Lab Assignment 3 Markdown

Michael Sullivan

2025-02-22

```
### Packages------------------------------------------------------------------

library(dplyr)
library(ggplot2)
```

```
### Data Cleaning and Setup---------------------------------------------------

original_data <- read.csv("GACTT_RESULTS_ANONYMIZED_HW3.csv")
data <- read.csv("GACTT_RESULTS_ANONYMIZED_HW3.csv")

# Data exploration
head(na.omit(data))
sapply(data, class)
sapply(data, function(x) sum(is.na(x)))

table(data$monthly_spending)
table(data$age_num)
table(data$self_experience)
table(data$children)
table(data$most_for_cup)
table(data$equipment_spent)
table(data$marital_status)
table(data$ethnicity_race[!is.na(data$monthly_spend_mdpt)])

# Creating numerical variables
data <- data %>%
  mutate(age_mdpt = case_when(
    age_num == 18 ~ 19.5,
    age_num == 21 ~ 25.5,
    age_num == 30 ~ 35,
    age_num == 40 ~ 45,
    age_num == 50 ~ 55,
    age_num == 60 ~ 62.5,
    age_num == 65 ~ 70
  ))

data <- data %>%
  mutate(monthly_spend_mdpt = case_when(
    monthly_spending == "<$20" ~ 10,
    monthly_spending == "$20-$40" ~ 30,
    monthly_spending == "$40-$60" ~ 50,
    monthly_spending == "$60-$80" ~ 70,
```

```
    monthly_spending == "$80-$100" ~ 90,
    monthly_spending == ">$100" ~ 120
  ))

data <- data %>%
  mutate(equip_spent_mdpt = case_when(
    equipment_spent == "Less than $20" ~ 10,
    equipment_spent == "$20-$50" ~ 35,
    equipment_spent == "$50-$100" ~ 75,
    equipment_spent == "$100-$300" ~ 200,
    equipment_spent == "$300-$500" ~ 400,
    equipment_spent == "$500-$1000" ~ 750,
    equipment_spent == "More than $1000" ~ 1500
  ))

data <- data %>%
  mutate(children_mdpt = case_when(
    children == "None" ~ 0,
    children == "1" ~ 1,
    children == "2" ~ 2,
    children == "3" ~ 3,
    children == "More than 3" ~ 5
  ))

data <- data %>%
  mutate(most_for_cup_mdpt = case_when(
    most_for_cup == "Less than $2" ~ 1,
    most_for_cup == "$2-$4" ~ 1,
    most_for_cup == "$4-$6" ~ 5,
    most_for_cup == "$6-$8" ~ 7,
    most_for_cup == "$8-$10" ~ 9,
    most_for_cup == "$10-$15" ~ 12.5,
    most_for_cup == "$15-$20" ~ 17.5,
    most_for_cup == "More than $20" ~ 25
  ))

data <- data %>%
  mutate(race4 = case_when(
    ethnicity_race == "Black/African American" ~ "Other",
    ethnicity_race == "Other (please specify)" ~ "Other",
    ethnicity_race == "Native American/Alaska Native" ~ "Other",
    ethnicity_race == "White/Caucasian" ~ "White",
    ethnicity_race == "Asian/Pacific Islander" ~ "Asian/Pacific Islander",
    ethnicity_race == "Hispanic/Latino" ~ "Hispanic/Latino"
  ))
```

```
### Question 1. Monthly Spending----------------------------------------------

### Part 1. Simple Linear Regression----------------------------------------------

# Regressing monthly spend against age
model <- lm(monthly_spend_mdpt ~ age_mdpt, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = monthly_spend_mdpt ~ age_mdpt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.404 -14.661   1.486   5.339  78.325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.91781    1.96075  19.338  < 2e-16 ***
## age_mdpt     0.19266    0.04862   3.962  7.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.92 on 2846 degrees of freedom
##   (432 observations deleted due to missingness)
## Multiple R-squared:  0.005486,   Adjusted R-squared:  0.005137
## F-statistic:  15.7 on 1 and 2846 DF,  p-value: 7.603e-05
```

Question 1

Part 1 - Simple Linear Regression

Part 1.1

I ran a regression between monthly spending on coffee and age. The coefficient of the predictor is 0.193 with a standard error of 0.049. This coefficient shows a positive relationship between age and monthly spending, with an additional year of age corresponding to an additional $0.19 of monthly spending. This result has a p-value of 7.6e-05, which is below 0.05 and therefore statistically significant

