Lab-4-Markdown

Michael Sullivan

2025-03-07

```
# Packages-------
library(lme4)
library(readxl)
library(ggplot2)
library(plm)
library(dplyr)
library(performance)

# Setup----------
data <- read_excel("GACTT_RESULTS_ANONYMIZED_HW4.xlsx")

class(data$most_willing_for_cup)
```

```
## [1] "numeric"
```

```
class(data$hh_income)
```

```
## [1] "numeric"
```

```
class(data$age_cat)
```

```
## [1] "character"
```

```
class(data$state)
```

```
## [1] "character"
```

```
data$age_cat <- factor(data$age_cat)
data$state <- factor(data$state)

table(data$most_willing_for_cup)
```

```
##
##   1   3   5   7   9  13  17  25
##   6  42 177 357 481 359 194 368
```

```r
table(data$age_cat)
```

```
##
##   <35 35-44 45-54 55-64   65<
##  1296   466   127    66    29
```

```r
table(data$state)
```

```
##
##  AZ  CA  CO  FL  GA  IL  MD  MI  MN  NC  NY  OH  OR  PA  TX  UT  VA  WA
##  51 399 109  80  58 103  51  63  53  74 209  80  65 107 179  61  82 160
```

```r
sum(is.na(data$most_willing_for_cup))
```

```
## [1] 0
```

```r
sum(is.na(data$hh_income))
```

```
## [1] 0
```

```r
sum(is.na(data$age_cat))
```

```
## [1] 0
```

```r
sum(is.na(data$state))
```

```
## [1] 0
```

```r
# The data appears to already be clean for the purposes we need it for
```

Question 1: Normal OLS Model

Part 1. Fit OLS Model

```r
# Question 1-----
model <- lm(most_willing_for_cup ~ hh_income, data = data)
summary(model)
```

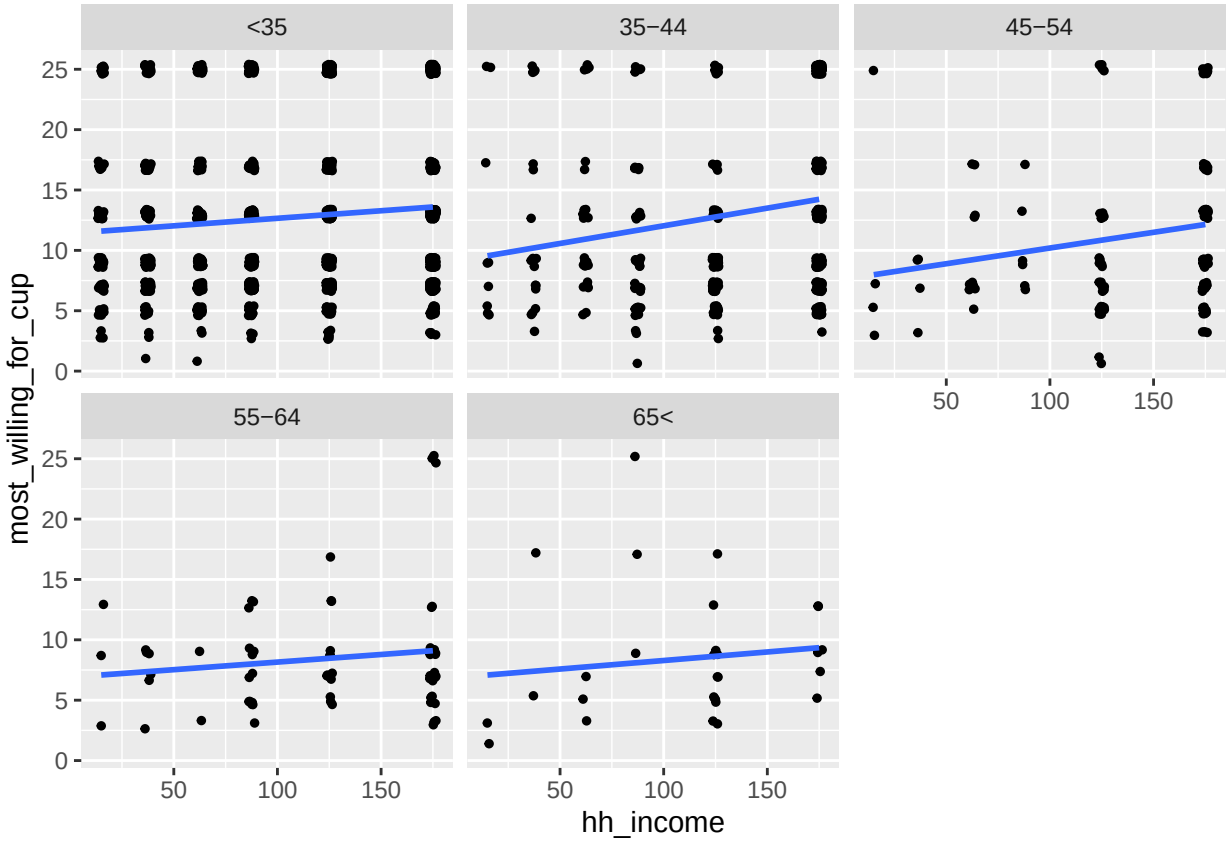
```
##
## Call:
## lm(formula = most_willing_for_cup ~ hh_income, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.657  -5.068  -2.676   3.558  14.070
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.694751   0.381791   28.01  < 2e-16 ***
## hh_income    0.015697   0.002875    5.46 5.37e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.789 on 1982 degrees of freedom
## Multiple R-squared:  0.01482,    Adjusted R-squared:  0.01432
## F-statistic: 29.81 on 1 and 1982 DF,  p-value: 5.373e-08
```

