Final Project

Introduction to Applied Statistics & Data Science

Sophie Ennis

2025-03-11

# Does Money Buy Happiness?

## Question 1

### Research Question

Does a higher family income predict higher life satisfaction, in consideration of various social positions?

## Question 2

### Background

Richard Easterlin (1973) argued that while richer societies are no happier than poor ones, within countries, richer people are happier than poorer people. However, recent studies have found that “economic growth does materially increase a country’s collective sense of well-being and that differences in well-being within a country are not significantly related to income” (Lane, 1993). For most Americans, there is “no substantial relation between income and well-being,” but money does appear to buy greater happiness and well-being for the poor (Lane, 1993). Lane (1993) writes that income and life satisfaction are not closely correlated, but how does this contend with the scholarly discrepancies over the effect of income on wellbeing? Boyce et al. (2010) found that the relative rank of an individual’s income, rather than their absolute income, better predicted their life satisfaction. The unequal outcomes brought on by disparities in income, rather than the income itself, are thus the likely drivers behind diminished life satisfaction. Income disparities may lead to stress, which may then lead to diminished life satisfaction. What then, are these drivers of disparities in income that could influence life satisfaction? Do disparities in income necessitate diminished life satisfaction? How might different social categories influence disparities in income, which may influence life satisfaction in different ways? Do all disadvantaged social categories experience diminished life satisfaction, and could this be due to lower income if so? All these questions lead me to ask if money buys happiness, and more specifically, if a higher family income predicts higher life satisfaction through a variety of social conditions.

## Question 3

### A description of the data. Describe the sample and the population it is drawn from. Describe how key covariates are measured or operationalized. Include a table of descriptive statistics, including valid N, means, standard deviations, and minimum and maximum values of each variable. Include only variables that you will utilize in an analysis in this report, not all the variables available.

gss$Female <- ifelse(gss$Sex=="Female", 1, 0)   
gss$RaceBlack <- ifelse(gss$Race=="Black", 1, 0)   
gss$RaceOther <- ifelse(gss$Race=="Other", 1, 0)

table1(~LifeSat + FamIncome10k + SES + RaceBlack + RaceOther + Female + Age + YearsEd, data=gss)

|  | Overall (N=2348) |
| --- | --- |
| **r's rating of life overall now from 0-10** |  |
| Mean (SD) | 7.42 (1.61) |
| Median [Min, Max] | 8.00 [0, 10.0] |
| Missing | 935 (39.8%) |
| **family income in constant $** |  |
| Mean (SD) | 3.37 (3.12) |
| Median [Min, Max] | 2.50 [0.0227, 12.0] |
| Missing | 196 (8.3%) |
| **r's socioeconomic index (2010)** |  |
| Mean (SD) | 46.9 (23.0) |
| Median [Min, Max] | 41.0 [10.6, 92.8] |
| Missing | 100 (4.3%) |
| **RaceBlack** |  |
| Mean (SD) | 0.164 (0.370) |
| Median [Min, Max] | 0 [0, 1.00] |
| **RaceOther** |  |
| Mean (SD) | 0.115 (0.319) |
| Median [Min, Max] | 0 [0, 1.00] |
| **Female** |  |
| Mean (SD) | 0.552 (0.497) |
| Median [Min, Max] | 1.00 [0, 1.00] |
| **age of respondent** |  |
| Mean (SD) | 49.0 (18.1) |
| Median [Min, Max] | 48.0 [18.0, 89.0] |
| Missing | 7 (0.3%) |
| **highest year of school completed** |  |
| Mean (SD) | 13.7 (2.97) |
| Median [Min, Max] | 14.0 [0, 20.0] |
| Missing | 3 (0.1%) |

lifesat\_n <- length(gss$LifeSat[!is.na(gss$LifeSat)])   
famincome10k\_n <- length(gss$FamIncome10k[!is.na(gss$FamIncome10k)])   
ses\_n <- length(gss$SES[!is.na(gss$SES)])   
black\_n <- length(gss$RaceBlack[!is.na(gss$RaceBlack)])   
other\_n <- length(gss$RaceOther[!is.na(gss$RaceOther)])   
female\_n <- length(gss$Female[!is.na(gss$Female)])   
age\_n <- length(gss$Age[!is.na(gss$Age)])   
yearsed\_n <- length(gss$YearsEd[!is.na(gss$YearsEd)])   
  
