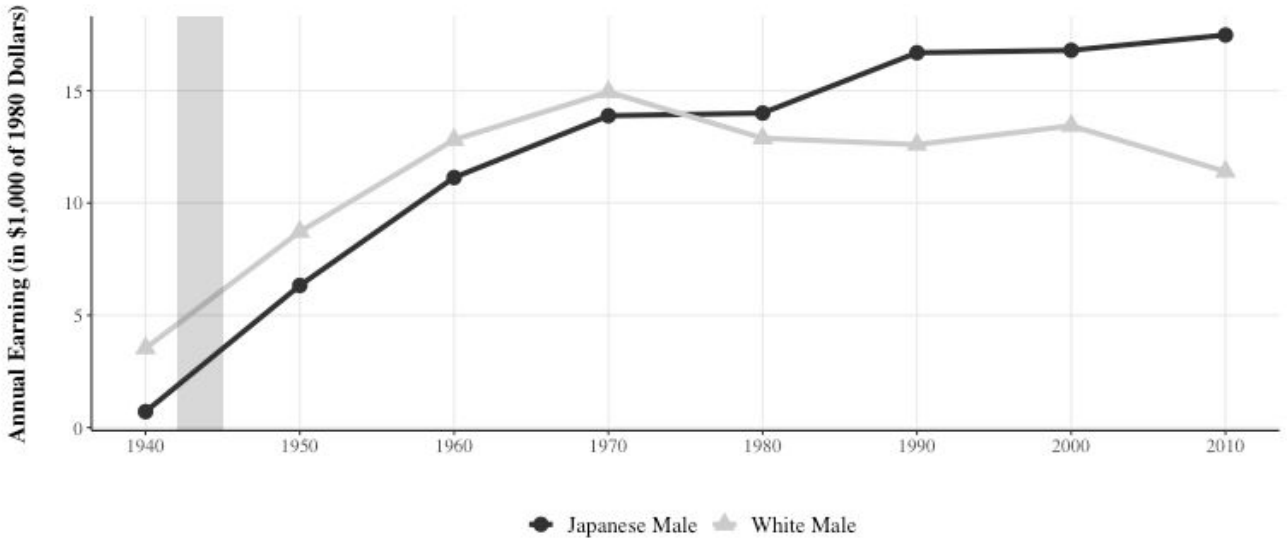
Data Science Capstone:

