Regression Analysis of Depression Data

Anton Hung, Bofan Chen, Digvijay Yadav, Kevin Rouse, Joe Gyorda (Team Zoey 101)

Overview

- **Introduction and Motivation**
- **Exploratory Data Analysis**
  - Missing values
  - Correlations
  - Distributions
- **Identification of Most Informative Predictors**
  - Stepwise Regression
  - LASSO Regression
- **Comparison of Regression Models**
  - GLM
  - GLMM
  - GEE
- **Discussion**

Introduction and Motivation

Motivation

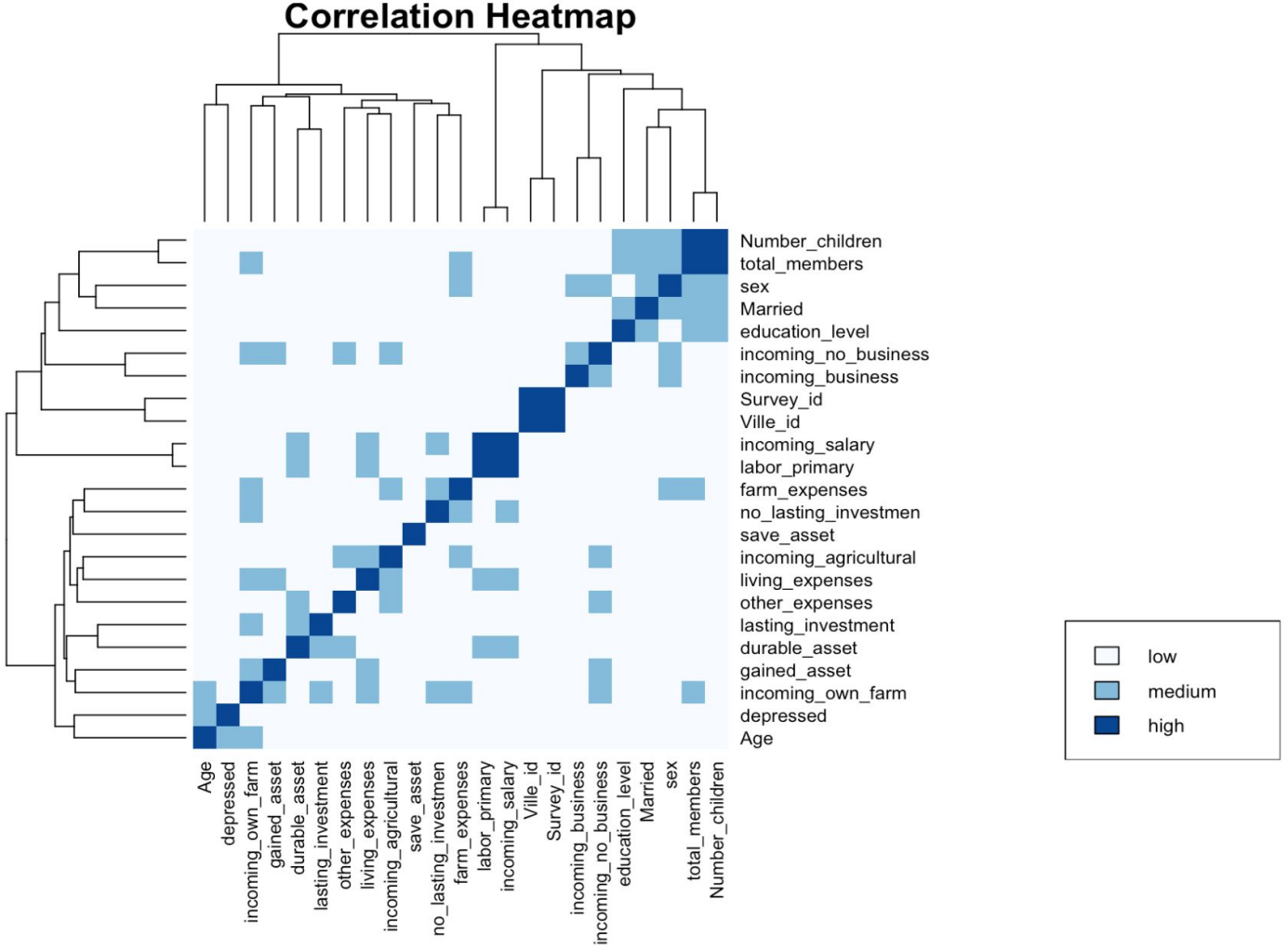
- Analyzing depression in rural areas
- An estimated 10% of adults suffer from depression globally
- What are some sociodemographic factors that we can target in order to combat the rising rates of depression?

