Predicting Alumni Income Based on University-level Data

Niko Miller (717199), Akseli Manninen (665306), Santeri Löppönen (730273)

Project Introduction

- Past studies have concluded that a strong connection exists between higher education and income[1]
- We further examine this link by regressing median earnings of students 10 years after entry on university-level covariates
- We extend previous studies by considering demographic and geographic factors
- Data on 6681 U.S. universities spans over 9 regions, 59 state post codes, 2430 cities
- Bayesian models:
  - Pooled model for all universities
  - Separate and hierarchical model with region-grouping

[1](Card, 1999)

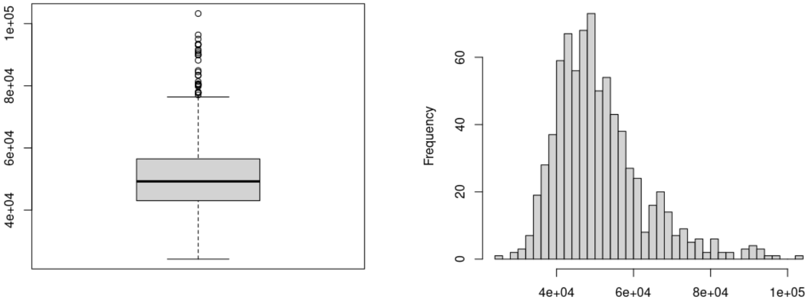
Data Description

- Most recent institutional-level college scorecard data from U.S. Department of Education[1]
- 6681 observations/universities and 2989 variables
- Aggregate data for each university
    - Institutional characteristics, enrollment, student aid, costs and student outcomes
- Data quality is quite good, but many variables have lots of NAs
- Why this dataset?

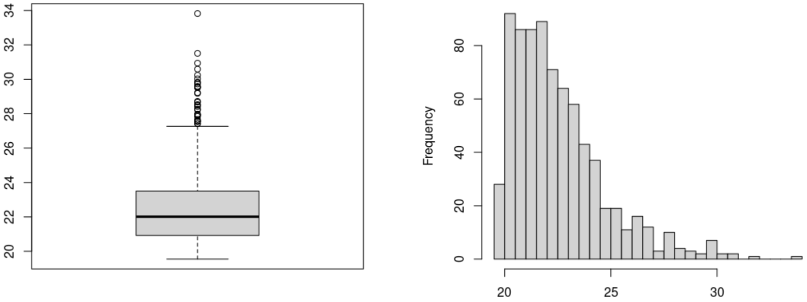
[1]Available at: https://collegescorecard.ed.gov/data

Data Description

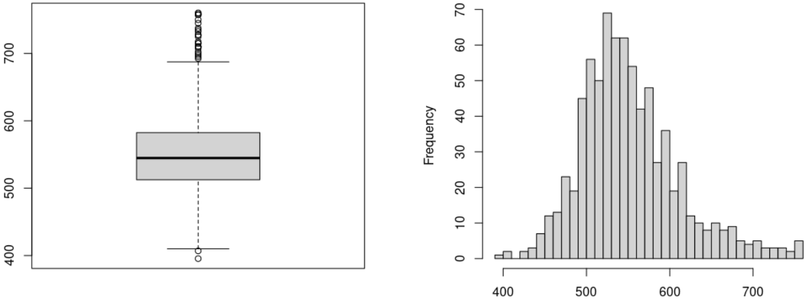
Dependent variable / Median earnings 10 yr. after entry



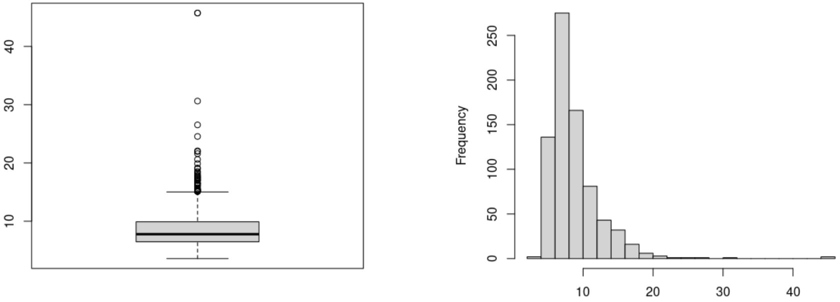
Age of entry



SAT score



Poverty rate



Analysis Problem

- Predict median earnings of alumni with university level covariates
- Multivariate regression setting