The Condition of Internment Camps on Educational Attainment

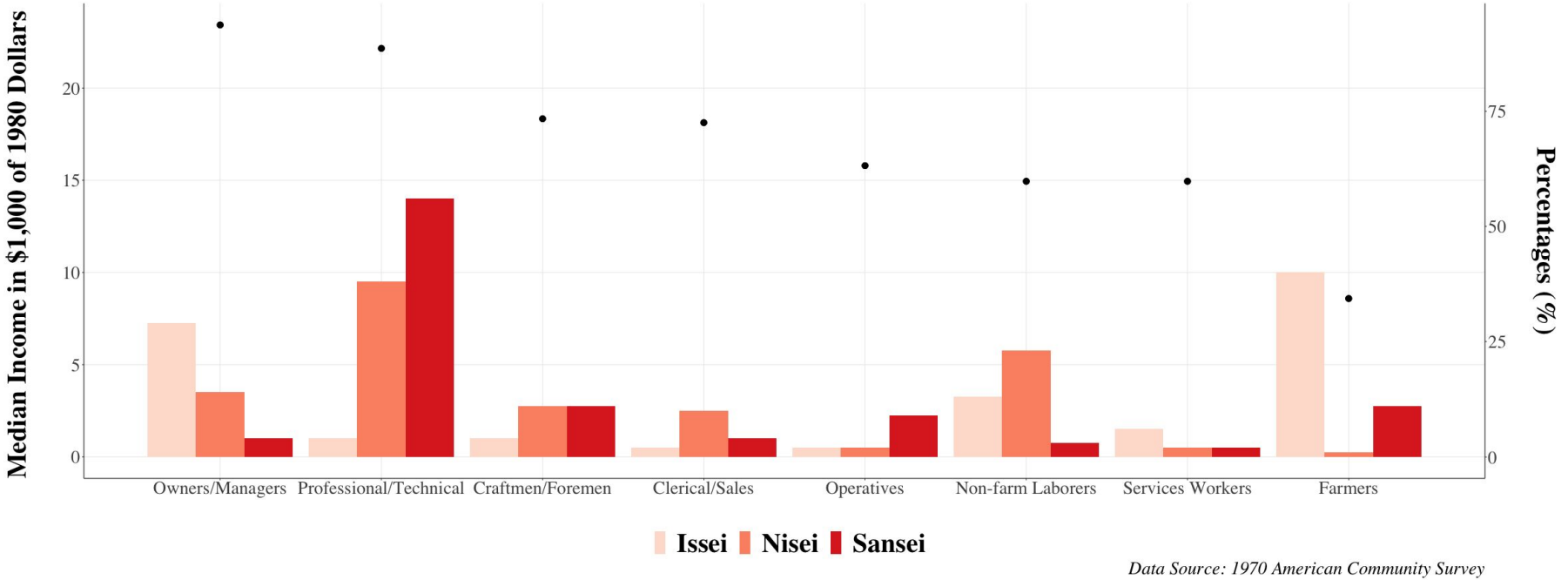
--Miranda Miao

Thesis: The Evolution of the Japanese-White Real Earnings Gap

Figure I: Real Earnings of Japanese and White Men, Median

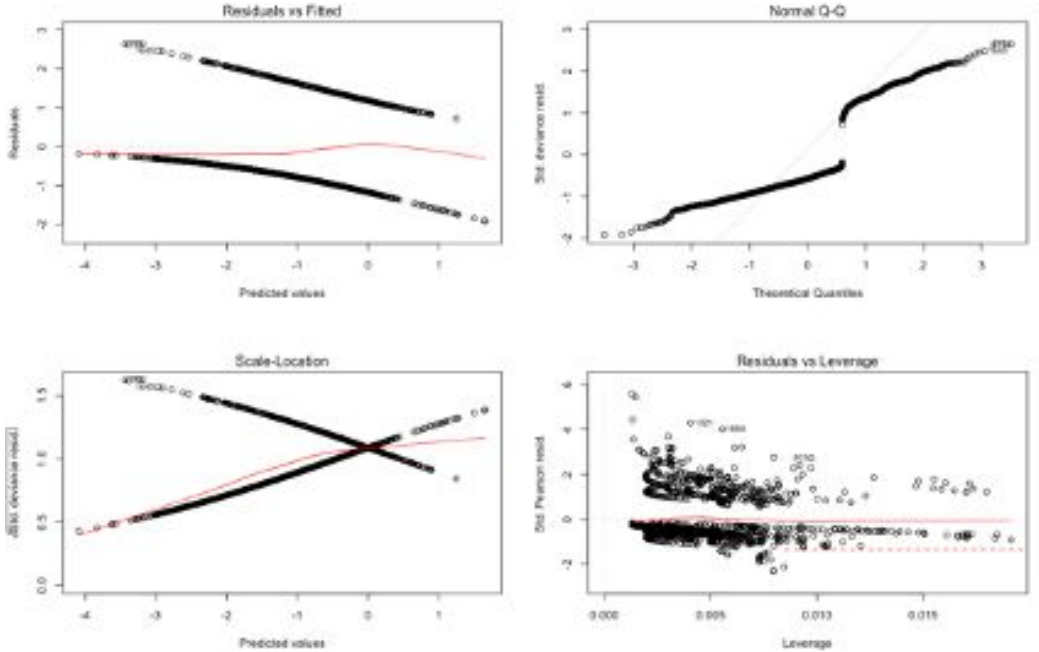


Data Source: American Community Survey



Data Source: 1970 American Community Survey

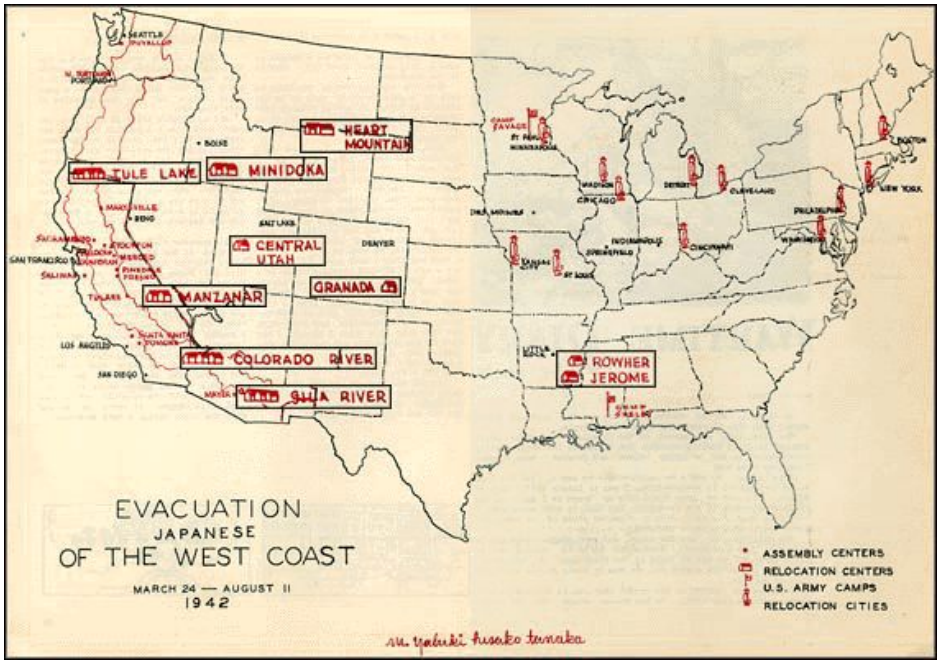
318 Project: Correlation of Internment on College Education Attainment



- Logistics Regression
- Internment had statistically significant and negative correlation on education using the 0.05 threshold