n\_table <- data.frame(  
 Variable = c("LifeSat", "FamIncome10k", "SES", "RaceBlack",   
 "RaceOther", "Female", "Age", "YearsEd"),  
 n = c(lifesat\_n, famincome10k\_n, ses\_n, black\_n, other\_n, female\_n, age\_n, yearsed\_n))  
n\_table

## Variable n  
## 1 LifeSat 1413  
## 2 FamIncome10k 2152  
## 3 SES 2248  
## 4 RaceBlack 2348  
## 5 RaceOther 2348  
## 6 Female 2348  
## 7 Age 2341  
## 8 YearsEd 2345

I draw on data from the 2018 General Social Survey. The 2018 GSS draws a sample of 2,348 respondents from the U.S. population ages 18 and over. The GSS is a nationally representative survey of the United States that aims to monitor behavioral, attitudinal, and demographic trends.

The LifeSat variable is operationalized as the respondents’ ranking of their current overall life satisfaction on a scale of 0 to 10. The respondents in the sample have a mean rating of life satisfaction of 7.42 on a scale of 0 to 10, 10 being the highest life satisfaction, with a standard deviation of 1.61.

The FamIncome10k variable is operationalized as the respondents’ family income in constant dollars by $10,000, which will be centered around its mean. Respondents’ mean family income divided by 10,000 dollars is 3.37, indicating a mean household income of 33,700 dollars, with a standard deviation of 3.12 (31,200 dollars). The minimum family income is 0.0227 (227 dollars) and the maximum is 12.0 (120,000 dollars).

The SES variable is operationalized as the respondents’ socioeconomic index, which will be centered around its mean. Respondents’ mean socioeconomic status is 46.9 with a standard deviation of 23, with the minimum being 10.6 and the maximum being 92.8.

Race is operationalized to represent the complex social category of race as three independent categories: White, Black, and Other. Black and Other are dummy variables of Race centered around their means. The mean of 0.164 for RaceBlack represents the proportion of the sample that is coded as 1 (Black), which is thus 16.4%. The “average” person in the sample is thus not Black, exemplified by the median of 0 (Not Black). The mean of 0.115 for RaceOther represents the proportion of the sample that is coded as 1 (Other Race), which is thus 11.5%. The “average” person in the sample is thus not of the ‘Other’ race, exemplified by the median of 0 (Not Other Race).

The Female variable is operationalized as a binary indicator of sex where Female is 1 and Male is 0. The mean gender is 0.552, and with Female being 1 and Male being 0, the sample is more Female, with a standard deviation of 0.497. This is reflected by the median of 1, representing Female.

The Age variable is operationalized as the age of the respondent, with a minimum of 18 because the General Social Survey only surveys adults. The Age variable will be centered around its mean. The mean age of respondents is 49 years old, with a minimum of 18 years old and a maximum of 89 years old.

The YearsEd variable is operationalized as the highest year of school completed, which will be centered at 12 years of education. Respondents have completed a mean of 13.7 years of education.

## Question 4

### Create a visualization that is descriptive of an important aspect of your data, relevant to your research question (but not based on regression modeling).

I grouped family income by low, middle, and high income.

gss <- gss %>%  
 mutate(IncomeGroup = case\_when(  
 FamIncome10k <= 3 ~ "Low Income",   
 FamIncome10k > 3 & FamIncome10k <= 7 ~ "Middle Income",   
 FamIncome10k > 7 ~ "High Income",   
 TRUE ~ NA\_character\_   
 ))  
gss$IncomeGroup <- factor(gss$IncomeGroup, levels= c("Low Income", "Middle Income", "High Income"))

