Final Project – EDA & Preliminary Results
Claire Yang, Saman De Silva, & Hans Elasri
2025-11-25

1. Background & Primary Research Questions

In this project, we use the 2022 wave of the China Family Panel Survey (CFPS) to study how adults' perceived income satisfaction, perceived income relative to others, and perceived social status are related to depressive symptoms, as measured by the CES-D8 scale. Our primary research questions are: (1) Are higher levels of subjective socioeconomic status associated with lower depression scores, after adjusting for demographic and household characteristics, and if so is there an interactive effect between the subjective scores? (2) How much of the variation in depression occurs between individuals versus between families, communities, and counties? and (3) To what extent to these associations remain after controlling for proxies of socioeconomic status such as education, employment status, and household composition?

2. Data Sources & Structure

Our data come from the China Family Panel Study (CFPS) 2022, a nationally representative household survey administered by the Institute of Social Science Survey at Peking University. The raw person-level dataset contains 27,001 individuals, each linked to hierarchical identifiers for families, communities, counties, and providences. After recoding special missing codes and cleaning key SES and demographic variables, removing structurally invalid IDs, our dataset reduces to 23,727 individuals nested in 9,663 families, 3,220 communities, 760 counties, and 31 provinces. For the analytic sample, we drop respondents missing any of the three SES predictors of interest, missing depression scores, or under age 18. This produces a final dataset of 14,575 adults nested in 8,123 families, 2,856 communities, 702 counties, and all 31 providences. We believe that missingness is largely missing completely at random based on the user guide, but more work is needed to verify this assumption. All variables come from survey interviers (no administrative data or observational data) and we created no synthetic measures other than centering continuous covariates and contructing binary indicators. The major data decisions made were (1) treating CFPS special codes as true missingness, (2) dropping households with invalid/missing geographic IDs, and (3) restricting to respondents with complete SES and outcome information. These steps ensure comparability across respondents and preserve the multilevel structure needed for our 4 level model.

3. EDA Plots

See Appendix A.

4. & 5. Variance Decomposition, Data Trends

From our visualizations (see: Appendix A), we observe that there appears to be a negative relationship between CESD8 Depression scores (where higher indicates more depressive symptoms) and perceived income status, perceived social status, and personal income satisfaction

(where higher indicates higher status/satisfcation). This gives justification to our original question of how these three factors are related to mental health outcomes, and shows that from a completely pooled level, there is a relationship, and in particular it appears to be strongest by income satisfaction. Our boxplot for CESD Score for 20 Randomly Selected County Clusters also gives justification for the need for mutlilevel models, as we see that even on the county level, there is a large amount of variation (and also different sample sizes) in the CESD8 scores (and this is true for the family and community level, although the plots are not shown).

In terms of variance, we observe that the residual variance by far exceeds the variance from any other clusters, indicating that within individual variance dominates, taking up 80% of the total variance (Appendix B). For the other structures, family is the most important clustering structure in the data, making up around 15% of the total variance. We observe that around 20% of the total variance can be attributed to any grouping at all (family, community, or county), giving justification for the use of a hierarchical models (Appendix B). A possible sub-question we could look into is whether true income position or perceived income position is a better predictor for CESD8 scores.

6. Mathmatical Model

Our model has four clustered levels: $i =$ individuals, $h =$ households/families, $c =$ communities, $k =$ counties. Our primary outcome variable is a (0-24) depression score: Y_{ihck}.
There are three primary predictors of interest (all grand mean centered Likert scores):

- Income Satisfaction: $IS^*_{ij} = income\_satis_{ij} - \overline{income\_satis}$
- Perception of Relative Income: $IR^*_{ij} = income\_rel_{ij} - \overline{income\_rel}$
- Perception of Relative Social Status: $IS^*_{ij} = status\_relij - \overline{status\_rel}$
- Let $\boldsymbol{X}_{ihck}$ be a column vector of control variables, mostly level 1 (e.g. age, gender, ethnicity, rural/urban, CCP party member, retired, pension, etc)
- Let $\boldsymbol{\delta}$ is the corresponding vector of fixed effect coefficients