R squared generally tells us how much of the variation in the response variable is explained by the predictors. In this case it tells us that 0.5% of the variation in monthly spending is explained by age.

```
### Part 2. Building Nested Models-----------------------------------------------

# Incorporating equipment spend
model <- lm(monthly_spend_mdpt ~ age_mdpt + equip_spent_mdpt, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = monthly_spend_mdpt ~ age_mdpt + equip_spent_mdpt,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.624 -13.592  -5.216  11.932  86.267
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      26.47399    2.08731  12.683  < 2e-16 ***
```

```
## age_mdpt          0.20088    0.04959   4.051 5.29e-05 ***
## equip_spent_mdpt  0.02281    0.00186  12.263  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.39 on 2143 degrees of freedom
##   (1134 observations deleted due to missingness)
## Multiple R-squared:  0.07234,    Adjusted R-squared:  0.07147
## F-statistic: 83.56 on 2 and 2143 DF,  p-value: < 2.2e-16
```

```
# Incorporating "most for cup"
model <- lm(monthly_spend_mdpt ~ age_mdpt +
              equip_spent_mdpt +
              most_for_cup_mdpt,
            data = data)
summary(model)
```

```
##
## Call:
## lm(formula = monthly_spend_mdpt ~ age_mdpt + equip_spent_mdpt +
##     most_for_cup_mdpt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.277 -13.519  -3.928  10.498  86.666
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       14.049126   2.288828   6.138 9.93e-10 ***
## age_mdpt           0.227284   0.048357   4.700 2.77e-06 ***
## equip_spent_mdpt   0.018980   0.001837  10.334  < 2e-16 ***
## most_for_cup_mdpt  1.415138   0.119807  11.812  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.72 on 2136 degrees of freedom
##   (1140 observations deleted due to missingness)
## Multiple R-squared:  0.1291, Adjusted R-squared:  0.1279
## F-statistic: 105.5 on 3 and 2136 DF,  p-value: < 2.2e-16
```

```
# Incorporating number of children
model <- lm(monthly_spend_mdpt ~ age_mdpt +
              equip_spent_mdpt +
              most_for_cup_mdpt +
              children_mdpt,
            data = data)
summary(model)
```

```
##
## Call:
## lm(formula = monthly_spend_mdpt ~ age_mdpt + equip_spent_mdpt +
##     most_for_cup_mdpt + children_mdpt, data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.458 -13.939  -3.993  10.750  86.819
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       15.103594   2.434797   6.203 6.67e-10 ***
## age_mdpt           0.194521   0.055526   3.503 0.000469 ***
## equip_spent_mdpt   0.018675   0.001876   9.954  < 2e-16 ***
## most_for_cup_mdpt  1.409770   0.120988  11.652  < 2e-16 ***
## children_mdpt      0.923186   0.574426   1.607 0.108177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.65 on 2051 degrees of freedom
##   (1224 observations deleted due to missingness)
## Multiple R-squared:  0.1298, Adjusted R-squared:  0.1281
## F-statistic: 76.49 on 4 and 2051 DF,  p-value: < 2.2e-16
```

Part 2 - Building Nested Models

Part 2.1

In the above models I add a series of covariates that I believe could be potential confounders with age. In particular my first concern was that the relationship between age and coffee spending was potentially actually a result of a relationship between wealth and coffee spending. Wealth could predict age because older people have more time to develop in their careers and accumulate assets, and wealth could predict coffee spending as it provides a greater abillity to spend on coffee.

Based on this idea, the first two covariates I add are for spending on equipment and for the most a person would spend for a cup of coffee. These are not direct measures of wealth but we would expect that they are linked to it. People who have more money are certainly going to be more likely to spend significant amounts on equipment, and are more likely to be willing to pay higher amounts for consumable goods.

The third covariate I add is number of children. I believe this could be a potential confounder as well. Number of children certainly would predict age. Young adults are less likely to have children or to have many children than older adults. It also could potentially (though this is less certain) predict coffee consumption. Adults with many children will potentially tend to be busier and less likely to get enough sleep, which could lead to greater coffee consumption.

Part 2.2

Looking at the model results, we see that after each additional control variable is added the p-value for age remains well below 0.05 and so its relationship remains significant. The coefficient varies slightly with the addition of new variables, but mostly hovers around 0.2 and does not move significantly. This tells us that, at least based on these proxies for wealth, the relationship between age and monthly coffee spending is not solely due to the increased spending power of older people, or solely due to their having more children.