- Mathematical model:
  - Y is the dependent variable
  - X is the data matrix of covariates
  - β is the regression coefficient vector
  - ε is the residual error
  - Hatted variables $\hat{Y}, \hat{\beta}$ indicate LS[1] estimates

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1q} \\ 1 & x_{21} & x_{22} & \dots & x_{2q} \\ 1 & x_{31} & x_{32} & \dots & x_{3q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

[1]LS= Least Squares

Feature Selection

- Raw data had ~3000 variables → feature selection was a focus area
- Feature selection in 3 phases:
    1. Select initial set of features based on common intuition and literature
    2. Assess correlations, linear relationships, and dummy effects of categorical variables
    3. Use stepwise regression to suggest final subset of features

Feature Selection – Phase 1

Initial set of features

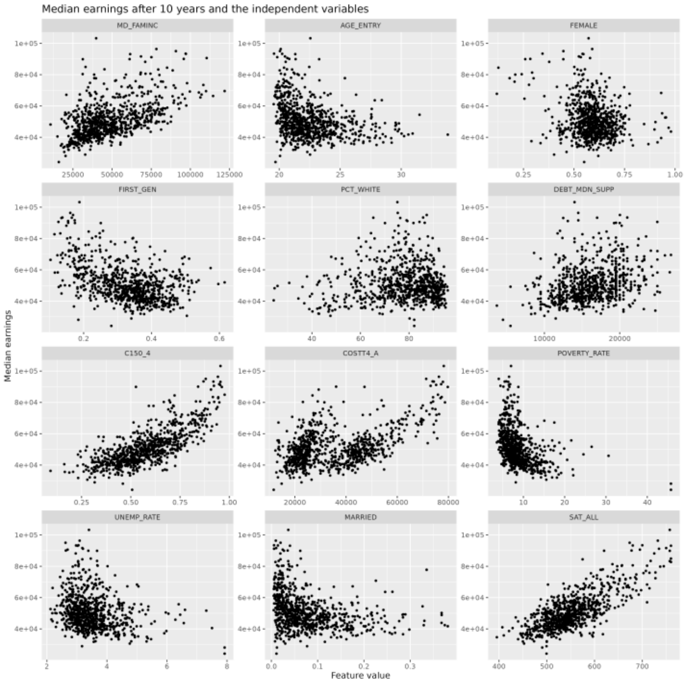
	Name	Data type	Description
1	SATVRMID	integer	Midpoint of SAT scores at the institution (critical reading)
2	SATMTMID	integer	Midpoint of SAT scores at the institution (math)
3	SATWRMID	integer	Midpoint of SAT scores at the institution (writing)
4	MD_FAMINC	double	Median family income
5	AGE_ENTRY	double	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

Comments

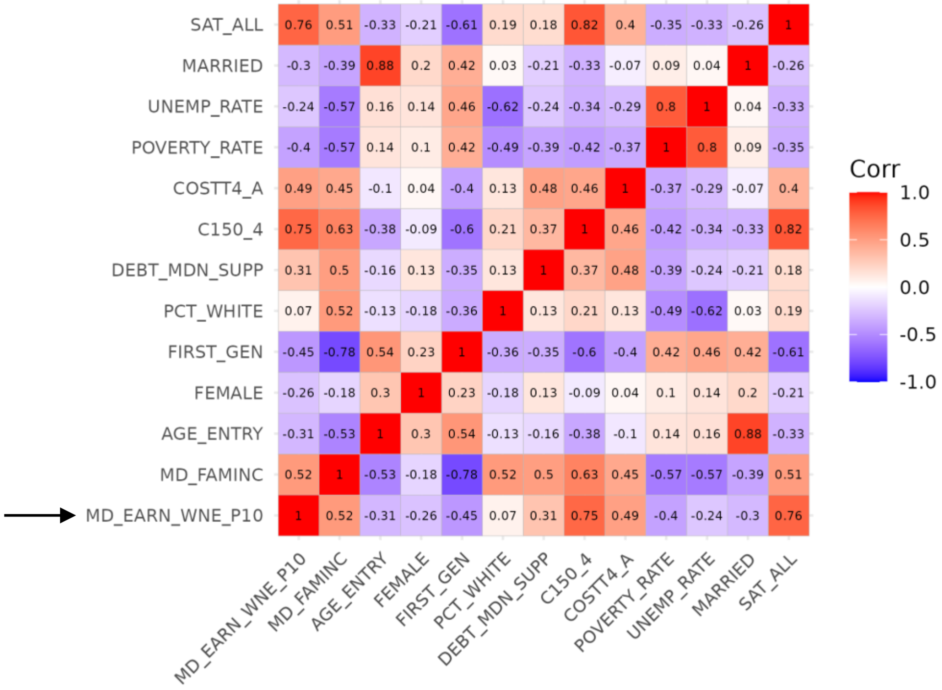
- Ability assumed to have positive effect
- Privileged background assumed to have positive effect
- Age of entry assumed to have adverse effect
- Gender split assumed to have an effect due to empirically observed gender gap
- Marriage and veteran rates assumed to have adverse affect
- Cost and debt assumed to have positive effect
- Assumption that location, control, classification has some effect