- The Odds ratio for internment is 0.766

Japanese Internment

- Japanese Internment: 1942 - 1945
  - 110,000 Interned; 66,000 US Citizens; 1862 Deaths
  - Forced evacuation; Property damage
  - Regional Randomized Selection of camp site
- Violence and force at Camp
  - Deaths, fights, lack of privacy, and sexual assault
  - Definition of violence, force, and strike



Research Questions

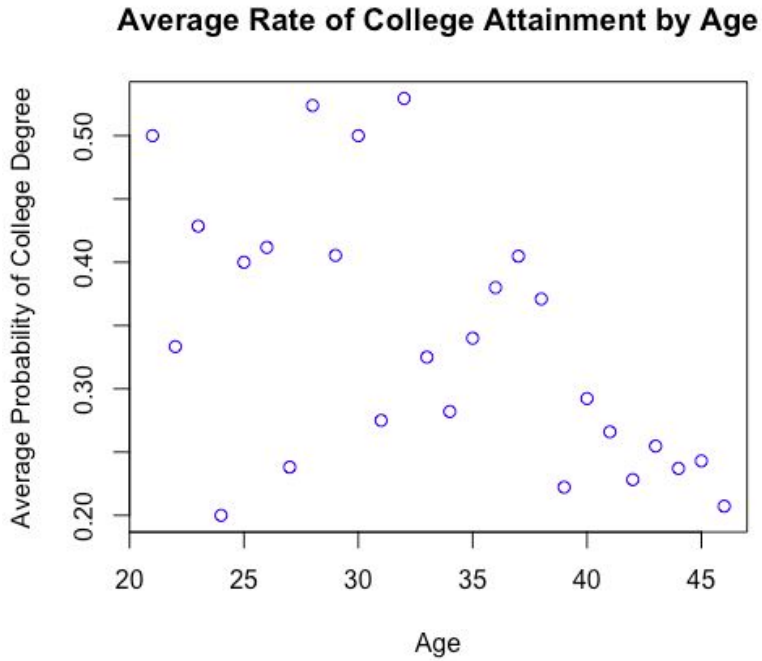
Research Question: Given that if someone was interned, their internment camp site assignment is random conditional on geography, how does the condition of the assigned camp affect one’s college educational attainment?