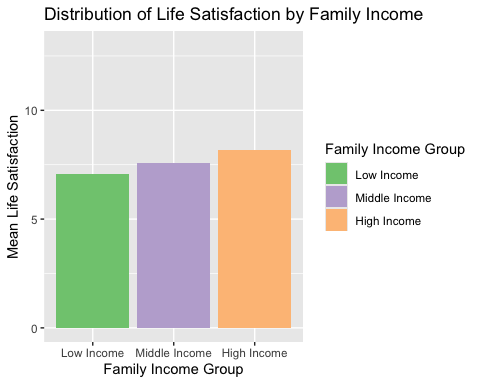
I calculated the mean life satisfaction for each income group.

lifesat\_income <- gss %>%  
 group\_by(IncomeGroup) %>%  
 summarise(LifeSatMean = mean(LifeSat, na.rm = TRUE))  
  
lifesat\_income\_clean <- lifesat\_income %>%  
 filter(!is.na(LifeSatMean) & !is.na(IncomeGroup))  
lifesat\_income\_clean

## # A tibble: 3 × 2  
## IncomeGroup LifeSatMean  
## <fct> <dbl>  
## 1 Low Income 7.07  
## 2 Middle Income 7.59  
## 3 High Income 8.17

I plotted these differences in mean life satisfaction by family income group.

ggplot(lifesat\_income\_clean, aes(  
 x = IncomeGroup,   
 y = LifeSatMean,  
 fill = IncomeGroup)) +  
 geom\_bar(stat = "identity", na.rm = TRUE) +  
 ylim(0, 13) +  
 scale\_fill\_brewer(palette = "Accent") +  
 labs(  
 title = "Distribution of Life Satisfaction by Family Income",  
 x = "Family Income Group",  
 y = "Mean Life Satisfaction",  
 fill = "Family Income Group")



## Question 5

### Begin with a naïve regression model that includes your primary predictor of interest. Interpret the coefficient, standard error, and p-value.

base\_model <- summ(lm(LifeSat ~ FamIncome10k, data=gss))  
base\_model

A white paper with black text

AI-generated content may be incorrect.

The coefficient for FamIncome10k represents the expected change in life satisfaction for each 10,000 dollar increase in family income. Therefore, expected (average) life satisfaction increases by 0.12 when one’s family income increases by one unit, which is represented by 10,000 dollars. The standard error (0.01) measures the degree of variability or uncertainty in the coefficient estimate. It reflects how much the estimated coefficient could vary across different samples drawn from the same population. A smaller standard error indicates that the estimated coefficient is more precise, meaning that the true population parameter is likely to be close to the estimated value. Following the traditional convention in the social sciences to use a critical value of 0.05 (α = 0.05) (Jean, Lecture Day 6, Slide 15), the p-value is less than 0.05, indicating that the relationship between family income and life satisfaction is statistically significant.

## Question 6

### Add three to five covariates to your model. Interpret the new coefficients, and the change (if any) to the coefficient for the predictor from your naïve model.

model1 <- summ(lm(LifeSat ~ FamIncome10k + SES + Race + Female + YearsEd + Age, data=gss))  
model1

A screenshot of a computer code

AI-generated content may be incorrect.

The coefficient for FamIncome10k decreased from 0.12 to 0.10. The expected (average) life satisfaction now increases by 0.10 rather than 0.12 when one’s family income increases by one unit, which is represented by 10,000 dollars, while holding all other variables constant. This main effect of family income on life satisfaction, holding other variables constant, remains statistically significant, with a p-value of less than 0.05. As for the new covariates, the coefficient for SES is 0.00, indicating that SES has a very small positive effect on life satisfaction, holding other variables constant. Although the p-value is 0.03, which is statistically significant, the small effect size suggests that the impact of SES on life satisfaction is minimal. Being Black is associated with a small positive increase in life satisfaction (0.05), holding other variables constant, but the effect is not statistically significant, with a p-value over 0.05. Belonging to the ‘Other’ racial category is associated with a small negative effect on life satisfaction (-0.04), holding other variables constant. Again, this effect is not statistically significant, with a p-value over 0.05. Being female is associated with a positive change of 0.24 in life satisfaction, holding other variables constant. This is statistically significant with a p-value of 0.01, indicating that, on average, females report higher life satisfaction than males. The coefficient for YearsEd is -0.00, indicating that years of education has a very small negative effect on life satisfaction, holding other variables constant. However, this effect is not statistically significant, with a p-value over 0.05. As age increases by one year, life satisfaction increases by 0.01 points. This main effect of age on life satisfaction is statistically significant, with a p-value of less than 0.05.