U.S. News. (2022, September 19). *Depression Affects Almost 1 in 10 Americans*. U.S. News & World Report L.P.

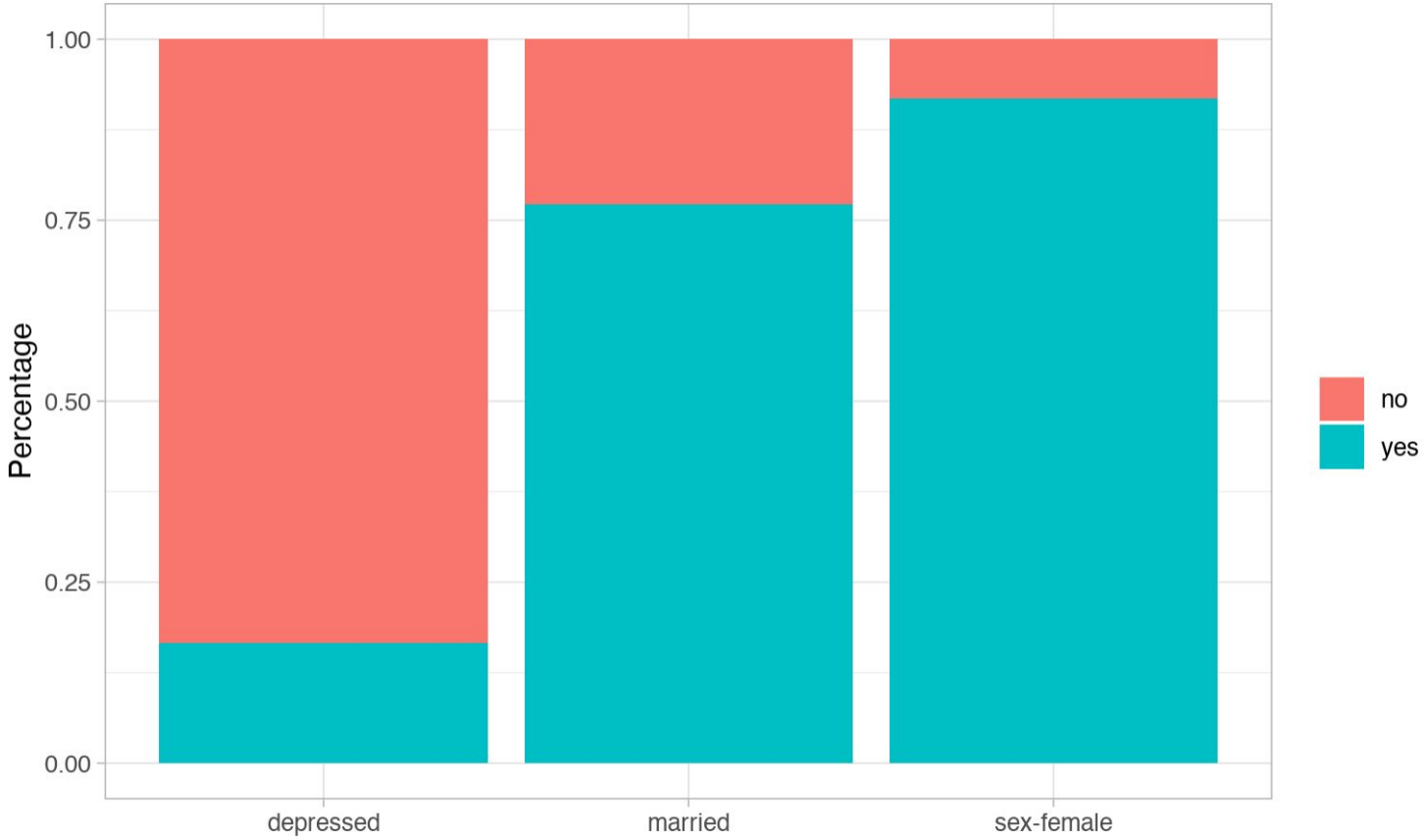
Introduction to Dataset and Depression Analysis