Part 2. Interpret Coefficient

The coefficient of household income on most_willing_for_cup is 0.0157 with a standard error of 0.0029. This indicates that the relationship is statistically significant, and the specific interpretation is that an increase of household income by one thousand dollars is associated with an increase of $0.0157 on the most someone is willing to pay for a cup. Additionally, the adjusted R^2 score of 0.0143 is very low, indicating that income explains only a small portion of the variation in willingness to pay.

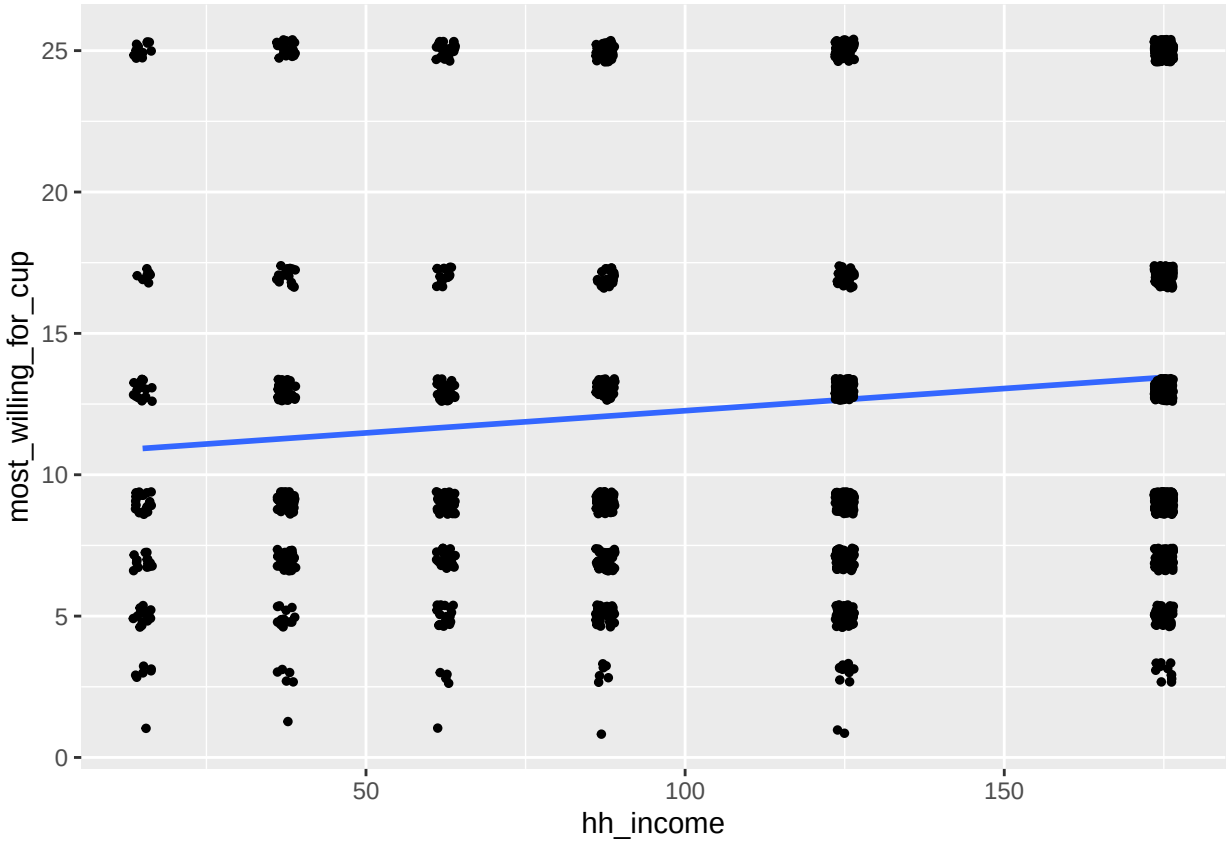
```
ggplot(data, aes(x = hh_income, y = most_willing_for_cup)) +
  geom_jitter(width = 1.5, height = 0.4, size = 1) +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~ age_cat)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(data, aes(x = hh_income, y = most_willing_for_cup)) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_jitter(width = 1.5, height = 0.4, size = 1)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Part 3. Plot Relationship

The above plots show the trend by and across age groups. They show that the relationship between household income and most willing to pay varies meaningfully by age group. The variability in trends across age groups suggests that a pooled OLS model, which assumes a single relationship for all individuals, may not be appropriate.

Question 2: Fixed Effects Models

Part 1. One-Way Fixed Effects Model