Feature Selection – Phase 2

Numerical variables

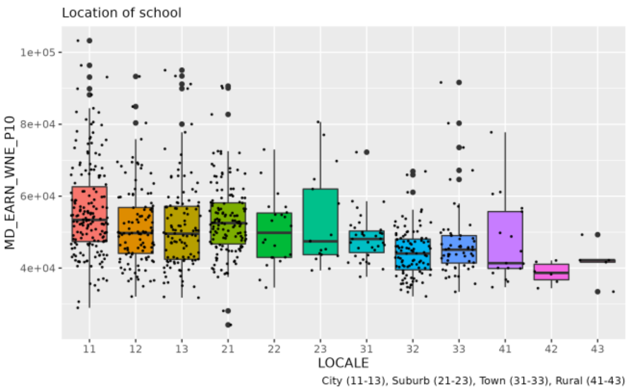


Pearson's correlation

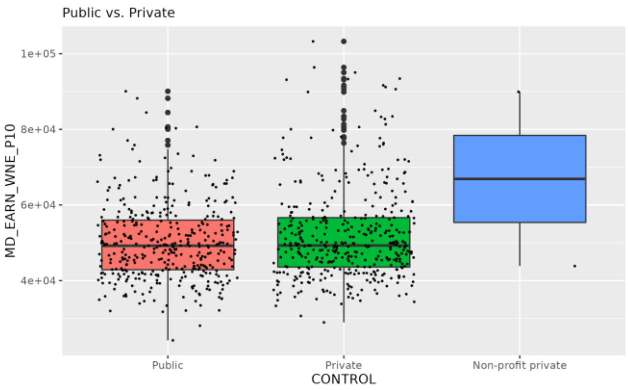


Feature Selection – Phase 2

University locale



University control



Comments

- We created a dummy variable URBAN if locale is not rural

Comments

- We created a dummy variable for private institutions (for-profit and non-profit)

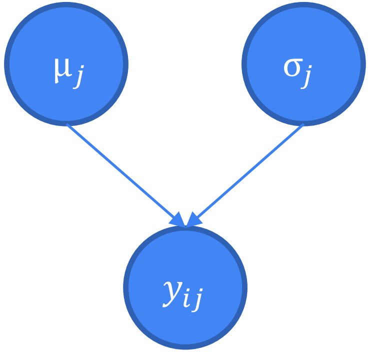
Feature Selection – Phase 3

- Lastly, we used stepwise regression with backward elimination to finetune our model
- Final model suggested by the stepwise regression model:

$$
\begin{aligned}
\boldsymbol{Y} = {} & \beta_{SAT\_ALL}\boldsymbol{x}_{SAT\_ALL} + \beta_{MD\_FAMINC}\boldsymbol{x}_{MD\_FAMINC} + \beta_{COSTT4\_A}\boldsymbol{x}_{COSTT4\_A} \\
& + \beta_{POVERTY\_RATE}\boldsymbol{x}_{POVERTY\_RATE} + \beta_{URBAN}\boldsymbol{x}_{URBAN} + \beta_{PRIVATE}\boldsymbol{x}_{PRIVATE} + \varepsilon
\end{aligned}
$$

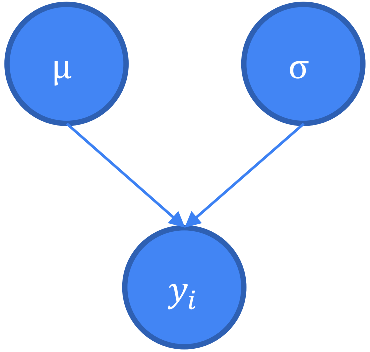
- We proceeded to Stan with this final model