## Question 7

### Do you have reason to believe that any of your covariates might mediate the relationship between your key predictor and your outcome variable? If so, provide statistical and/or theoretical evidence of this.

SES may explain how FamIncome10k affects LifeSat, because higher family income can lead to greater access to resources and opportunities, which in turn contribute to a higher socioeconomic status, which then can lead to higher life satisfaction. YearsEd may also explain how FamIncome10k affects LifeSat, because higher family income may allow for greater access to educational opportunities like tutoring, and more education can contribute to better job prospects, higher earnings, and higher job satisfaction, which may lead to higher life satisfaction. YearsEd may be less of an intuitive mediator than SES, but could be a potential mediator of the relationship between FamIncome10k and LifeSat nonetheless.

## Question 8

### Do you have reason to believe that any of your covariates might moderate the relationship between your key predictor and your outcome variable? Test two of your covariates for this possibility. Interpret the new coefficients in your model and describe any moderating relationships.

Race may be a moderator in the relationship between FamIncome10k and LifeSat because the effect of family income on life satisfaction might be stronger or weaker or change the direction of the effect depending on the racial category. For instance, the racial discrimination that Black individuals and individuals of other minority races face, no matter their income, may lower their life satisfaction.

Sex may be another moderator in the relationship between FamIncome10k and LifeSat because the effect of family income on life satisfaction might be stronger or weaker or change the direction of the effect depending on an individual’s sex. For instance, the sexual harassment and discrimination that women face, no matter their income, may lower their life satisfaction.

I will test if the variables Race and Female might moderate the relationship between FamIncome10k and LifeSat.

moderation\_model1 <- summ(lm(LifeSat ~ FamIncome10k + SES + Race + Female + YearsEd + Age + FamIncome10k\*Race, data=gss))  
moderation\_model1

A screenshot of a computer code

AI-generated content may be incorrect.

The coefficient for the main effect of family income was 0.10 in model1 and is now 0.11, meaning that for every 10,000 dollar increase in family income, life satisfaction now increases by 0.11 points, 0.01 more points in life satisfaction than before. However, this only applies to the White reference group. The coefficient for the interaction term for FamIncome10k and RaceBlack, -0.01, indicates that for Black individuals, the effect of family income on life satisfaction (when family income is 0 and increases by one unit) is slightly lower (by 0.01) than for the reference group, but this is not statistically significant due to the p-value of 0.81, which is higher than the conventional 0.05. The coefficient for the interaction term for FamIncome10k and RaceOther, -0.06, indicates that for individuals of other racial groups, the effect of family income on life satisfaction (when family income is 0 and increases by one unit) is slightly lower (by 0.06) than for the reference group, but this is not statistically significant due to the p-value of 0.25, which is higher than the conventional 0.05. In the original model, RaceBlack and RaceOther had non-significant coefficients, meaning race alone had little effect on life satisfaction when family income is 0, although RaceOther flipped from negative to positive and the effect more than tripled, and RaceBlack remained positive but nearly doubled in effect size. But again, because these coefficients remain insignificant, we can conclude that race alone still does not have a significant effect on life satisfaction when controlling for other factors. The other variables remain nearly unchanged, meaning that the inclusion of the interaction term does not significantly impact their effects. All this said, there is no statistically significant evidence that Race moderates the relationship between FamIncome10k and LifeSat.

moderation\_model2 <- summ(lm(LifeSat ~ FamIncome10k + SES + Race + Female + YearsEd + Age + FamIncome10k\*Female, data=gss))  
moderation\_model2

A screenshot of a computer

AI-generated content may be incorrect.