```
fe_one_way <- plm(most_willing_for_cup ~ hh_income,
                  data = data,
                  index = "age_cat",
                  model = "within")
summary(fe_one_way)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = most_willing_for_cup ~ hh_income, data = data,
##     model = "within", index = "age_cat")
##
## Unbalanced Panel: n = 5, T = 29-1296, N = 1984
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -11.3533  -4.8316  -2.3022   3.2155  16.9423
##
## Coefficients:
##            Estimate Std. Error t-value  Pr(>|t|)
## hh_income 0.0163570  0.0029224  5.5971 2.484e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    90596
## Residual Sum of Squares: 89184
## R-Squared:      0.015591
## Adj. R-Squared: 0.013103
## F-statistic: 31.3273 on 1 and 1978 DF, p-value: 2.4838e-08
```

Part 2. Two-Way Fixed Effects Model

```
age_cat_list <- unique(data$age_cat)
state_list <- unique(data$state)

agg_data <- data.frame(
  age_cat = rep(age_cat_list, length(state_list)),
  state = rep(state_list, each = length(age_cat_list))
)

agg_data <- agg_data %>%
  left_join(
    data %>%
      group_by(age_cat, state) %>%
      summarise(
        most_willing_for_cup = mean(most_willing_for_cup),
```

```
        hh_income = mean(hh_income),
        .groups = "drop"
      ),
    by = c("age_cat", "state")
  )

fe_two_way <- plm(most_willing_for_cup ~ hh_income,
                  data = agg_data,
                  index = c("state", "age_cat"),
                  model = "within",
                  effect = "twoways")
summary(fe_two_way)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = most_willing_for_cup ~ hh_income, data = agg_data,
##     effect = "twoways", model = "within", index = c("state",
##         "age_cat"))
##
## Unbalanced Panel: n = 18, T = 4-5, N = 83
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -5.47998 -0.91408 -0.10374  1.15159  5.27059
##
## Coefficients:
##           Estimate Std. Error t-value Pr(>|t|)
## hh_income 0.039650   0.015207  2.6073   0.0115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    414.8
## Residual Sum of Squares: 372.59
## R-Squared:      0.10177
## Adj. R-Squared: -0.22758
## F-statistic: 6.79795 on 1 and 60 DF, p-value: 0.011498
```

Part 3. Compare Results

Once we include fixed effects the coefficient for household income increases. When we include age group as a fixed effect is increases to 0.0164, when we include both age group and state it increases to 0.0397.

These changes indicate that after controlling for age and state we find a more sizeable effect of household income on willingness to pay. They therefore indicate that the age groups and states for which respondents have higher willingness to pay tend to also be age groups and states which have lower incomes, which would dampen the original observed relationship between income and willingness to pay. Once we hold age and state constant we see that income has an even stronger relationship because the effect of these counter-acting relationships is removed.

We also observe here an R^2 of 0.1018 for the two-way fixed effects model, which is significantly higher than the R^2 of the pooled OLS model (0.0143), indicating that the fixed effects model has greater explanatory power.

Question 3: Mixed Effects Models

Part 1. Random Intercept Model