Model Descriptions



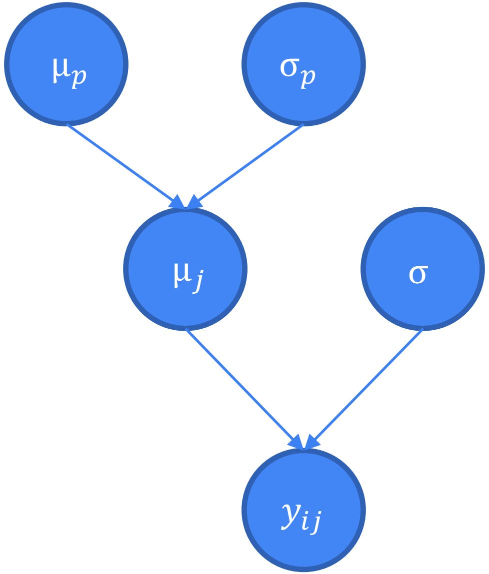
Separate Model

$$y_{ij} \mid \mu_j, \sigma_j \sim N(\mu_j, \sigma_j^2),$$
$$\mu_j = \alpha_j + \boldsymbol{X\beta}_j$$
$$\alpha_j, \beta_j, \mu_j, \sigma_j \sim N$$



Pooled Model

$$y_i \mid \mu, \sigma \sim N(\mu, \sigma^2),$$
$$\mu = \alpha + \boldsymbol{X\beta}$$
$$\alpha, \beta, \mu, \sigma \sim N$$



Hierarchical Model

$$y_{ij} \mid \mu_j, \sigma \sim N(\mu_j, \sigma^2),$$
$$\mu_j = \alpha_j + \boldsymbol{X\beta}_j$$
$$\alpha_j, \beta_j \mid \mu_P, \sigma_P \sim N(\mu_P, \sigma_P^2)$$
$$\mu_P, \sigma_P \sim N$$

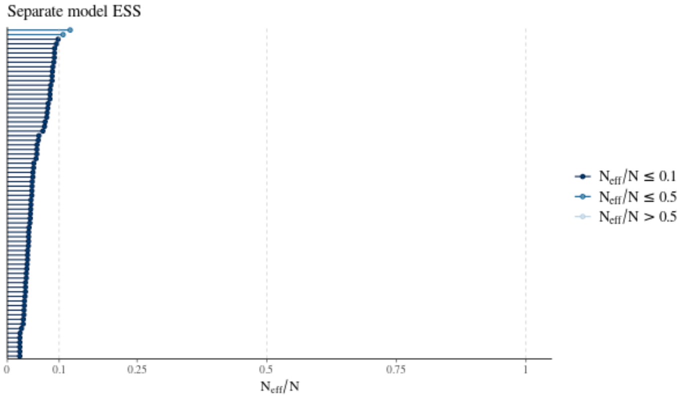
Choice of Priors

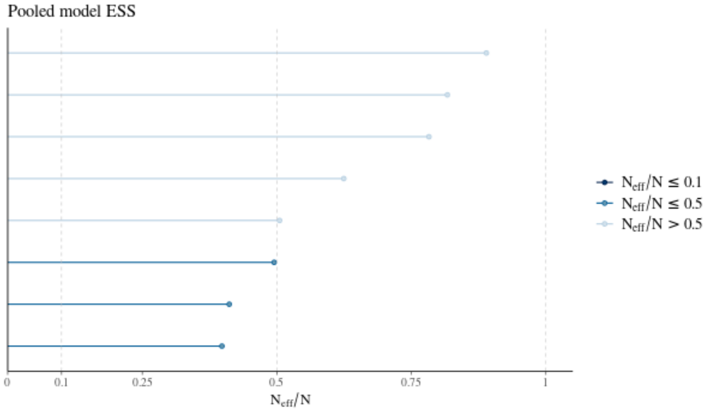
- We tried to conduct background research on the prior choices whenever possible
    - E.g., mean on SAT effect was based on national averages on income and SAT scores
- If no meaningful analysis on effect could be done, we assumed no effect on average
- We chose sufficiently large standard deviations to prevent informativeness in the priors
- We tried to avoid prior-data conflict by keeping the level of informativeness between location and scale parameters constant

