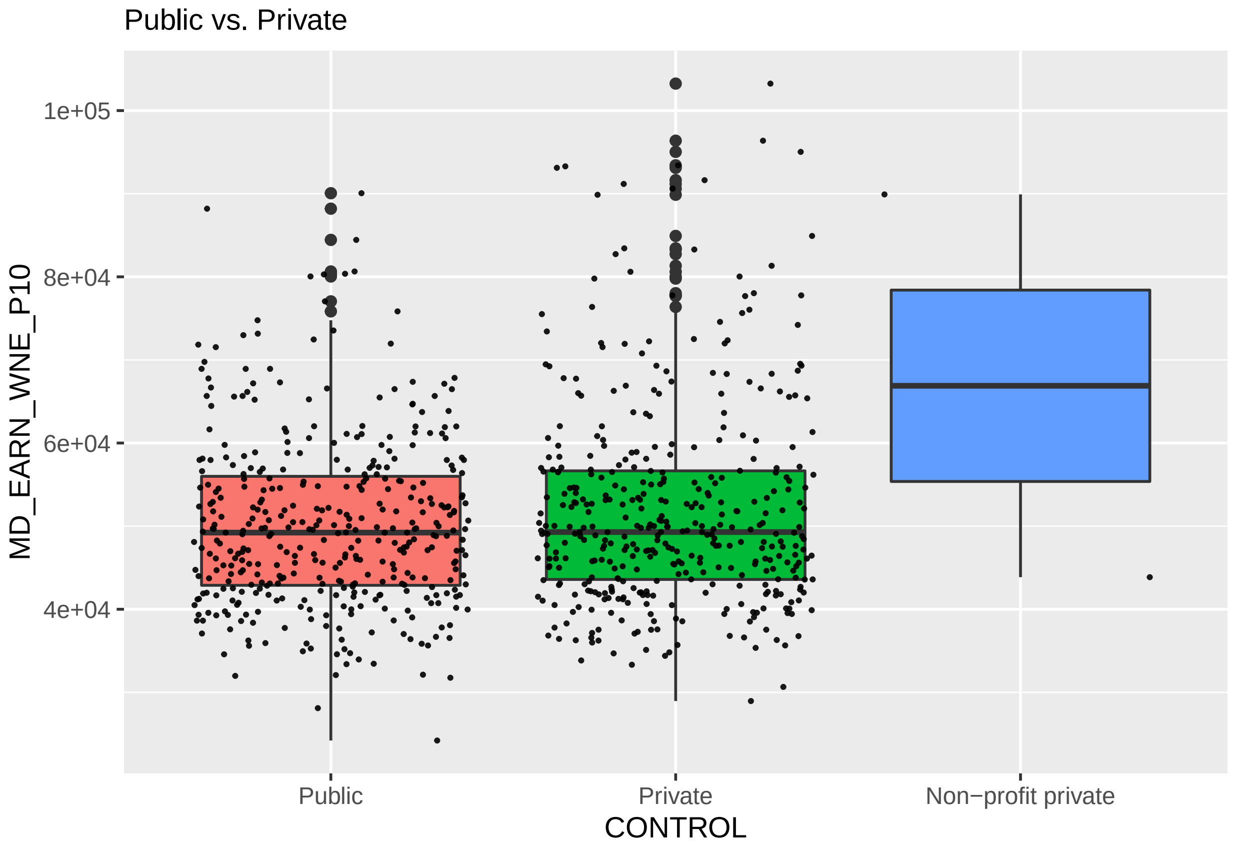
Based on an analysis on the variable LOCALE (whether college located in metropolis, city, suburb, town, or rural area) a new binary variable URBAN was generated, where value 1 represents any area that is not categorized as rural.

```
ggplot(melted.categorial.locale, aes(x=value, y=MD_EARN_WNE_P10, fill=value)) +
  geom_boxplot() +
  geom_jitter(color="black", size=0.4, alpha=0.9) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("Location of school") +
  xlab("LOCALE") +
  labs(caption = "City (11-13), Suburb (21-23), Town (31-33), Rural (41-43)")
```



The categorical variable CONTROL had three classes: public, for-profit private and nonprofit private. CONTROL was modified into a binary variable PRIVATE representing if the school is private or public school. There were only few non-profit private observations, so those were merged together with the private observations.

```
ggplot(melted.categorial.control, aes(x=value, y=MD_EARN_WNE_P10, fill=value)) +
  geom_boxplot() +
  geom_jitter(color="black", size=0.4, alpha=0.9) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("Public vs. Private") +
  xlab("CONTROL") +
  scale_x_discrete(labels = c("Public","Private","Non-profit private"))
```



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

Phase 3 - Feature selection with stepwise regression

TARKISTETTAVA PITÄÄKÖ ALLAOLEVA TEKSTI PAIKKANSA

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 CITY 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.