The coefficient for the main effect of family income was 0.10 in model1 and remains as such in this model, meaning that for every 10,000 dollar increase in family income, life satisfaction increases by 0.10 points, holding other variables constant. This coefficient, however, now represents the main effect of family income for the male reference group. The new interaction term suggests that for women, the effect of a one unit increase in income (when income is 0) on life satisfaction is slightly lower by 0.01 points, but this is not statistically significant. In the original model, the main effect of female suggested that women have higher life satisfaction than men, by 0.24 points, holding other variables constant. This was statistically significant, with a p-value of 0.01. In the new model, the effect size increases slightly as women have higher life satisfaction than men by 0.26 points, increasing from the original model by 0.02 points, but the p-value also increases to 0.05, remaining significant but barely so. The other variables remain virtually unchanged, meaning that adding the interaction did not strongly affect the relationships of these variables with life satisfaction. All this said, there is no statistically significant evidence that Female moderates the relationship between FamIncome10k and LifeSat.

## Question 9

### Make any additional changes you believe would help better address the research question. You may use any of the techniques we learned this quarter. Show us what you’ve learned! You may also include one additional variable in your model from the GSS (you may use one we’ve examined before, or clean a new one yourself). Describe the reasoning behind these changes and present your final model.

I decided to center the variables FamIncome10k, SES, RaceBlack, RaceOther, Female, and Age around their means and YearsEd at a meaningful number of 12 years of education. This decision was rooted in a desire for a more meaningful intercept. Without centering, the intercept represents the expected life satisfaction when all the predictors are equal to zero, which may not be meaningful. After centering, the intercept represents the expected life satisfaction when all the predictors are at their average values, which is more interpretable. This is a more meaningful baseline for understanding how changes in the predictors affect life satisfaction. I also decided to include the variable PolViews in the final model because I was curious about how political views might act as a mediator in the relationship between family income and life satisfaction. For instance, higher income might lead an individual to become fiscally conservative to protect their assets, which may affect their life satisfaction when the current administration aligns with their beliefs. Political views thus may be a mechanism in which family income affects life satisfaction, so I included the variable PolViews in the model to see how it may affect the coefficients of other covariates in the model.

gss$AgeC <- gss$Age - mean(gss$Age, na.rm=T)  
gss$SESC <- gss$SES - mean(gss$SES, na.rm=T)  
gss$FamIncome10kC <- gss$FamIncome10k - mean(gss$FamIncome10k, na.rm=T)  
gss$YearsEd12 <- gss$YearsEd - 12   
gss$FemaleC <- gss$Female - mean(gss$Female, na.rm=T)   
gss$RaceBlackC <- gss$RaceBlack - mean(gss$RaceBlack, na.rm=T)   
gss$RaceOtherC <- gss$RaceOther - mean(gss$RaceOther, na.rm=T)

model\_centered <- summ(lm(LifeSat ~ FamIncome10kC + SESC + RaceBlackC + RaceOtherC + FemaleC + YearsEd12 + AgeC, data=gss))  
model\_centered

A screenshot of a computer code

AI-generated content may be incorrect.

gss$PolViewsC <- gss$PolViews - mean(gss$PolViews, na.rm=T)

I present the final model as such:

final\_model <- summ(lm(LifeSat ~ FamIncome10kC + SESC + RaceBlackC + RaceOtherC + FemaleC + YearsEd12 + AgeC + PolViewsC, data=gss))  
final\_model

A screenshot of a computer

AI-generated content may be incorrect.

## Question 10

### Compare the results of your final to the bivariate relationships described in our correlation matrix and your naïve model.

base\_model

A screenshot of a computer code

AI-generated content may be incorrect.

correlation\_matrix <- subset(gss, select= c(LifeSat, FamIncome10k, SES, RaceBlack, RaceOther, Female, YearsEd, Age))   
rcorr(as.matrix(correlation\_matrix))

A screenshot of a computer screen

AI-generated content may be incorrect.

The naive model only included the predictor variable FamIncome10k and the outcome variable of LifeSat. The correlation matrix represents the correlations between the variables LifeSat, FamIncome10k, SES, RaceBlack, RaceOther, Female, YearsEd, and Age. The final model includes the predictor variable, but centered, as FamIncome10kC, the outcome variable, LifeSat, and the covariates SESC, RaceBlackC, RaceOtherC, FemaleC, AgeC, and PolViewsC, all centered around their means, and YearsEd12, centered at 12 years of education.