- The dataset was obtained from Kaggle
- Data was collected in 2015 in rural areas in Western Kenya
- The dataset contains 23 variables for 1429 people
  - We are analyzing which variables are most related to depression
- There were only 20 missing values for “no lasting investment”; these individuals were not included in analysis

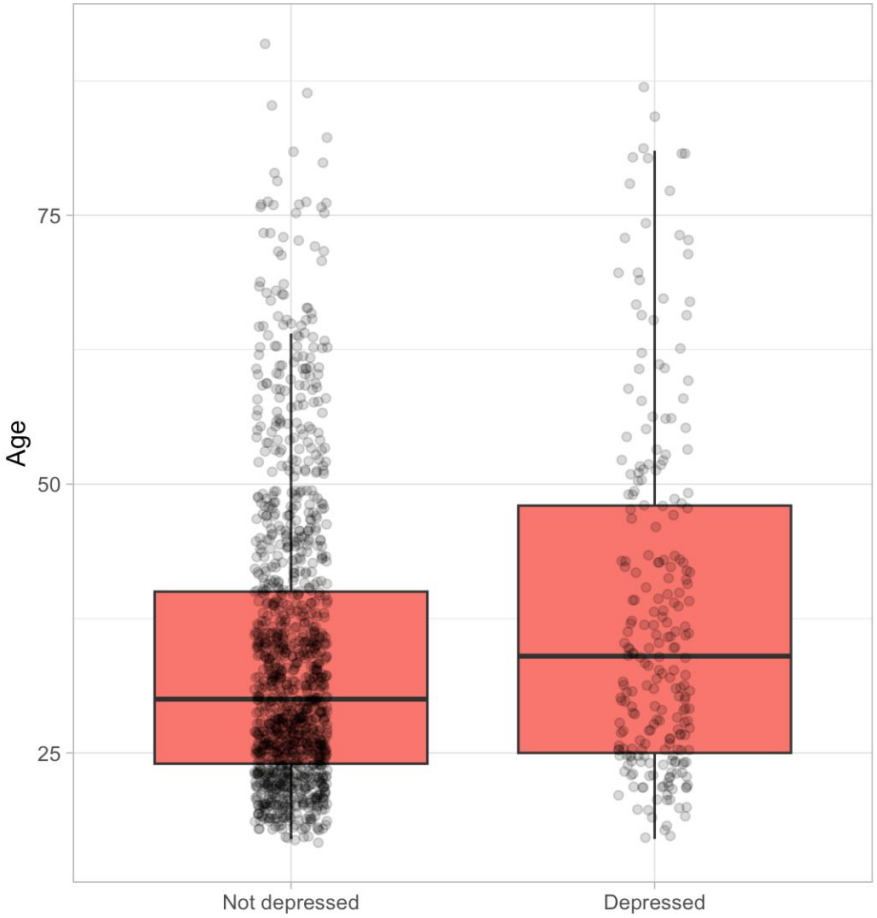
Exploratory Data Analysis


Correlation Heatmap
Number_children
total_members
sex
Married
education_level
incoming_no_business
incoming_business
Survey_id
Ville_id
incoming_salary
labor_primary
farm_expenses
no_lasting_investmen
save_asset
incoming_agricultural
living_expenses
other_expenses
lasting_investment
durable_asset
gained_asset
incoming_own_farm
depressed
Age
Age
depressed
incoming_own_farm
gained_asset
durable_asset
lasting_investment
other_expenses
living_expenses
incoming_agricultural
save_asset
no_lasting_investmen
farm_expenses
labor_primary
incoming_salary
Ville_id
Survey_id
incoming_business
incoming_no_business
education_level
Married
sex
total_members
Number_children
low
medium
high