From a hypothesis testing perspective we find a significant relationship between equipment expenditure and monthly spending, as well as between the most they would pay for a cup and monthly spending. This implies that there is likely a relationship between financial situation and monthly spend on coffee. Finally there is not a significant relationship between number of children and monthly spending on coffee, so the speculation above about how having children would impact coffee consumption may not hold (or may only hold in more specific cases, we don't have information on the ages of children or if they are still living with their

parents). Finally, looking at the R^2 value we see that it goes up significantly with the addition of equipment expenditure, then rises meaningfully again with "most for cup". This implies that these additional variables help to explain meaningful amounts of the variation in monthly spending. R^2 has almost no change with the addition of the children variable so we do not interpret that variable as having significant explanatory power.

Part 2.3

Of the models I examined I believe the model with age, equipment expenditure, and "most for cup" offers the best explanation of monthly spending. I believe this because the variables all have statistically significant relationships with the response variable, and have clear theoretical reasons to explain these connections. I believe that the fourth model, with children, is not preferable because the children variable adds almost nothing to the R^2, meaning it adds little to the explanation of variation by the model. It also has a weaker theoretical justification for inclusion.

Controlling for additional variables does change my interpretation from the original model. I originally suspected that the effect I saw of age on monthly spending was largely a result of wealth. After including the control variables I have greater reason to believe that there is an effect from age that is not only based in the better financial position of older people.

```
### Part 3. Modeling Interaction Effects----------------------------------------

model <- lm(monthly_spend_mdpt ~ age_mdpt +
              equip_spent_mdpt +
              most_for_cup_mdpt +
              age_mdpt * children_mdpt,
            data = data)
summary(model)
```

```
##
## Call:
## lm(formula = monthly_spend_mdpt ~ age_mdpt + equip_spent_mdpt +
##     most_for_cup_mdpt + age_mdpt * children_mdpt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.568 -13.770  -3.995  10.788  86.844
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             14.670661   2.617601   5.605 2.37e-08 ***
## age_mdpt                 0.206655   0.061708   3.349 0.000826 ***
## equip_spent_mdpt         0.018633   0.001879   9.917  < 2e-16 ***
## most_for_cup_mdpt        1.407876   0.121085  11.627  < 2e-16 ***
## children_mdpt            2.000305   2.456188   0.814 0.415514
## age_mdpt:children_mdpt -0.022549   0.049992  -0.451 0.652004
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.66 on 2050 degrees of freedom
##   (1224 observations deleted due to missingness)
## Multiple R-squared:  0.1299, Adjusted R-squared:  0.1278
## F-statistic: 61.21 on 5 and 2050 DF,  p-value: < 2.2e-16
```

Part 3 - Modeling Interaction Effects

Part 3.1-3.2

Given my discussion above in which I noted that having children does not necessarily entail having young children, or the children still living with their parents, I decided to add an interaction term between age and children. My hope is that this term captures some of that dynamic, because of parents with children, the younger the parent is the more likely it is that they have a young child and that their children still live at home.