Given these general “rules”:

|r| < 0.3 is a weak correlation

0.3 < |r| < 0.7 is a moderate correlation  
|r| > 0.7 is a strong correlation

I make these following analyses:

In the final model, the coefficient for FamIncome10kC is 0.10 and is statistically significant. In the naive model, the coefficient for FamIncome10k is 0.12 and is also statistically significant. The naive model estimated a slightly stronger effect (0.12) than the final model (0.10), suggesting that controlling for other variables slightly reduces the estimated impact of income. The correlation matrix shows a weak positive correlation that is statistically significant between FamIncome10k and LifeSat (0.24).

SES and Life Satisfaction correlate at 0.17, which is statistically significant, indicating a weak positive association. In the final model, SESC has a significant (but small) effect on LifeSat with a coefficient of 0.01. SES is moderately correlated with FamIncome10k, at 0.40. This being the strongest correlation within the correlation matrix indicates that this is an important bivariate relationship.

RaceBlack has a weak negative correlation (-0.04) with LifeSat, and coupled with its nonsignificant effect on LifeSat in the final model with a coefficient of 0.10 in the final model, belonging to the Black racial category may not significantly predict life satisfaction. RaceBlack has a stronger, but still weak, negative significant correlation with RaceOther (-0.16), which makes sense because they’re both racial groups, but it also has stronger, but still weak, negative significant correlations with FamIncome10k (-0.15) and SES (-0.13), indicating important bivariate relationships. With these correlations being negative, belonging to the Black racial group seems to lower family income and SES.

RaceOther also has a weak negative correlation (-0.03) with LifeSat, and coupled with its nonsignificant effect on LifeSat in the final model with a coefficient of -0.06 in the final model, belonging to the Other racial category may not significantly predict life satisfaction. RaceOther has a stronger, but still weak, negative significant correlation with RaceBlack (-0.16) as was previously mentioned, but it also has stronger, but still weak negative significant correlations with YearsEd (-0.13) and Age (-0.11), indicating important bivariate relationships. With these correlations being negative, belonging to the Other racial group seems to lower years of education and age.

Female has a weak but significant positive correlation (0.05) with LifeSat. FemaleC has a significant effect on LifeSat in the final model with a coefficient of 0.25, indicating that controlling for other variables reveals a clearer effect of sex on life satisfaction. Female has a slightly larger (though still weak) positive correlation with FamIncome10k that is significant (-0.07), indicating an important bivariate relationship in which being female increases family income.

YearsEd has a weak but significant positive correlation (0.12) with LifeSat. However, YearsEd has a nonsignificant effect on LifeSat in the final model with a coefficient of -0.01. The weak positive correlation suggests that YearsEd is associated with LifeSat, but in the final model, the coefficient turns negative, but it is nonsignificant, suggesting that once other variables are controlled for, years of education does not independently predict life satisfaction. YearsEd, however, has a moderate and significant correlation with FamIncome10k (0.38), the second strongest correlation in the correlation matrix, indicating an important bivariate relationship.

Age has a weak significant positive correlation (0.16) with LifeSat. In the final model, Age has a significant (but small) effect on LifeSat with a coefficient of 0.01. Age remains a significant predictor even after controlling for other factors. Age and LifeSat are the strongest bivariate relationship for Age, but Age and SES follow closely behind at a weak but significant correlation of 0.11, indicating another important bivariate relationship.

Overall, the strongest correlations exist between SES and FamIncome10k (0.40) and YearsEd and FamIncome10k (0.38), which are both moderate and significant correlations. These indicate important bivariate relationships.