Proportion of participants who are depressed, married, or female



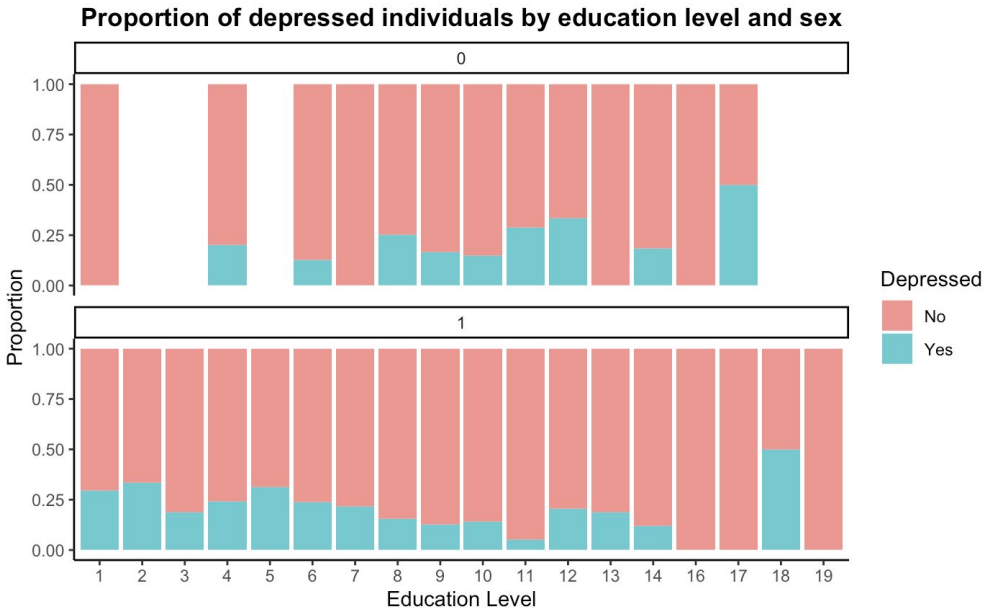
Age for depressed vs not depressed

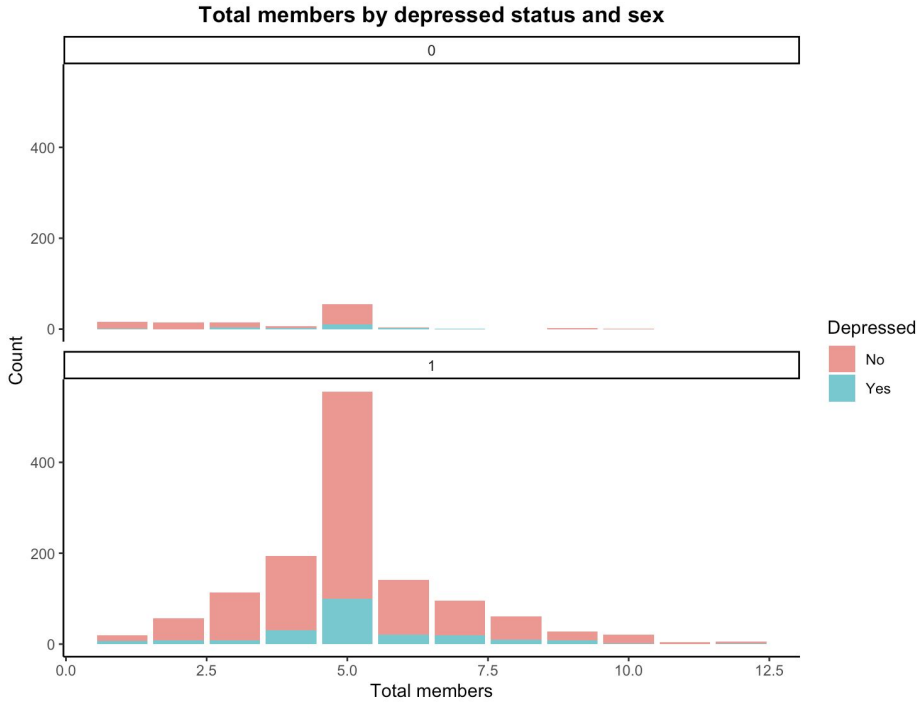


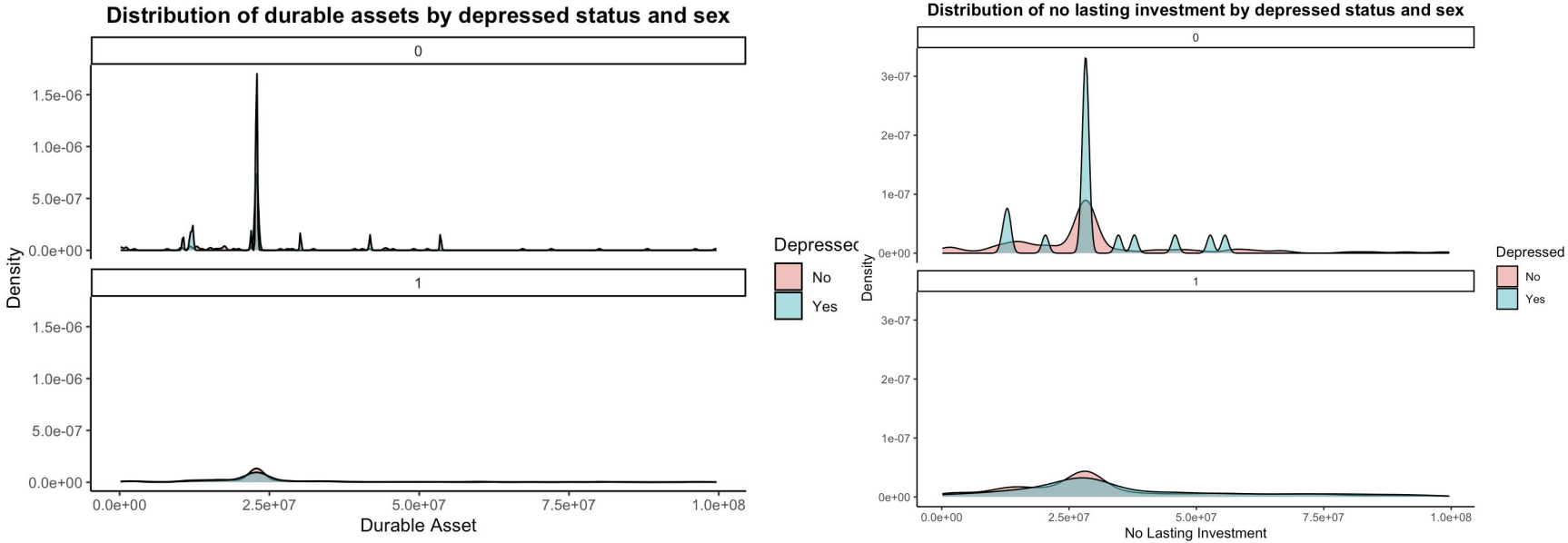
Proportion of depressed individuals by education

- Few males in the dataset
- Only 6 of the 1429 individuals have 17 or more years of education