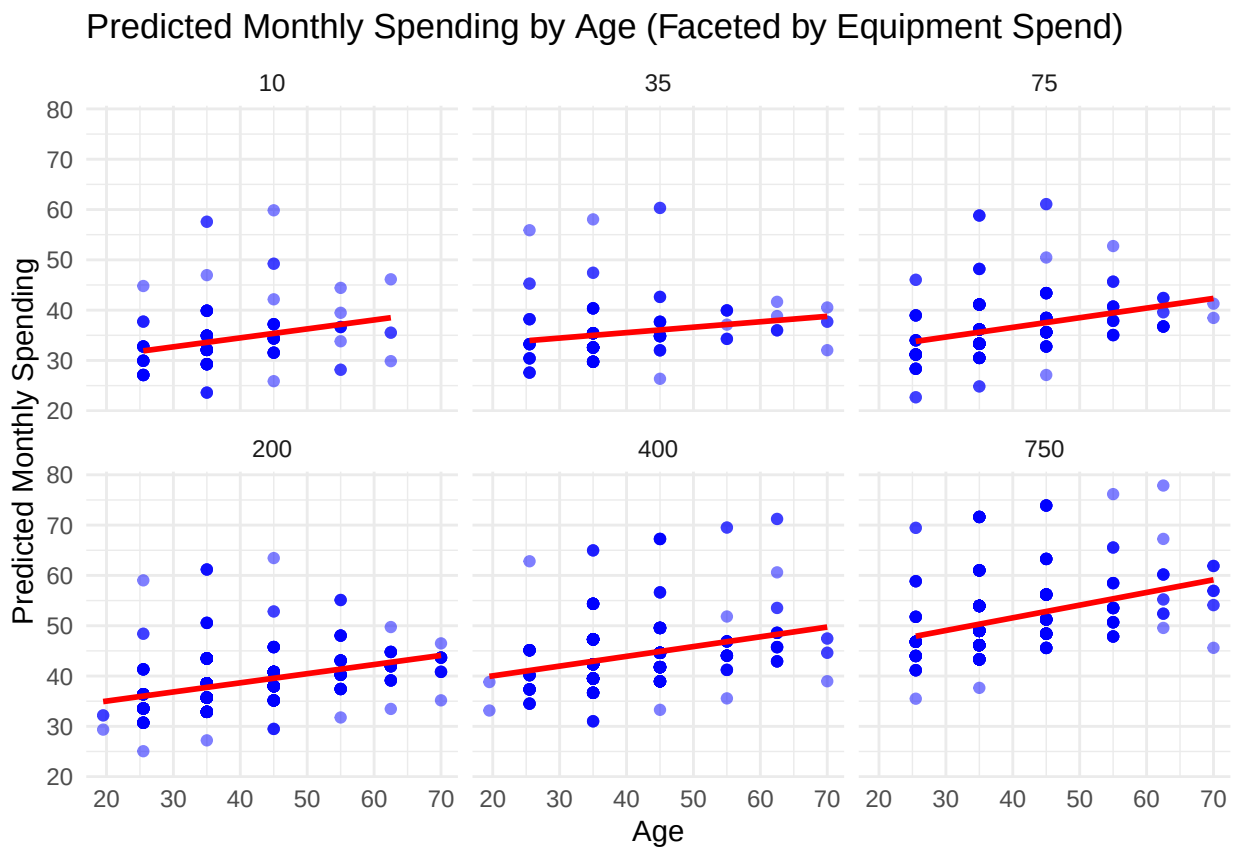
The interaction term also produces a non-statistically significant p-value, so we don't find an effect associated with the interaction of age and number of children. This strengthens my original conclusion about the relationship between the main predictor and monthly spending because it indicates that children failing to have a statistically significant effect is probably not just due to some children having moved out of their older parents' houses which would reduce the observed effect.

```
### Part 4. Visualization-------------------------------------------------------

model <- lm(monthly_spend_mdpt ~ age_mdpt +
              equip_spent_mdpt +
              most_for_cup_mdpt,
            data = data)
summary(model)
```