The multilevel model is then as follows:
Level 1 (Individual):

$$Y_{ihck} = \beta_{0hck} + \beta_1 IS^*_{ihck} + \beta_2 IR^*_{ihck} + \beta_3 SR^*_{ihck} + \mathbf{X}^T_{ihck}\boldsymbol{\delta} + \sum_{p=2}^{P} \beta_p^{(prov)} D_{p(ihck)} + \varepsilon_{ihck}, \quad \varepsilon_{ihck} \sim \mathcal{N}(0, \sigma^2)$$

where $D_{p(ihck)}$ is a dummy indicating that the individaul is in providence p.
Level 2 (Family):

$$\beta_{0hck} = \gamma_{00ck} + u_{0hck}, \quad u_{0hck} \sim \mathcal{N}(n, \tau^2_{family})$$

Level 3 (Community):

$$\beta_{00ck} = \gamma_{000k} + v_{0ck}, \quad v_{0ck} \sim \mathcal{N}(n, \tau^2_{community})$$

Level 4 (County):

$$\beta_{000k} = \gamma_{0000} + w_{0k}, \quad w_{0k} \sim \mathcal{N}(n, \tau^2_{county})$$

Combined:

$$Y_{ihck} = \gamma_{0000} + \beta_1 \, \mathrm{IS}^*_{ihck} + \beta_2 \, \mathrm{IR}^*_{ihck} + \beta_3 \, \mathrm{SR}^*_{ihck} + \mathbf{X}^\top_{ihck}\boldsymbol{\delta} + \sum_{p=2}^{P} \beta_p^{(\mathrm{prov})} D_{p(ihck)} + w_{0k} + v_{0ck} + u_{0hck} + \varepsilon_{ihck}.$$

7. Initial Findings

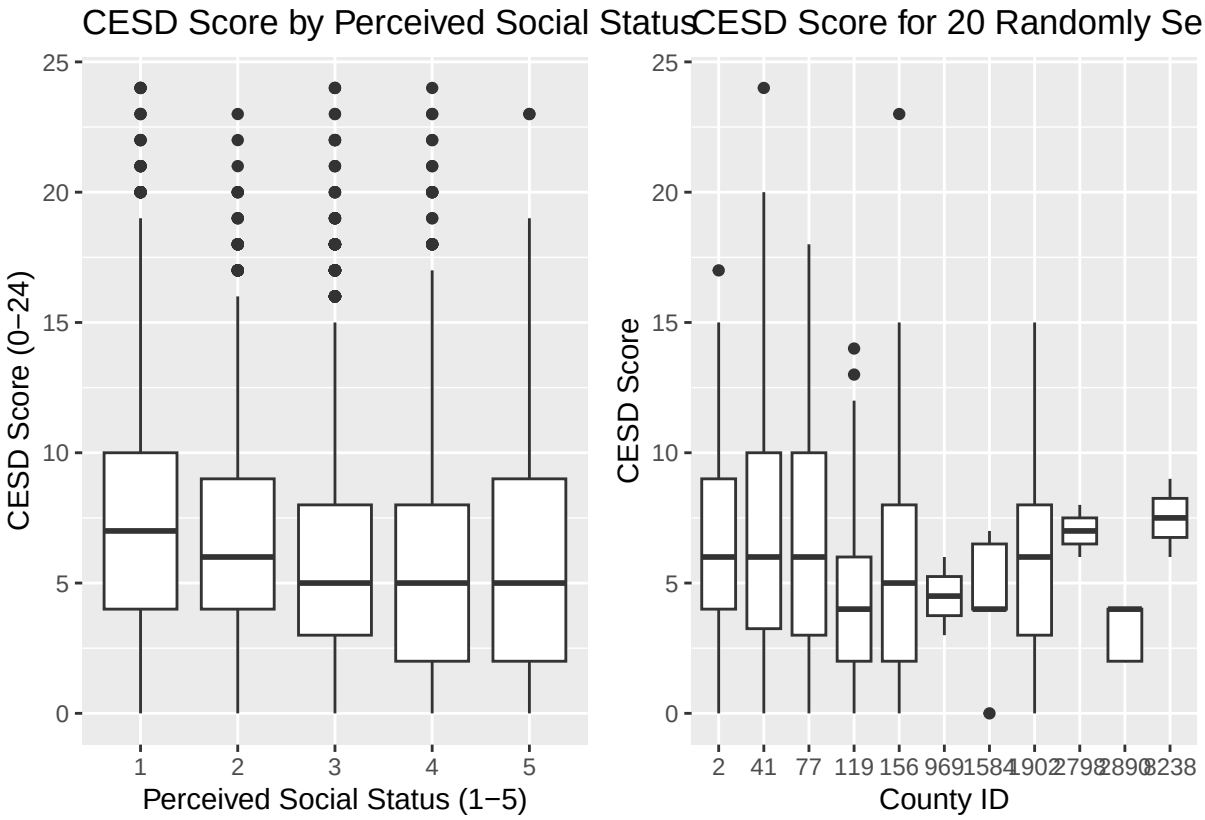
(References for these findings can be found Appendix Items B and D.) The cross-classified model estimates how perceived socioeconomic status measures relate to depression as measured by the CES-D8 survey among Chinese adults while controlling for demographic stats, household characteristics, and providence fixed effects. After adjusting for the covariates, all three predictors of relative socioeconomic measures are statistically significant at the 1% level ($p < 0.001$). Income satisfaction shows the strongest association with depression: a one unit increase (relative to the sample mean) predicts a 0.53 point decrease in CES-D8 scores (this corresponds loosely to "half of a symptom"). Perceived relative income ($\beta = -0.17$) and perceived social status ($\beta = -0.33$) are also economically significant. Taken together, the findings suggest that individuals who feel better off or higher status than their peers tend to experience fewer depressive symptoms. Random-effects estimates show that most of the clustering occurs at the family level (var ≈ 2.31) with smaller but still meaningful clustering happening at the community (var ≈ 0.29) and county (var ≈ 0.20) levels which suggests heterogeneity in depressive symptoms across families and communities. The results seem to indicate that subjective socioeconomic well being is a robust, consistent predictor of mental health in the CFPS 2022 data, but more work is needed to determine if these effects have any interactive effects between each other.