```
data <- data %>%
  mutate(hh_income_scaled = scale(hh_income))

random_intercept_model <- lmer(most_willing_for_cup ~ hh_income + (1 | age_cat), data = data)
summary(random_intercept_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: most_willing_for_cup ~ hh_income + (1 | age_cat)
##    Data: data
##
## REML criterion at convergence: 13207.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6837 -0.7178 -0.3751  0.4849  2.4306
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  age_cat  (Intercept)  4.272   2.067
##  Residual             45.089   6.715
## Number of obs: 1984, groups:  age_cat, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  8.95375    1.04269   8.587
## hh_income    0.01643    0.00292   5.627
##
## Correlation of Fixed Effects:
##           (Intr)
## hh_income -0.353
```

```
# Producing scaled version because the random-slope model failed to converge when unscaled
random_intercept_model_scaled <- lmer(most_willing_for_cup ~ hh_income_scaled + (1 | age_cat), data = da
summary(random_intercept_model_scaled)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: most_willing_for_cup ~ hh_income_scaled + (1 | age_cat)
##    Data: data
##
## REML criterion at convergence: 13199.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6837 -0.7178 -0.3751  0.4849  2.4306
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
```

```
##  age_cat  (Intercept)  4.272   2.067
##  Residual             45.089   6.715
## Number of obs: 1984, groups:  age_cat, 5
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)       10.9543     0.9757  11.227
## hh_income_scaled   0.8713     0.1549   5.627
##
## Correlation of Fixed Effects:
##             (Intr)
## hh_ncm_scld -0.013
```

Part 2. Random Slope Model

```
# Model with scaled income
random_slope_model_scaled <- lmer(most_willing_for_cup ~ hh_income_scaled + (hh_income_scaled | age_cat)
summary(random_slope_model_scaled)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: most_willing_for_cup ~ hh_income_scaled + (hh_income_scaled |
##     age_cat)
##    Data: data
##
## REML criterion at convergence: 13197.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6555 -0.7513 -0.3656  0.4987  2.4415
##
## Random effects:
##  Groups   Name             Variance Std.Dev. Corr
##  age_cat  (Intercept)       4.0771  2.0192
##           hh_income_scaled  0.1485  0.3853   0.04
##  Residual                  45.0104  6.7090
## Number of obs: 1984, groups:  age_cat, 5
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)       10.9093     0.9558  11.413
## hh_income_scaled   1.0234     0.2839   3.604
##
## Correlation of Fixed Effects:
##             (Intr)
## hh_ncm_scld 0.007
```

Part 3. Compare Random Intercept To One-Way Fixed Effects

Note that the second random intercept model above and the random slope model have normalized versions of the income variable. The random slope model failed to converge with the raw data so I've normalized it and done so for the random-intercept model as well to compare them.

The one-way fixed effects model controls for unobserved heterogeneity at the age group level, focusing on within-group variation. The coefficient for hh_income in this model (0.0163) reflects the within-age-group relationship between household income and willingness.

The random intercept model allows each age group to have its own baseline level of willingness while assuming that these intercepts come from a common distribution. The estimated coefficient for hh_income in this model (0.0164) is similar to that of the fixed effects model which is what we would expect because both coefficients represent the association of hh_income with willingness to pay when you remove the difference across groups. As a result both are reasonable choices if we are specifically interested in the within-group relationships.

Part 4. Compare Random Intercept To Random Slope

As discussed above, in the random intercept model, the coefficient for hh_income represents the mean effect of income on willingness across all age groups, with each group having distinct intercepts. The coefficient in the scaled version of the model is 0.8713 which indicates that an increase in income of 0.8713 of a standard deviation is associated with a one dollar increase in willingness to pay.

In the scaled random slope model, the coefficient for hh_income_scaled (1.023) also represents the mean effect, but in this case the model also accounts for variations in this effect across age groups while assuming the same baseline willingness to pay. The different results indicate that when we assume the same baseline, the relationship between income and willingness to pay is slightly higher on average.

The variance of the random slope (0.1485) indicates how much this relationship changes between age groups. The correlation (0.04) between the random intercept and the random slope suggests there is a very weak relationship between the baseline willingness and the effect of income across age groups.

I believe that the random slope model is better justified because it has a more realistic assumption. For the random intercept model to be preferable we would assume that the relationship between income and willingness to pay is static across all age ranges. Theoretically I find this difficult to justify as an individuals relationship to money typically varies widely across their lifetime. Conversely, the random slope model only assumes that at the intercept, when someone's income is at 0, they will share the same initial willingness to pay. While this is also probably not literally and universally true, the idea that across ages if someone has little to no income they will share similar minimal willingness to pay seems much more plausible.

Question 4: More Mixed Effects Model