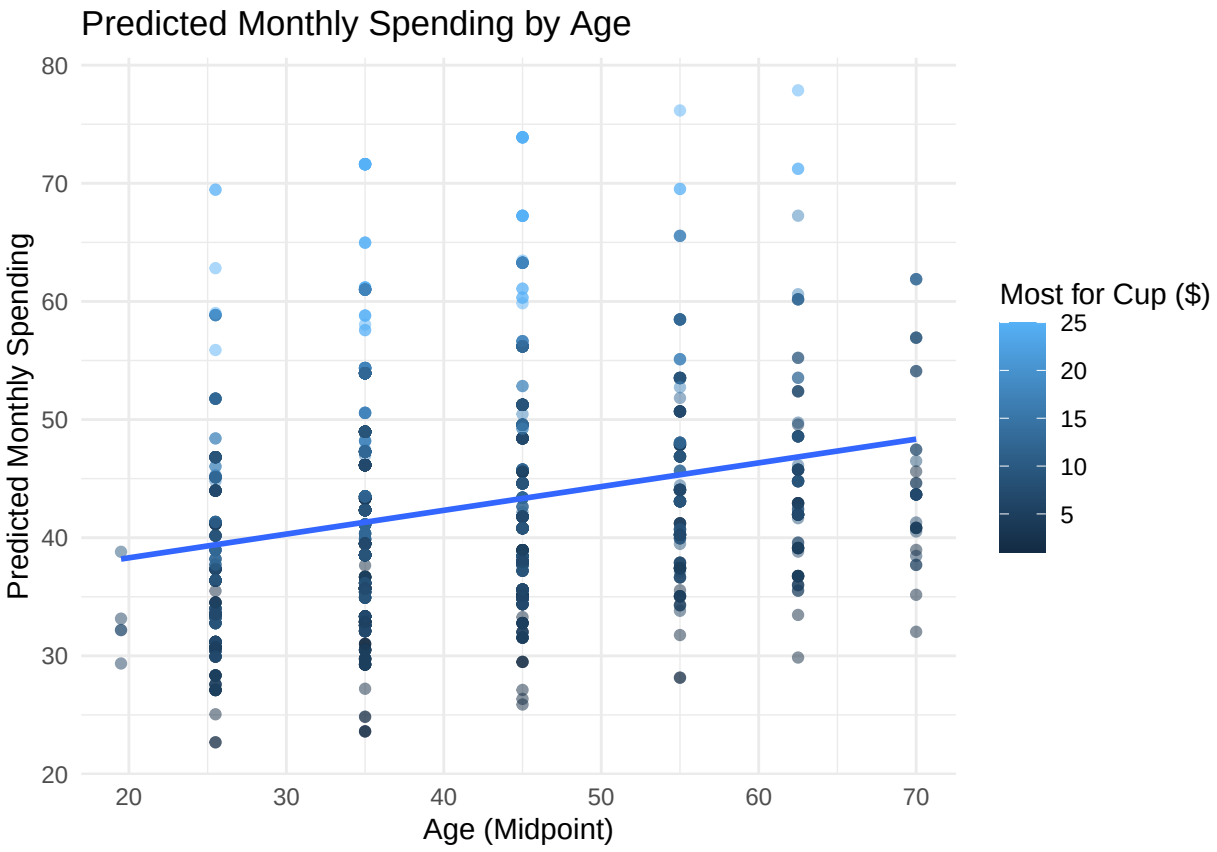
```
##
## Call:
## lm(formula = monthly_spend_mdpt ~ age_mdpt + equip_spent_mdpt +
##     most_for_cup_mdpt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.277 -13.519  -3.928  10.498  86.666
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       14.049126   2.288828   6.138 9.93e-10 ***
## age_mdpt           0.227284   0.048357   4.700 2.77e-06 ***
## equip_spent_mdpt   0.018980   0.001837  10.334  < 2e-16 ***
## most_for_cup_mdpt  1.415138   0.119807  11.812  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.72 on 2136 degrees of freedom
##   (1140 observations deleted due to missingness)
## Multiple R-squared:  0.1291, Adjusted R-squared:  0.1279
## F-statistic: 105.5 on 3 and 2136 DF,  p-value: < 2.2e-16
```

```
data$predicted_spending <- predict(model, newdata = data)


ggplot(
  data %>% filter(!is.na(equip_spent_mdpt)), # Remove NA values
  aes(x = age_mdpt, y = predicted_spending)
) +
```

```
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  facet_wrap(~equip_spent_mdpt) + # Facet by equipment spending
  labs(
    title = "Predicted Monthly Spending by Age (Faceted by Equipment Spend)",
    x = "Age",
    y = "Predicted Monthly Spending"
  ) +
  theme_minimal()
```




```
ggplot(data, aes(x = age_mdpt,
                 y = predicted_spending,
                 color = most_for_cup_mdpt)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Predicted Monthly Spending by Age",
    x = "Age (Midpoint)",
    y = "Predicted Monthly Spending",
    color = "Most for Cup ($)"
  ) +
  theme_minimal()
```



Part 4 - Visualization

Above are visualizations based on the model that includes equipment spend and "most for cup". I am using this model because of the lack of relationship we found above for number of children and the interaction term with number of children. Both visualizations show the relationship between age and monthly spend. The first also incorporates the equipment spend, with separate facets for each bracket of equipment spend, showing that monthly spend is higher for the higher equipment spend brackets. THe second visualization incorporates the "most for cup" data in the color of the scatterplot, with lighter dots displaying for higher values. Again we see that with higher values the monthly spending increases.

```
### Question 2. Coffee Preference Analysis-----------------------------------

### Part 1. Binary Indicator for Coffee Preference---------------------------

data <- data %>%
  mutate(a_pref_bin = case_when(
    abc_prefer == "Coffee A" ~ 1,
    abc_prefer == "Coffee B" ~ 0,
    abc_prefer == "Coffee C" ~ 0
  ))
```