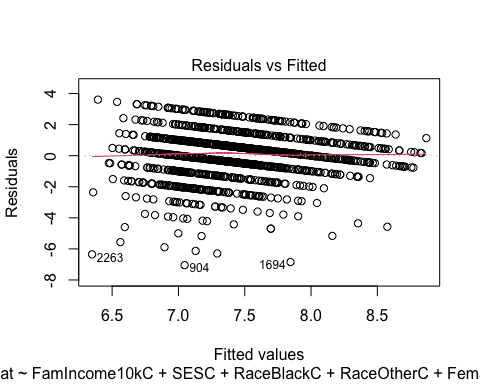
## Question 11

### Test the statistical assumptions underlying your final regression model, and note any concerns.

final\_model\_test <- lm(LifeSat ~ FamIncome10kC + SESC + RaceBlackC + RaceOtherC + FemaleC + YearsEd12 + AgeC + PolViewsC, data=gss, na.action = na.exclude)

### Linearity

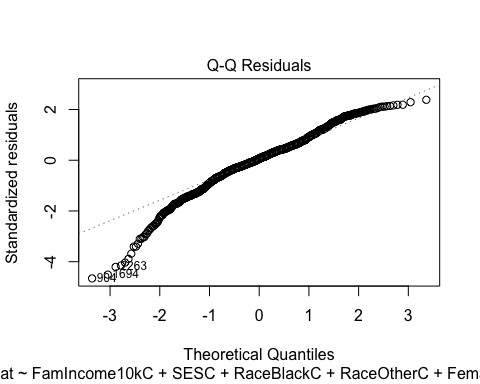
plot(final\_model\_test, 1)



The absence of major bends in the line suggests that the linearity assumption of the regression model is satisfied.

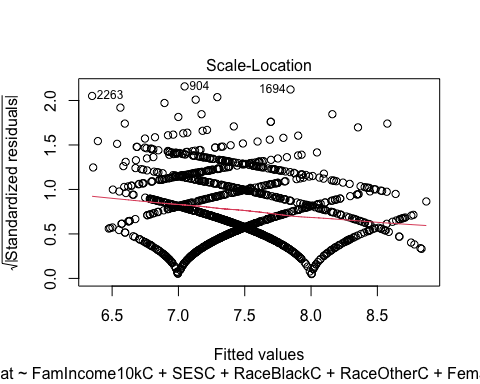
### Normality

plot(final\_model\_test, 2)

 Deviations at the negative tail suggest some non-normality of errors. This may be a point of concern, but moderate deviations are generally considered to be acceptable with large sample sizes like that of the General Social Survey.

### Homoscedasticity

plot(final\_model\_test, 3)



The residuals do not appear to be randomly scattered and there are distinct bands forming, which could indicate issues with the variance of residuals. The line is flat without major bending, but it is not horizontal, as the variance of residuals appears to decrease as fitted values increase. This all suggests potential heteroscedasticity, which may be a point of concern, as this will affect the accuracy of standard errors.

### Independence

table1(~Race, data=gss)

|  | Overall (N=2348) |
| --- | --- |
| **Race** |  |
| White | 1693 (72.1%) |
| Black | 385 (16.4%) |
| Other | 270 (11.5%) |

The General Social Survey uses a complex random sampling technique that involves clustering, which may lead to a violation of the independence assumption. Clustering is common in social science research, however, so this potential violation may not be as much of a concern.

The sample is predominantly White (72.1%), with a mode of White at 1693 observations, but this number is generally reflective of the proportion of White individuals of the U.S. population, where White individuals represent 75.3% of the population (United States Census, 2024). It should be noted that when excluding White individuals who identify as Hispanic or Latino, this number drops to 58.3%, but because the 2018 General Social Survey only included the three racial categories of White, Black, and Other, such nuances may be disregarded. The sample is also 16.4% Black and 11.5% other races collapsed into the category ‘Other’. According to the US Census, Hispanics and Latinos alone make up 19.5% of the U.S. population, and Asians make up another 6.4%, on top of other smaller racial categories, so the ‘Other’ racial group may be undersampled. On the other hand, Black individuals make up 13.7% of the U.S. population, so the General Social Survey may have oversampled the Black racial group. Therefore, the Black racial group appears to be oversampled and the Other racial group appears to be undersampled.