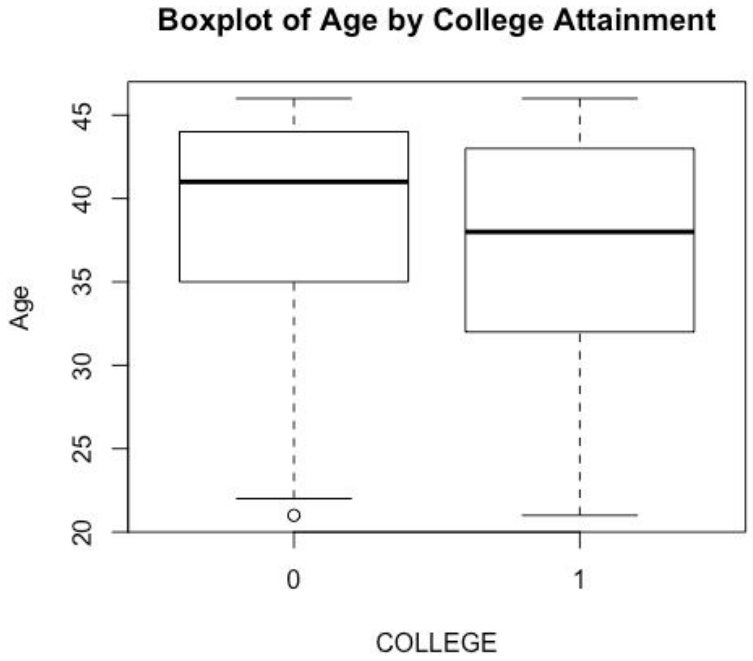
Data Exploration and Method

Data

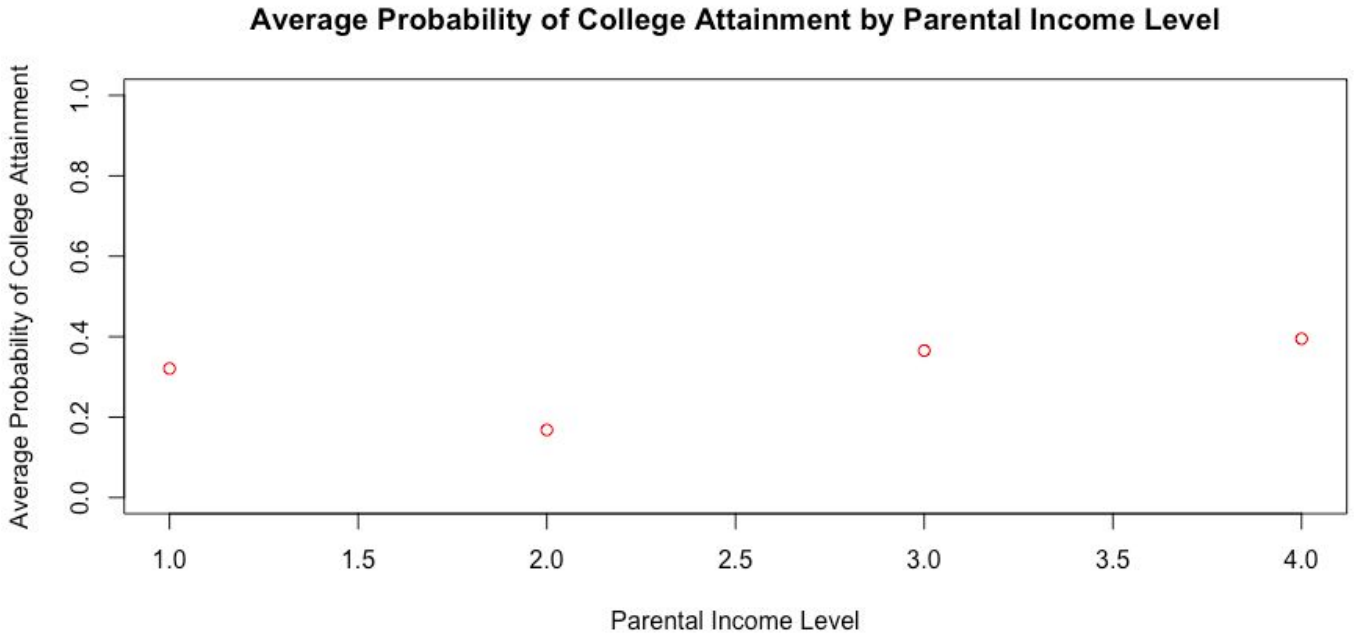
- Japanese American Research Project
  - Socio-historical survey randomized by county across the united States
  - Randomly selected Issei (first-generation Japanese Americans)
  - Traced their children (Nisei) and grandchildren (Sansei)
- Useful Information
  - Internment status, education attainment, family ID, internment site, and property compensation, etc.
- Only Nisei and those interned and Age < 47