8. Next Steps

Our next steps are to refine the model and assess robustness. We plan to test interactions among the three subjective SES measures, examine subgroup differences (e.g., by gender or age), and compare our current four-level random-intercept model with alternatives such as cross-classified structures or limited random slopes. A remaining concern will be if adding complexity to the model will make estimates unstable or slow to fit. We would appreciate feedback on which extensions are most appropriate for this type of dataset and leverage the 4 levels in a more comprehensive way.

Appendix

Item A: Visualizations



Item B: Model Summary

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: cesd8 ~ 1 + (1 | family_id) + (1 | community_id) + (1 | county_id)
##    Data: dat
##
## REML criterion at convergence: 82173.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0855 -0.7027 -0.1183  0.5645  4.5744
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  family_id    (Intercept)  2.3938  1.5472
##  community_id (Intercept)  0.3499  0.5915
##  county_id    (Intercept)  0.5890  0.7674
##  Residual                 13.7571  3.7091
## Number of obs: 14575, groups:
## family_id, 8123; community_id, 2856; county_id, 702
##
## Fixed effects:
```

```
##             Estimate Std. Error t value
## (Intercept)  5.87116    0.06417   91.49
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00232387 (tol = 0.002, component 1)
```

Item C: ICC/Variance Analysis

```
## [1] 0.8049889
```

```
## [1] 0.1400719
```

```
## [1] 0.1950111
```

Item D: Model Output

Table 1: Effects of Subjective SES on Depression (Multilevel Model)

Term	Estimate	SE	t stat
Income satisfaction (centered)	-0.527	0.033	-16.192
Relative income (centered)	-0.165	0.041	-4.064
Perceived social status (centered)	-0.331	0.040	-8.296
Age (centered)	0.021	0.003	6.579
Male	-0.900	0.094	-9.555
Family SD	1.522	NA	NA
Community SD	0.538	NA	NA
County SD	0.452	NA	NA