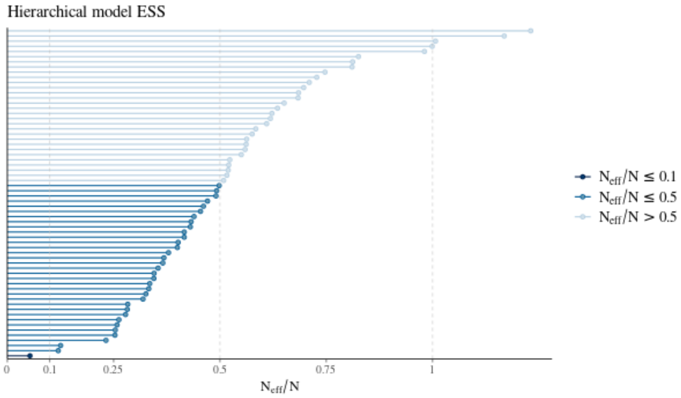
Stan Setup

- Computing: JupyterLab
- Stan interface: cmdstanr
- Chains: 4 (default)
- Iterations per chain: 2 000 (default: 1 000 warmup, 1 000 sampling)
- Seed for RNG: 1234

Convergence Diagnostics - ESS



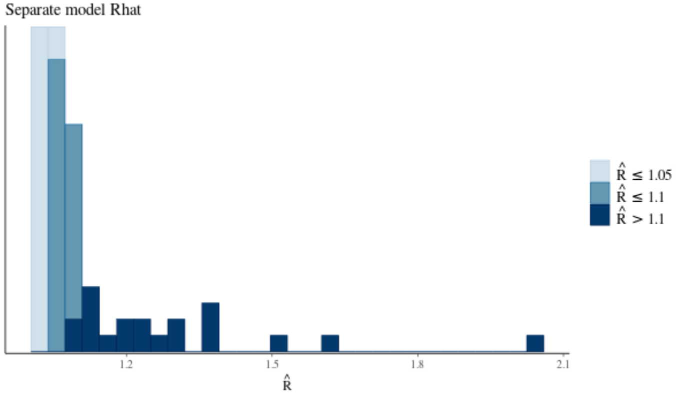


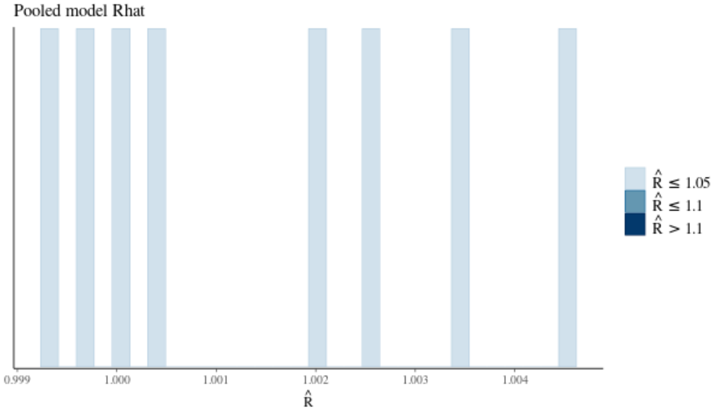


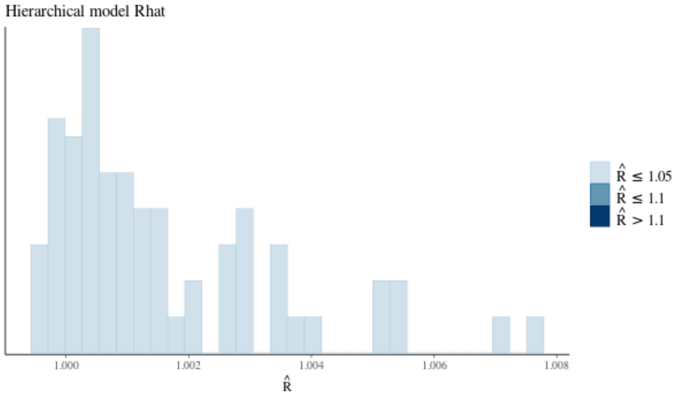
Ratios of effective sample size to total sample size

- *light*: between 0.5 and 1 (high)
- *mid*: between 0.1 and 0.5 (good)
- *dark*: below 0.1 (low)