```
data <- data %>%
  mutate(hh_income_scaled = scale(hh_income))

random_intercept_slope_model <- lmer(most_willing_for_cup ~ hh_income_scaled + (1 + hh_income_scaled | a
summary(random_intercept_slope_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: most_willing_for_cup ~ hh_income_scaled + (1 + hh_income_scaled |
##     age_cat)
##    Data: data
##
## REML criterion at convergence: 13197.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6555 -0.7513 -0.3656  0.4987  2.4415
##
## Random effects:
##  Groups   Name             Variance Std.Dev. Corr
##  age_cat  (Intercept)       4.0771  2.0192
##           hh_income_scaled  0.1485  0.3853   0.04
##  Residual                  45.0104  6.7090
## Number of obs: 1984, groups:  age_cat, 5
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)       10.9093     0.9558  11.413
## hh_income_scaled   1.0234     0.2839   3.604
##
## Correlation of Fixed Effects:
##             (Intr)
## hh_ncm_scld 0.007
```

```
compute_plm_aic <- function(plm_model) {
  logLik_value <- as.numeric(logLik(plm_model))  # Extract log-likelihood
  k <- length(coef(plm_model))  # Number of estimated parameters
  AIC_value <- -2 * logLik_value + 2 * k  # AIC formula
  return(AIC_value)
}

# Compute AIC for all models
aic_values <- data.frame(
  Model = c("Pooled OLS", "One-way Fixed Effects", "Two-way Fixed Effects",
            "Random Intercept", "Random Slope", "Random Intercept & Slope"),
  AIC = c(
    AIC(model),  # Pooled OLS Model
    compute_plm_aic(fe_one_way),  # One-way Fixed Effects
    compute_plm_aic(fe_two_way),  # Two-way Fixed Effects
    AIC(random_intercept_model),  # Random Intercept Model
    AIC(random_slope_model_scaled),  # Random Slope Model
```

```
    AIC(random_intercept_slope_model)  # Random Intercept & Slope Model
  )
)

# Print AIC values
print(aic_values)
```

```
##                       Model        AIC
## 1               Pooled OLS 13234.3266
## 2    One-way Fixed Effects 13182.6317
## 3    Two-way Fixed Effects   362.1792
## 4         Random Intercept 13215.2196
## 5             Random Slope 13209.4629
## 6 Random Intercept & Slope 13209.4629
```

Part 1. Compare Models

Much of these distinctions have come up in previous answers, but the high level set of differences across the models is that the random intercept model (3.1) allows each age group to have its own baseline level of willingness but assumes a constant effect of hh_income across all groups. The random slope model (3.2) permits the effect of hh_income to vary across age groups while maintaining a common baseline willingness. The random intercept and slope model (4.1) combines both approaches, allowing each age group to have its own baseline willingness and a unique income effect, capturing both individual differences in willingness to pay and variations in how income influences willingness to pay.

Part 2. Best Fit

The two-way fixed effects model provides the best fit for explaining the relationship between household income and willingness to pay for coffee. Among all models estimated, its AIC (362.18) is substantially lower than that of the pooled OLS model, one-way fixed effects model, and all mixed effects models. This improvement in fit suggests that both age group and state capture meaningful variation that the other models fail to account for. By including these fixed effects, the model isolates the effect of household income more effectively than alternatives that ignore these differences.

From a theoretical perspective, age group and state make sense to include as fixed effects because they are stable, meaningful categories that we would expect to influence consumer behavior. Different age groups have distinct spending habits, disposable income levels, and preferences, which could affect willingness to pay. Controlling for these differences ensures that the estimated relationship between income and willingness to pay is not confounded by these factors or generational effects. Similarly, state-level differences in cost of living, coffee culture, and market conditions can affect pricing and consumer behavior. A model that does not account for these factors could potentially attribute regional variations in willingness to pay to income alone, giving misleading conclusions about the associations with income.

Going through the alternatives, the pooled OLS model fails to properly account for group-level heterogeneity, which can lead to biased estimates, while the one-way fixed effects model, using age group only, somewhat improves the fit but still fails to capture the state-level variation. Mixed effects models, which allow for random variation across groups, provide a better alternative to OLS but still underperform when compared to the two-way fixed effects model. The fact that mixed effects models have significantly higher AIC values suggests that treating age group and state as fixed rather than random effects better accounts for any unobserved heterogeneity in the data. Given both the statistical evidence and theoretical justification, the two-way fixed effects model is the most appropriate choice.