Modeling Assumption - Age





Model Assumption - Parent Income Level



Modeling - Selecting Predictor Subset

Full Model Before Predictor Selection:

Expected Log Odds of Attending College <- $\beta_0 + \beta_1$*Sex + β_2*Married + β_3 *Age + β_4*Age_Squared + β_5*Property_Compensation + β_6 *Parent_Income2 + β_7 *Parent_Income3 + β_8 * Parent_Income4 + β_9* Violence + β_{10} * Strike + β_{11}* Force

AIC Criterion Model: step(k=2)

Expected Log Odds of Attending College <- $\beta_0 + \beta_1$*Sex +β_2*Married + β_3 *Age + β_4*Property_Compensation + β_5 *Parent_Income2 + β_6 *Parent_Income3 + β_7* Parent_Income4 + β_8* Strike + β_9* Force

AUC Criterion Model: AUCRF Package

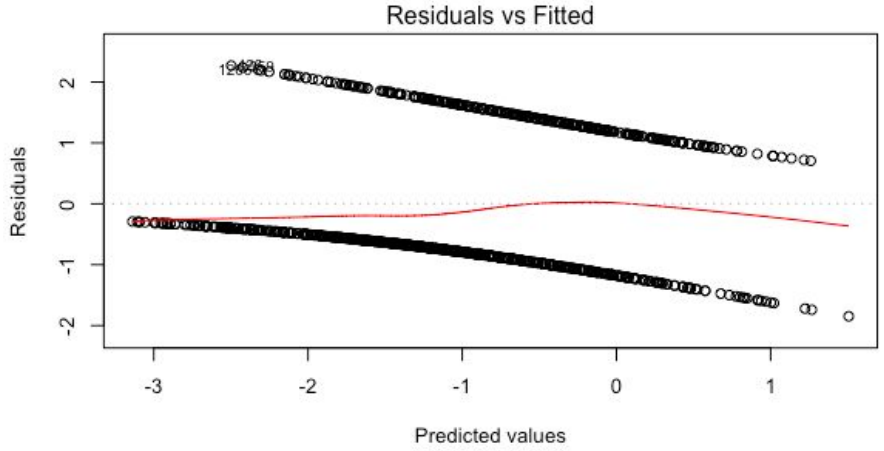
Expected Log Odds of Attending College <- $\beta_0 + \beta_1$*Sex + β_2 *Age + β_3*Age_Square + β_4*Parent_Income2 + β_5*Parent_Income3 + β_6* Parent_Income4 +β_7* Strike