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	*AGE_ENTRY*	*float*	*Average age of entry*
4	*COSTT4_A*	*float*	*Average cost of attendance (academic year institutions)*
5	*POVERTY_RATE*	*float*	*Poverty rate*
6	*PRIVATE*	*binary*	*Carnegie Classification − −basic*

3. Description of the models

MATEMATIIKKA TARKISTETTAVA

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.

Description of the linear model

MIKÄ TÄÄ ON?

4. Informative or weakly informative priors, and justification of their choices.

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.

Ehdotus

This is a variable we will try to transform to logarithmic scale as impact of say, 10 000$ more income can be far greater for a low-income family than for a high-income family.

AGE_ENTRY ~ Normal(0, 2500)

Justification: We don't have a strong expectation of direction or magnitude of effect on income. As the age of entry could be somewhere in the range of 15 - 50 without considering outliers, a few dozen years difference could have dramatic changes in income either way, the standard deviation is set 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)

Tää ei oikein täsmää. Jos tää sais ison painon pienillä arvoilla niin se tarkottais että kun liikutaan esim 2% köyhyydestä 3% köyhyyteen niin on tosi isot vaikutukset tuloihin. Todellisuudessa alueet on varmaan aika samanlaisia hyvinvoivia alueita. Nollaköyhyys on intercept-arvo josta sitten lähetään askeleittain kattomaan että miten tulot muuttuu kun alueen köyhyys lisääntyy. RATKAISUEHDOTUS: Lasketaan sd:tä. Ei tarvii olla yhtä dramaattinen kuin binäärisillä muutujilla.

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.

MASTER ~ Normal(0, 2500)

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.

5. Stan code

TÄHÄN JOTAIN TRAIN TEST SPLITISTÄ?

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:

Pooled model

Summary of model fit for main parameters:

Hierarchical model

Summary of model fit for main parameters:

6. How to the Stan model was run, that is, what options were used.

This is also more clear as combination of textual explanation and the actual code line.

We used the following options:

1. Seed: 1234
2. Chains: 4
3. Iterations per chain: 2000

7. Convergence diagnostics (rhat, ESS, divergences) and what was done if the convergence was not good with the first try.

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.

Separate model

Pooled model

```r
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

```r
rhat.df <- tibble()

params <- c("alpha[1]", "beta_SAT_ALL[1]", "beta_MD_FAMINIC[1]",
            "beta_AGE_ENTRY[1]", "beta_COSTT4_A[1]", "beta_POVERTY_RATE[1]",
            "beta_MASTER[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
```

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.

As can be seen on the tables below, for the separate model the values for the first region are range from 54 to 96. In this region, there are the sample size N is xxx. This means that…

For the pooled model the ESS values for parameters range from 508 to 954. This means that most of the samples for all parameters are independent, although there is also autocorrelation in most cases. According to (Stan) if the effective sample size is larger than the number of samples, this can be due to antithetic Markov Chains which have negative autocorrelations on odd lags.

For the hierarchical model the ESS values for parameters are between 266 and 1080. This means that for some parameters there samples are more autocorrelated samples than independent ones. The analysis for the highest values is the same as for pooled.

Source: https://mc-stan.org/docs/2_19/reference-manual/effective-sample-size-section.html

Separate Model

```
params <- separate.model$variables()$parameters %>% names()
ess.df <- tibble()

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

for (param in params) {
  ess <- extract_variable_matrix(separate.fit$draws(), variable = param) %>% apply(2, ess_basic)
  row <- tibble("Parameter" = param,
                "ESS" = ess)
  ess.df <- rbind(ess.df, row)
}
ess.df <- ess.df  %>% group_by(Parameter) %>% summarise(ESS = sum(ESS)/n(),)

ess.df
```

Pooled Model

```
params <- pooled.model$variables()$parameters %>% names()
ess.df <- tibble()

for (param in params) {
  ess <- extract_variable_matrix(pooled.fit$draws(), variable = param) %>% apply(2, ess_basic)
  row <- tibble("Parameter" = param,
                "ESS" = ess)
  ess.df <- rbind(ess.df, row)
}
ess.df <- ess.df  %>% group_by(Parameter) %>% summarise(ESS = sum(ESS)/n(),)
ess.df
```

Hierarchical Model