Distributions of other Variables by Depression






Distribution of durable assets by depressed status and sex
0
1
Density
1.5e-06
1.0e-06
5.0e-07
0.0e+00
0.0e+00
2.5e+07
5.0e+07
7.5e+07
1.0e+08
Durable Asset
Depressed
No
Yes
Distribution of no lasting investment by depressed status and sex
3e-07
2e-07
1e-07
0e+00
No Lasting Investment

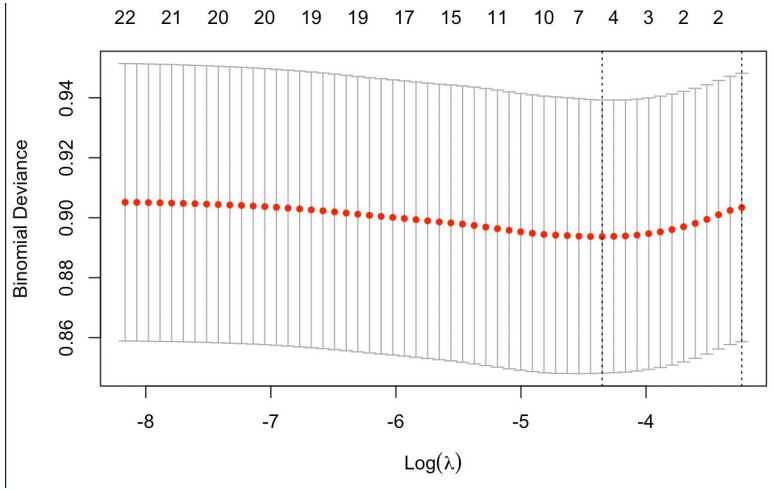
Identification of Most Informative Predictors