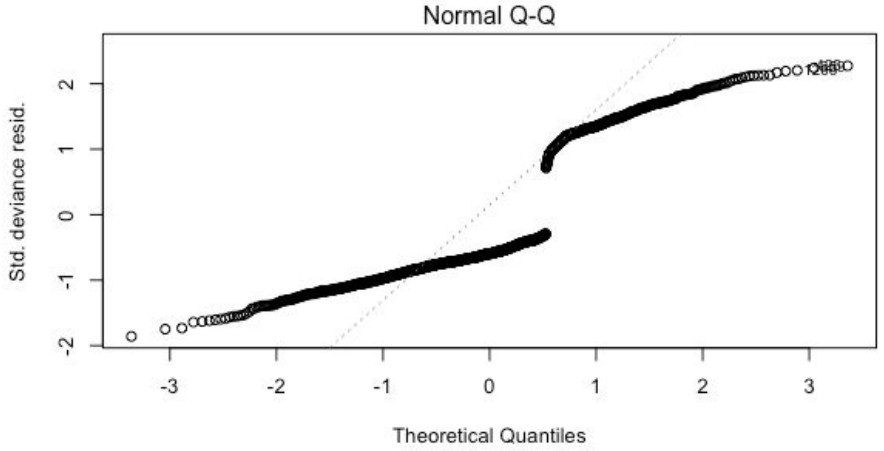
Modeling - Cross Validation

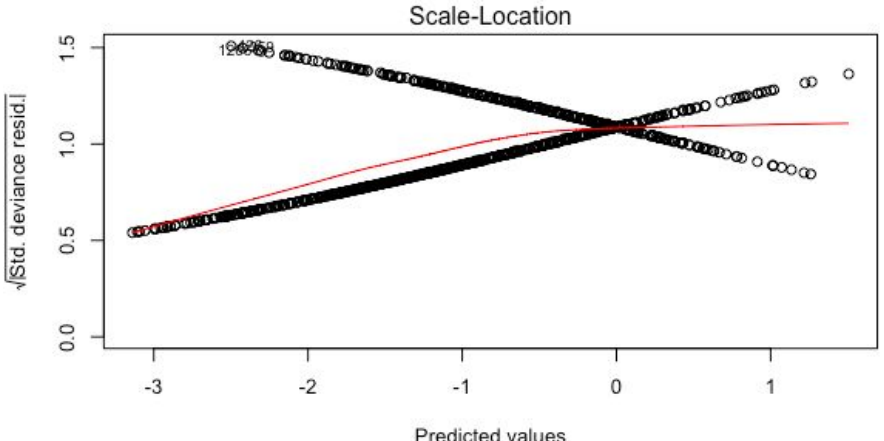
Implemented five-fold cross validation to determine predictive power

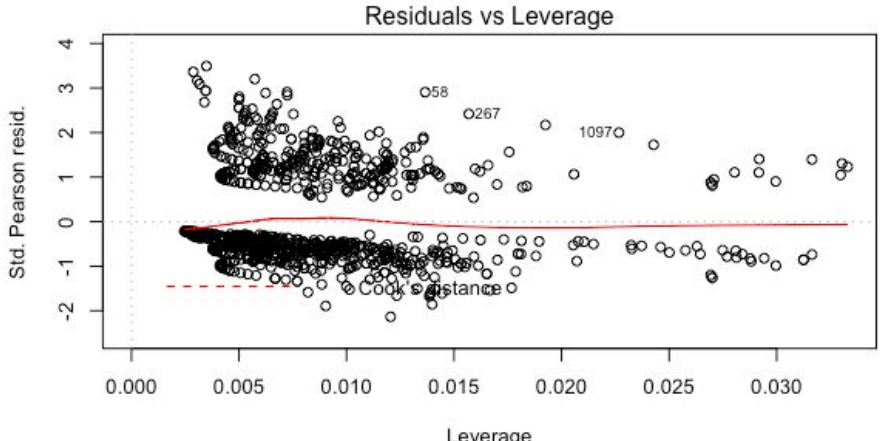
- Criterion: Average AUC over 5 validations
- AIC: Average AUC is 0.713
- AUCRF: Average AUC is 0.706