Part 1 - Binary Indicator

Part 1.1

I created the binary indicator above. ### Part 1.2 I select race as the demographic predictor.

```
### Part 2. OLS Modeling-------------------------------------------------------

# Regress preference for A against race
model <- lm(a_pref_bin ~ race4, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = a_pref_bin ~ race4, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4683 -0.4683 -0.4419  0.5317  0.6296
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.461988   0.026949  17.143   <2e-16 ***
## race4Hispanic/Latino -0.020128   0.046587  -0.432   0.6657
## race4Other           -0.091618   0.050657  -1.809   0.0706 .
## race4White            0.006348   0.029046   0.219   0.8270
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4984 on 2761 degrees of freedom
##   (515 observations deleted due to missingness)
## Multiple R-squared:  0.001872,   Adjusted R-squared:  0.0007873
## F-statistic: 1.726 on 3 and 2761 DF,  p-value: 0.1595
```

Part 2 OLS Modeling

Part 2.1

I chose to look at the relationship between race and preference for coffee A. As part of my data cleaning I consolidated the race variable into 4 categories because some of the original categories had very low numbers of observations. I find that there is not a statistically significant relationship between any of the racial categories and preference for coffee A.

```
# Regress preference for A against race with controls for age and # of children
model <- lm(a_pref_bin ~ race4 + age_mdpt + equip_spent_mdpt, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = a_pref_bin ~ race4 + age_mdpt + equip_spent_mdpt,
##     data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6018 -0.4181 -0.3281  0.5310  0.7834
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.787e-01  5.279e-02   9.068  < 2e-16 ***
## race4Hispanic/Latino  1.071e-02  5.423e-02   0.198    0.843
## race4Other           -2.989e-02  5.724e-02  -0.522    0.602
## race4White            4.570e-02  3.409e-02   1.341    0.180
## age_mdpt             -4.501e-03  1.100e-03  -4.091 4.46e-05 ***
## equip_spent_mdpt      2.562e-04  4.117e-05   6.223 5.88e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4882 on 2087 degrees of freedom
##   (1187 observations deleted due to missingness)
## Multiple R-squared:  0.02784,    Adjusted R-squared:  0.02551
## F-statistic: 11.95 on 5 and 2087 DF,  p-value: 1.98e-11
```

Part 2.2

For control variables I added age and equipment spend. I chose these because as we explored above they predict the outcome, and I believe they likely predict the treatment because different racial and ethnic groups have systemically different levels of wealth and population-level aging curves due to demographic differences, so these would meet the criteria for confounders.

Part 2.3

Comparing the results of the two models, the added results don't significantly alter my interpretation of the results. The main predictor fails to produce a statistically significant relationship in both cases. The second model does help to increase my confidence in this result because it indicates that the control variables I included were not confounding the original result.

```
### Part 3. Binary Logistic Regression----------------------------------------

model_logit <- glm(a_pref_bin ~ race4 + age_mdpt + equip_spent_mdpt,
  data = data,
  family = binomial
)
summary(model_logit)
```

```
##
## Call:
## glm(formula = a_pref_bin ~ race4 + age_mdpt + equip_spent_mdpt,
##     family = binomial, data = data)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.0651114  0.2241650  -0.290    0.771
## race4Hispanic/Latino  0.0457906  0.2285618   0.200    0.841
## race4Other           -0.1351174  0.2468471  -0.547    0.584
```

```
## race4White            0.1928833  0.1441933   1.338    0.181
## age_mdpt             -0.0192833  0.0047486  -4.061 4.89e-05 ***
## equip_spent_mdpt      0.0010596  0.0001725   6.143 8.12e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2855.1  on 2092  degrees of freedom
## Residual deviance: 2796.2  on 2087  degrees of freedom
##   (1187 observations deleted due to missingness)
## AIC: 2808.2
##
## Number of Fisher Scoring iterations: 4
```

Part 3 - Binary Logistic Regression

Part 3.1

The difference in interpretation of OLS and logistic regression here, is that OLS coefficients represent linear change in the outcome variable for a given change in the predictor variables, while logistic regression coefficients represent a change in the log-odds of the outcome variable for a given change in the predictor variables.

Part 3.2

In this case logistic regression is more appropriate. This is because the outcome variable is a binary variable, so the linear change in outcome variable that is associated with the coefficients of the predictors is not particularly meaningful. A change in a predictor times its coefficient can lead to outcome values less than 0 or greater than 1, which don't meaningfully represent the value of the binary outcome, or the probability of the binary outcome. In the case of logistic regression we can look at a change in a predictor as changing the log-odds of the outcome in a consistent and meaningful way. "‘