Convergence Diagnostics – Rhat







Rhat values

- *light*: below 1.05 (good)
- *mid*: between 1.05 and 1.1 (ok)
- *dark*: above 1.1 (too high)

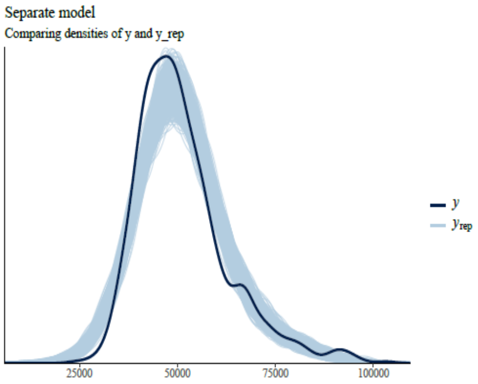
Convergence Diagnostics – HMC specific

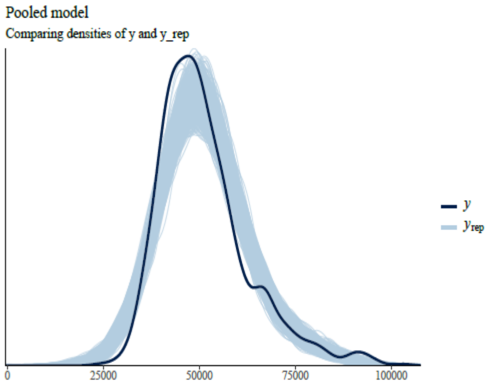
- Treedepth:
  - For pooled and hierarchical models satisfactory
  - In separate model all transitions hit the max treedepth

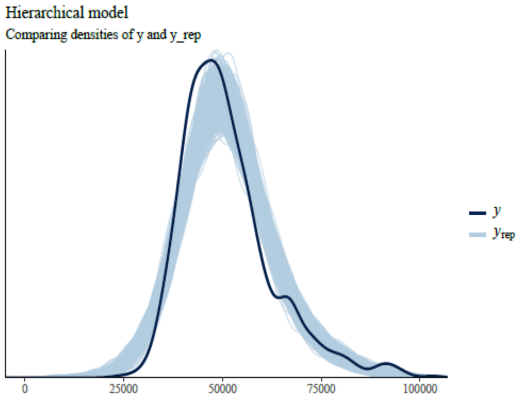
- Divergences:
  - For separate and pooled models no divergent transitions
  - In hierarchical model ≈ 4% divergent transitions

Posterior Predictive Checks

- Observed values vs. posterior predictions







Model Comparison - LOO

- Model performance order based on log pointwise predictive density (ELPD)
  1. **Best:** Separate model (-7700)
  2. Hierarchical model
  3. Pooled model

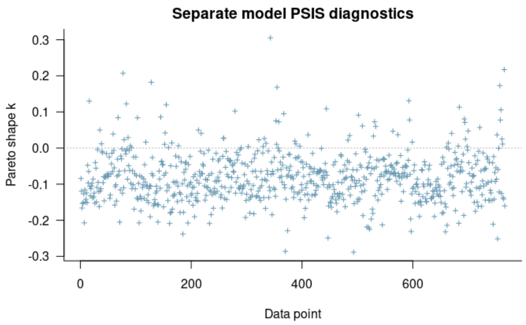
	elpd_diff	se_diff
separate	0.0	0.0
hierarchical	-49.7	6.2
pooled	-7450000.0	75700.0

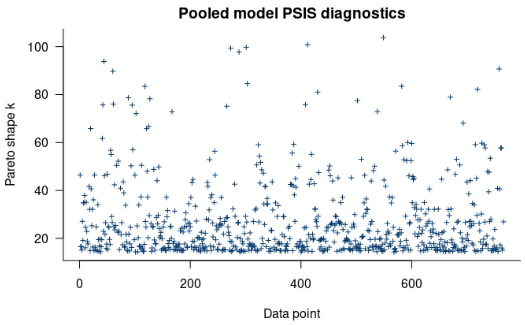
elpd_loo is an estimate of the expected log pointwise predictive denstity (ELPD). elpd_loo sums individual pointwise log predictive densities.[1]

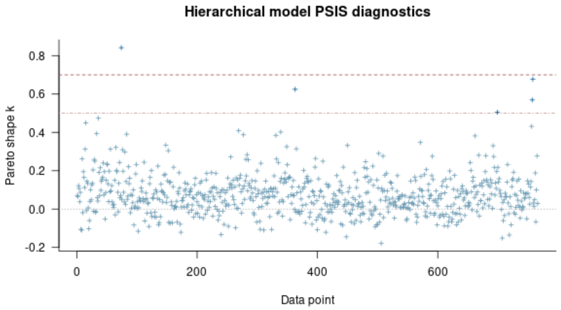
[1](Stan, n.d.)