Stepwise Regression

- Data contained 22 different variables (aside from depression variable)
- Used backwards stepwise regression to determine most informative predictors
- Variables were excluded in each step based on AIC
- Found that these variables were: Total members, Ville ID, No lasting investment, Educational Level, Age

LASSO Regression

- Performed LASSO regression to compare to stepwise regression
- Lowest binomial deviance associated with a lambda value where 5 variables were included
- Variables obtained from LASSO were the same besides LASSO excluding ville ID and including durable asset
- Decided to use a combination of results from step and LASSO for a total of 6 variables



Variables of Interest

- **Ville ID -** Unique identifier for the community the person lives in
- **Age -** Age of individual
- **Education Level -** Grade level of highest education level reached
- **Total Members -** Total members in the person's family
- **No Lasting Investment -** Amount of money the person has that is not in long-term investments
- **Durable Asset-** Indicates the value of durable assets (such as a house or car) that the individual owns

Comparison of Regression Models

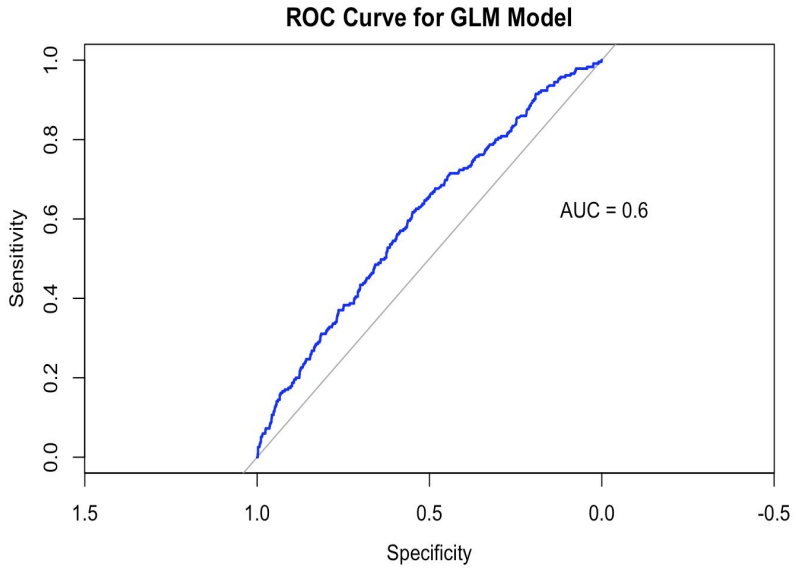
Generalized Linear Model (GLM)

- Age, education, no lasting investments were significant predictors

```
# Modeling
```{r}
#not including survey-id as it is an identifier col

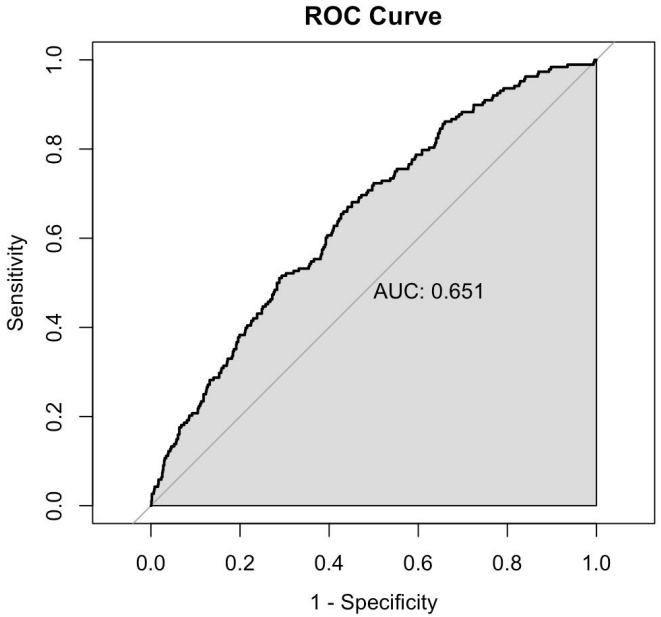
glm_model = glm(depressed ~ Age + education_level + total_members + no_lasting_investmen + durable_asset,
family=binomial(link='logit'), data=dep_cleaned)

summary(glm_model)
```
```



Generalized Linear Mixed Model (GLMM)