Responses from the Black racial group may thus come from a more homogeneous subset of the population, which can create within-group correlation, meaning observations may not be truly independent, potentially violating the assumption of independence. Responses from the Other racial group may be less representative of the population, leading to underestimated variance. Residuals may be more correlated if they come from similar individuals, potentially violating the assumption of independence. We may thus have reason to be concerned about a violation of the independence assumption, leading to potential inaccuracies in Type I error and Type II error rates due to potential inaccuracies in p-values.

## Question 12

### Explain in plain language what you’ve learned about your research question from the analyses you presented in this exercise.

Higher income appears to be linked to higher life satisfaction, but the effect is relatively small. In the final model, for every $10,000 increase in family income, life satisfaction rises by about 0.10 points on a 10-point scale when holding other variables constant. While this is statistically significant, it suggests that income is just one of many factors influencing life satisfaction.

While having a higher SES is associated with slightly greater life satisfaction, its effect is minimal. The correlation between SES and family income was the strongest in the correlation matrix, suggesting that higher family income is closely linked to higher SES, reinforcing the idea that income plays a crucial role in shaping overall socioeconomic standing. However, both of these factors have small, but significant, effects on life satisfaction.

Race does not significantly change the relationship between income and life satisfaction. Although Black and “Other” racial groups have lower average incomes, this doesn’t appear to strongly affect their overall life satisfaction in a statistically significant way.

Female was statistically significant throughout the models, consistently showing a positive effect on life satisfaction. This means that, on average, women reported higher life satisfaction than men, even when controlling for other factors like income, socioeconomic status, education, race, and age. However, sex may not significantly change the effect of income on life satisfaction. The nonsignificant interaction term between the variables representing sex and family income suggests that the effect of income on life satisfaction may not differ by gender, meaning that the effect of income on life satisfaction may not differ significantly between men and women. The main effect of female being significant throughout suggests that gender itself, independent of income, may play an important role in shaping life satisfaction. There could be several explanations for this, such as differences in social support networks, coping mechanisms, or expectations women have regarding life satisfaction. This finding reinforces that while income plays a role, social factors like gender also have a meaningful impact on life satisfaction.

Education does not appear to strongly or significantly predict life satisfaction. While higher education levels might lead to better job prospects, they may not directly translate to greater happiness.

Age appears to have a positive significant effect on life satisfaction, though this effect is not very strong. This means that older individuals may tend to report very slightly higher life satisfaction than younger individuals.

The implications of these analyses are that income matters for life satisfaction, but not as much as we might assume. Other social factors (such as gender, SES, and age) also play a role that is potentially separate from income. The fact that higher income doesn’t have a large impact on life satisfaction aligns with past research suggesting that beyond a certain point, money doesn’t necessarily buy happiness. Disparities in income, particularly among racial groups, exist, but they don’t seem to directly lower life satisfaction in a significant way. This suggests that other factors, like social support or resilience, might buffer the impact of lower income. Overall, while earning a higher income can slightly boost happiness, it’s not the biggest predictor of life satisfaction. Social conditions, identity factors, and life circumstances all contribute to how satisfied people feel with their lives.

## Question 13

### Create a visualization that presents a key finding from one of your regression analyses described above.

ggplot(gss, aes(x = FamIncome10k, y = LifeSat, color = Sex)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 facet\_wrap(~ Race) +  
 xlim(0,13) +  
 ylim(0,13) +  
 scale\_color\_manual(values = c("Male" = "deepskyblue2", "Female" = "hotpink")) +  
 labs(  
 title = "Life Satisfaction by Income, Sex, and Race",  
 x = "Family Income (in 10k)",  
 y = "Life Satisfaction Score",  
 color = "Sex")



## Bibliography

Boyce, C. J., Brown, G. D. A., & Moore, S. C. (2010). Money and Happiness: Rank of Income, Not Income, Affects Life Satisfaction. *Psychological Science, 21*(4), 471-475.

Easterlin, R. A. (1973). Does Money Buy Happiness? *The Public Interest*, 30, 3.

Lane, R. E. (1993). Does money buy happiness? *Public Interest*, 56.

QuickFacts United States. U.S. Census Bureau. (2024). <https://www.census.gov/quickfacts/fact/table/US/PST045224>