Regression Diagnostics


Residuals vs Fitted
Residuals
Predicted values


Normal Q-Q
Std. deviance resid.
Theoretical Quantiles


Scale-Location
√|Std. deviance resid.|
Predicted values


Residuals vs Leverage
Std. Pearson resid.
58
267
1097
Cook's distance
Leverage

Identifying Influential Outliers

- Using the delta deviance of coefficients to identify influential outliers
- Using the distribution of delta deviance at each data points as benchmark
    - 58 is an influential due to Strike; 267 is an influential outlier due to Parent_Inc3

```
> outliers
     (Intercept)         SEX         AGE PROPERTY_COMPEN     MARRIED  PARENT_INC2  PARENT_INC3  PARENT_INC4     STRIKE        FORCE
58    0.012892375 -0.01471394 -0.01130686     -0.01021479 -0.01342982 -0.009447340  0.005912919  0.005549265 0.01272637  0.009744724
267  -0.007867528 -0.03049179  0.03806342     -0.02429944 -0.02141389 -0.007083395 -0.020606910 -0.011279633 0.02986141 -0.066320751
1097 -0.005231070  0.03474164  0.01519625     -0.07069692  0.01505501 -0.006373795  0.004550003 -0.010293221 0.06520520 -0.031160414
> threshold
            (Intercept)         SEX         AGE PROPERTY_COMPEN     MARRIED PARENT_INC2 PARENT_INC3 PARENT_INC4      STRIKE        FORCE
Q1-1.5*IQR -0.04923637 -0.09008385 -0.06061404     -0.07117123 -0.04578402 -0.02001413 -0.02028241 -0.02147511 -0.09264240 -0.07313533
Q3+1.5*IQR  0.04894225  0.09390238  0.05763285      0.08010332  0.04724607  0.01697499  0.02114824  0.02179442  0.08402614  0.06364415
```

Likelihood Ratio Test:
For Nested Model Comparison

- Null Hypothesis: β For Parent Income Level 2 = 0
  1-pchisq(((1396.0 -1395.7),1) → 0.58
- Null Hypothesis: β for Force = β for Strike = 0
  1-pchisq((1399.6-1396.0),2) → 0.16
- Removal of all three variables

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.02640    0.41649   0.063  0.94947
SEX           0.90943    0.13462   6.756 1.42e-11 ***
AGE          -0.05182    0.01086  -4.771 1.83e-06 ***
PROPERTY_COMPEN  0.39691    0.14773   2.687  0.00721 **
MARRIED      -0.57995    0.17800  -3.258  0.00112 **
PARENT_INC3   1.24540    0.15956   7.805 5.94e-15 ***
PARENT_INC4   1.48951    0.22333   6.670 2.57e-11 ***
STRIKE       -0.19614    0.13765  -1.425  0.15418
FORCE        -0.17984    0.15317  -1.174  0.24035
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Likelihood Ratio Test: For Final Model Fit

Final Model:

Expected Log Odds of Attending College <-
$\beta_0 + \beta_1$*Sex +β_2*Married + β_3 *Age +
β_4*Property_Compensation +
β_5*Parent_Income3 + β_6* Parent_Income4

1-pchisq(1399.6 ,1277) → 0.009

```
glm(formula = COLLEGE ~ SEX + AGE + PROPERTY_COMPEN + MARRIED +
    PARENT_INC3 + PARENT_INC4, family = binomial(link = "logit"),
    data = jarp.clean)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8301  -0.8269  -0.5951   1.1470   2.2930

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.06728    0.41279  -0.163  0.87052
SEX             0.90102    0.13415   6.716 1.86e-11 ***
AGE            -0.05228    0.01085  -4.820 1.44e-06 ***
PROPERTY_COMPEN  0.39147    0.14730   2.658  0.00787 **
MARRIED        -0.57786    0.17783  -3.250  0.00116 **
PARENT_INC3     1.25177    0.15918   7.864 3.72e-15 ***
PARENT_INC4     1.49679    0.22323   6.705 2.01e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1563.2  on 1283  degrees of freedom
Residual deviance: 1399.6  on 1277  degrees of freedom
AIC: 1413.6
```

Data Ethics

- Data is anonymous and does not contain location information
- Some might argue that violent treatment doesn’t matter
- It is not statistically significant, but had negative effects!
- Education could be resilient but mental health might not be