```
params <- hierarchical.model$variables()$parameters %>% names()
ess.df <- tibble()

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

for (param in params) {
  ess <- extract_variable_matrix(hierarchical.fit$draws(), variable = param) %>% apply(2, ess_basic)
  row <- tibble("Parameter" = param,
                "ESS" = ess)
  ess.df <- rbind(ess.df, row)
}
ess.df <- ess.df  %>% group_by(Parameter) %>% summarise(ESS = sum(ESS)/n(),)

ess.df
```

HMC specific convergence diagnostics for all models

Based on the HMC specific convergence diagnostics, the separate model reached the initial maximum treedepth 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, treedepth, divergences, E-BFMI, effective sample size and split R-hat were satifactory.

For the hierarchical model, onyl 4 out of 4000 hit the maximum treedepth 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.

```
separate.fit$cmdstan_diagnose()

pooled.fit$cmdstan_diagnose()

hierarchical.fit$cmdstan_diagnose()
```

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.

```
separate.extract <- separate.fit$draws(variable = "y_rep")

y <- data.joined.stan$MD_EARN_WNE_P10

y_rep <- matrix(data = separate.extract[,1,], nrow = 1000)

separate.ppctitle <- ggtitle("Separate model",
                             "Comparing densities of y and y_rep")

ppc_dens_overlay(y, y_rep) + separate.ppctitle

# Pooled

pooled.extract <- pooled.fit$draws(variable = "y_rep")

y_rep <- matrix(data = pooled.extract[,1,], nrow = 1000)

pooled.ppctitle <- ggtitle("Pooled model",
                             "Comparing densities of y and y_rep")

ppc_dens_overlay(y, y_rep) + pooled.ppctitle

# Hierarchical

hierarchical.extract <- hierarchical.fit$draws(variable = "y_rep")

y_rep <- matrix(data = hierarchical.extract[,1,], nrow = 1000)

hierarchical.ppctitle <- ggtitle("Hierarchical model",
                             "Comparing densities of y and y_rep")

ppc_dens_overlay(y, y_rep) + hierarchical.ppctitle
```

LOO

The ELPD is the theoretical expected log pointwise predictive density for a new dataset (Eq 1 in VGG2017), which can be estimated, e.g., using cross-validation. elpd_loo is the Bayesian LOO estimate of the expected log pointwise predictive density (Eq 4 in VGG2017) and is a sum of N individual pointwise log predictive densities. Probability densities can be smaller or larger than 1, and thus log predictive densities can be negative or positive. For simplicity the ELPD acronym is used also for expected log pointwise predictive probabilities for discrete models. Probabilities are always equal or less than 1, and thus log predictive probabilities are 0 or negative.

LOO (Model comparison)

Yläpuolella kopsattu, kirjoita uusiks. Lähde alla

Lähde: https://mc-stan.org/loo/reference/loo-glossary.html

With $elpd_{loo-cv}$ we should select the model with the highest value. From the tables below we can see that in this case the suggested model would be separate model, the second best hierarchical model and the third best pooled model.

```
separate.loo <- separate.fit$loo()

pooled.loo <- pooled.fit$loo()

hierarchical.loo <- hierarchical.fit$loo()

loo_compare(separate.loo, pooled.loo, hierarchical.loo)
```

K hat

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

pareto_k_table(pooled.fit$loo())

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

```
plot(separate.fit$loo(), main = "Separate model PSIS diagnostics")

plot(pooled.fit$loo(), main = "Pooled model PSIS diagnostics")

plot(hierarchical.fit$loo(), main = "Hierarchical model PSIS diagnostics")
```

Based on $\hat{K}$ the predictions are … too optimistic/reliable? - Why pooled model looks like that - overfitting or something?

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. First we split the complete data set of 793 observations into a training set of 763 observations and a test set of 30 observations. Second we fit the models to the training set. Then we use the median of the posterior distribution for each parameter in our model to predict the dependent variable based on the values of the independent variables in the test set. Lastly, after obtaining the predicted values, we calculate the RMSE with respect to the actual observed values of the dependent variable in the test set.

Pooled model

```
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

}

y.hat <- y.hat %>%
  as.matrix()
```

Plot of squared errors:

```
se %>% plot(type="b", pch=20)
```

RMSE:

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

10. Sensitivity analysis with respect to prior choices (i.e. checking whether the result changes a lot if prior is changed). This should be reported for all models.

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.

Separate model

Summary of model fit for main parameters with wide priors:

Summary of model fit for main parameters with narrow priors:

Pooled model

Summary of model fit for main parameters with wide priors:

Summary of model fit for main parameters with narrow priors:

Hierarchical model

Summary of model fit for main parameters with wide priors:

Summary of model fit for main parameters with narrow priors:

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.

- **VOISI EHKÄ MAINITA JOTAIN PREDICTIVE ASSESSMENTISTA, ESIM. KAIKILLE MALLEILLE JONKUNLAINEN ASSESSMENT TAI MONIPUOLISEMPI**

12. Conclusion what was learned from the data analysis.

13. Self-reflection of what the group learned while making the project.

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 engineering 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.

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.

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

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")
```

```
## Start:  AIC=13491.39
## MD_EARN_WNE_P10 ~ SAT_ALL + MD_FAMINC + AGE_ENTRY + COSTT4_A +
##     POVERTY_RATE + URBAN + PRIVATE + DOCTORAL + MASTER
##
##                Df  Sum of Sq        RSS   AIC
## - AGE_ENTRY     1 1.6055e+06 3.3269e+10 13489
## - DOCTORAL      1 1.5259e+07 3.3283e+10 13490
## <none>                       3.3268e+10 13491
## - MASTER        1 1.1313e+08 3.3381e+10 13492
## - MD_FAMINC     1 1.9443e+08 3.3462e+10 13494
## - POVERTY_RATE  1 3.8991e+08 3.3658e+10 13498
## - URBAN         1 6.4142e+08 3.3909e+10 13504
## - PRIVATE       1 3.0432e+09 3.6311e+10 13556
## - COSTT4_A      1 5.2787e+09 3.8546e+10 13602
## - SAT_ALL       1 1.0261e+10 4.3529e+10 13695
##
## Step:  AIC=13489.42
## MD_EARN_WNE_P10 ~ SAT_ALL + MD_FAMINC + COSTT4_A + POVERTY_RATE +
##     URBAN + PRIVATE + DOCTORAL + MASTER
##
##                Df  Sum of Sq        RSS   AIC
## - DOCTORAL      1 1.4133e+07 3.3283e+10 13488
## <none>                       3.3269e+10 13489
## - MASTER        1 1.1259e+08 3.3382e+10 13490
## - MD_FAMINC     1 2.8949e+08 3.3559e+10 13494
## - POVERTY_RATE  1 3.9636e+08 3.3666e+10 13496
## - URBAN         1 6.4507e+08 3.3914e+10 13502
## - PRIVATE       1 3.4011e+09 3.6670e+10 13562
## - COSTT4_A      1 5.4513e+09 3.8721e+10 13604
## - SAT_ALL       1 1.0260e+10 4.3529e+10 13693
##
## Step:  AIC=13487.75
## MD_EARN_WNE_P10 ~ SAT_ALL + MD_FAMINC + COSTT4_A + POVERTY_RATE +
##     URBAN + PRIVATE + MASTER
##
##                Df  Sum of Sq        RSS   AIC
## <none>                       3.3283e+10 13488
## - MASTER        1 1.2072e+08 3.3404e+10 13488
## - MD_FAMINC     1 2.8022e+08 3.3564e+10 13492
## - POVERTY_RATE  1 4.0282e+08 3.3686e+10 13495
## - URBAN         1 7.3135e+08 3.4015e+10 13502
## - PRIVATE       1 3.7419e+09 3.7025e+10 13567
## - COSTT4_A      1 5.6028e+09 3.8886e+10 13605
## - SAT_ALL       1 1.1584e+10 4.4867e+10 13714
```

```
##
## Call:
## lm(formula = MD_EARN_WNE_P10 ~ SAT_ALL + MD_FAMINC + COSTT4_A +
##     POVERTY_RATE + URBAN + PRIVATE + MASTER, data = data.joined.model)
##
## Coefficients:
##  (Intercept)       SAT_ALL     MD_FAMINC      COSTT4_A  POVERTY_RATE
##   -1.232e+04     9.379e+01     4.465e-02     3.856e-01    -2.362e+02
##        URBAN       PRIVATE        MASTER
##    2.287e+03    -8.502e+03     8.419e+02
```