- Included Ville_ID as a random effect; same five fixed effects
- Did data-preprocessing including data imputation, convert variable type, scale numeric variables
- Coefficients different from GLM (i.e., odds ratios farther from 1)
- GLMM accounts for random effects and provide info about variability within and between groups




```
## GLMM Modeling
```{r}
# Train a GLMM model to predict depression
model <-  glmer(depressed ~  Age+ total_members+ durable_asset+no_lasting_investmen+education_level + (1 |
Ville_id) , data = train, family = binomial())

# Print the model summary
summary(model)
```
```

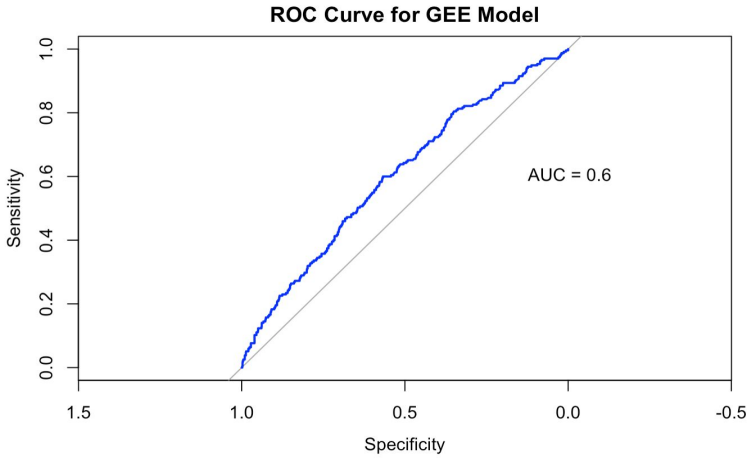
```
Correlation of Fixed Effects:
            (Intr) Age    ttl_mm drbl_s n_lst_
Age         -0.127
total_mmbrs -0.092  0.032
durable_sst -0.052 -0.034  0.045
n_lstng_nvs -0.040  0.050 -0.067 -0.023
educatn_lvl  0.102  0.400 -0.140 -0.033 -0.016
```

Generalized Estimating Equations (GEE)

- Used the “exchangeable” and “independence” correlation structures; id=Ville_id
- Minimal difference in coefficients between GEE models, same AUC (0.60)

```
# depression GEE model with exchangeable correlation structure
gee.dep1 = geeglm(depressed~Age+education_level+total_members+no_lasting_investmen,
                  family=binomial, data=depressed_data_complete,
                  id=Ville_id, corst="exchangeable")

# depression GEE model with independence correlation structure
gee.dep2 = geeglm(depressed~Age+education_level+total_members+no_lasting_investmen,
                  id=Ville_id, family=binomial, data=depressed_data_complete,
                  corst="independence")
```



Comparing Regression Model Performance

	GLM (AUC=0.60)	GLMM (AUC=0.651)	GEE (exch.) (AUC=0.60)	GEE (indep.) (AUC=0.60)
Age	1.013*	1.218*	1.014*	1.005*
Education Level	0.934*	0.785**	0.937*	1.027*
Total Family Members	1.072	1.219**	1.072	1.036
Durable Assets	1.000	1.104	1.000	1.000
No Lasting Investment	1.000*	1.133	1.000*	1.000*

Note: Values represent odds ratios (exponentiated model coefficients) *p-value < 0.05

Discussion

Summary of our Analysis

- Out of 23 variables, we identified 6 most informative variables in predicting binary depression status via stepwise and LASSO regression
- We tested three modeling approaches to predict depression: GLM (no random effects), GLMM (random effects for community), and GEE (population-averaged effects)
- The GLM and GEE models gave similar coefficients values, while the GLMM had more extreme coefficients (odds ratios farther from 1, AKA no change)
- Across all models, higher education level associated with lower odds of depression, and higher age/number of family members associated with higher odds of depression
- GLMM achieved the best predictive performance (**AUC=0.65**)

Thank YOU!

Any questions?