Model Comparison - $\hat{k}$

- PSIS-LOO estimates of the separate model can be considered reliable since Khat values < 0.5
- PSIS-LOO estimates of the pooled model are likely too optimistic (biased)[1] since Khat values >> 0.7
- PSIS-LOO estimates of the hierarchical model are mostly reliable but few Khat values > 0.7







[1](Gabry, Gelman, Vehtari, 2016.)

Predictive Performance Assessment

Approach

- Train/test split with 763/30 obs.
- Fit all models on train set
- Use fitted model to predict on test set
- Compute Root Mean Squared Error (RMSE)
- Assess model output for RMSE

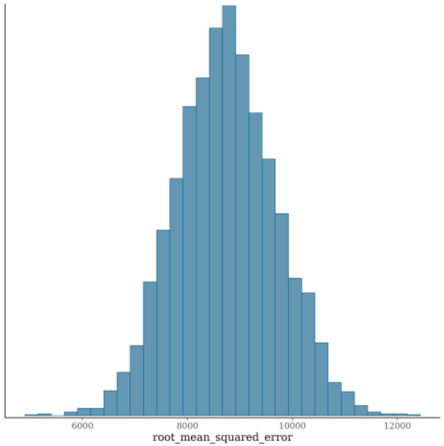
Model output

Description: df [1 × 10]

variable <chr>	mean <S3: AsIs>	median <S3: AsIs>	sd <S3: AsIs>	mad <S3: AsIs>	q5 <S3: AsIs>	q95 <S3: AsIs>	rhat <S3: AsIs>	ess_bulk <int>	ess_tail <int>
root_mean_squared_error	8740.40	8731.02	952.35	948.75	7231.03	10361.85	1.00	3843	4016

1 row

MCMC posterior histogram for RMSE



→ Decent predictions as the sample st.dev. in the complete set (train+test) is $11.6k for median earnings

1st Separate, 2nd Hierarchical, 3rd Pooled

Prior Sensitivity Analysis

- The width of priors had been a major point of discussion and uncertainty when setting original priors
- Experiment with priors of different widths:
  - Wide: 3x original prior sd
  - Narrow: 0.5x original prior sd
- Narrow priors perform better, Rhat, Khat, RMSE and ESS values improve
- For separate narrow model, all Rhat values reduced to below 1.01 from up to 1.12
- Estimates for coefficients are not affected much
- Wider priors lead to more convergence issues, worse predictive performance and lower ESS

Conclusion: Narrow priors perform better and should be used

Issues and Potential Improvements

- Narrowing down priors had a positive impact → Models could be improved by fine-tuning priors
- Pooled model had extreme k hat values and low elpd → Model could be rebuilt to prevent overfitting
- Separate model took ~1h to sample in Jupyter → Model could be made more efficient computationally
- Test set was small → Could be interesting to see how model fit and predictive performance is affected by altering the size of the test set

Conclusions

Project: Bayesian multivariate linear regression on median alumni earnings 10 years after entry in 793 universities in the U.S.

- Clear link found between education and future earnings; the model has decent predictive power in terms of RMSE

Three models: Pooled, Hierarchical and Separate

- Pooled model is prone to overfitting
- Separate model has issues with convergence and is slow to run, otherwise good performance, preferred by elpd and has lowest RMSE
- Hierarchical model has strong performance overall and doesn't have major issues

Prior sensitivity analysis shows that model performance can be improved by using narrower priors

Contact: *Niko Miller: niko.miller@aalto.fi, Akseli Manninen: akseli.manninen@aalto.fi, Santeri Löppönen: santeri.lopponen@aalto.fi*

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.
- Stan. (n.d) Effective Sample Size [Website]. Retrieved from: https://mc-stan.org/docs/2_19/referencemanual/effective-sample-size-section.html
- Gabry, Gelman, Vehtari. (2016) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.