ANA600 Final Exam

Charlene Wolthers

2024-02-04

INSTRUCTIONS

Perform basic exploratory data analysis (EDA) for the final exam. EDA consists of the procedures and concepts we have practiced throughout the course. You will be exploring your dataset, reviewing the variables, modeling an variable of interest, and interpreting results. Each section requires a complete paragraph explaining and interpreting the results within the R-Markdown above the code block required to analyze that section. All writing and code should be written as if presenting a report to your supervisor.

Provided data was selected to gather information about consumer attitudes and decisions to save, borrow, or make purchases. The researchers wanted to forecast changes in consumer behavior in the United States and better understand consumer confidence about personal finances, employment, price changes, and the perceived state of national business. Your job is to develop a model of income by sex to determine if there is a statistically significant difference between the two.

Below is a list of the variables included in the dataset:

- household: including yourself, how many members of your household are 18 years or older?
- kids: How many members of your household are 17 years or younger?
- vehicles: How many vehicles do you use, including those leased, owned, or provided by your employer for personal use?
- priceExpected: By what percent do you expect prices to go up/down on average during the next 12 months?
- incomeExpected: By what percent per year do you expect your (family) income to increase/decrease during the next 12 months?
- businessExpected: Considering business conditions in the country as a whole, do you think that during the next 12 months, we'll have good or bad times financially?
- financialStability: Would you say that you (and your family) are better off or worse off financially than you were a year ago?
- investments: What type of investments do you plan to add or shift money into or open during the next three months: mutual funds, savings accounts, stocks, bonds, retirement accounts?
- income: What is your household gross income (in thousands)?
- age: What is your age?
- employmentSector: What is your employment sector?
- region: What is your state of residence? (recoded into regions)
- hoursPerWeek: Approximately how many hours a week are you employed?

PREPARATION (10 Points)

Import the required file to a new dataframe and load necessary libraries

INTRODUCTION AND RESEARCH QUESTION (5 Points)

1. Introduce your purpose and scope, creating a story for the data generation process that might be responsible for the variation in income output variable.
2. Describe the research question, which is to examine income based on sex.

The primary purpose of this report is to conduct a comprehensive exploratory data analysis (EDA) on a dataset comprising of consumer attitudes and financial behaviors. The aim is to unravel the underlying patterns and correlations that might influence consumer decisions related to saving, borrowing, and making purchases. Variables in this dataset include a review of key variables such as sex, age, marital status, income, hours worked per week, race, citizenship status, health insurance coverage, and primary language. These variables were chosen for their potential impact on and relevance to consumer financial decisions and overall economic participation.The data generation process (DGP) responsible for the variation in the income output variable within the context of consumer behavior and demographics can be complex, involving multiple layers of influences ranging from individual characteristics to broader economic factors. Individuals enter the workforce with varying levels of education and skill sets, influenced by their age, sex, race, and language proficiency. Marital status and the decision to have children (not directly measured but inferred from marital status and age) introduce variations in income due to potential career breaks, the need for part-time employment, or the choice to prioritize one partner's career over the other. These initial conditions set the stage for their earning potential. In addition to providing insights into consumer behavior, a specific objective of this analysis is to develop a predictive model based on the dataset to assess whether there is a statistically significant difference in income between different sexes.

The research question at the core of this analysis is to examine the relationship between sex and income within the context of a comprehensive dataset capturing consumer attitudes and financial behaviors. Specifically, the investigation seeks to determine whether there is a statistically significant difference in income between males and females, and how such differences might reflect broader trends and disparities in economic participation and financial decision-making. To explore the potential gender disparity in income, the analysis will employ statistical modeling techniques to examine the relationship between sex and income while controlling for other variables in the dataset. This approach will enable us to isolate the effect of gender on income from other confounding factors, providing a clearer picture of any existing disparities. ***

QUESTION #1 (10 Points)

1. Enter code to produce the structure of your dataframe
2. Recode the Sex variable to 0=Female, 1=Male
3. Produce a crosstab table of observations for the race and sex variables
4. View the top five records in the dataframe
5. Write one paragraph describing the structure of the data frame and interpreting the produced table

The dataset under analysis comprises 1,000 observations across 9 variables: sex, age, married status, income, hours worked per week, race, US citizenship status, health insurance status, and primary language spoken. A specific table produced from this dataset provides a tally count of race by sex, facilitating an examination of demographic distributions within the dataset. In this table, the column values for sex are encoded as 1 and 0, where 0 represents female and 1 represents male. The rows categorize individuals by race, offering insights into the racial composition of the sample stratified by sex. The encoding of sex as binary values and the categorization of race into distinct rows serve to simplify the analysis, making it possible to discern patterns and disparities with clarity.

```
str(acsData)
```

```
## 'data.frame':    1000 obs. of  9 variables:
##  $ Sex            : int  0 1 1 0 1 1 1 1 0 0 ...
```

```
##  $ Age            : int  31 31 75 80 64 14 78 35 70 18 ...
##  $ Married        : int  0 0 0 0 1 0 1 0 1 0 ...
##  $ Income         : num  60 0.36 0 0 0 ...
##  $ HoursWk        : num  40 12 40 13.2 32.7 ...
##  $ Race           : chr  "white" "black" "white" "white" ...
##  $ USCitizen      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ HealthInsurance: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Language       : int  1 0 0 0 0 0 0 1 0 0 ...
```

```
recode(acsData$Sex, "Female" = 0, "Male" = 1)
```

```
## Warning in recode.numeric(acsData$Sex, Female = 0, Male = 1): NAs introduced by
## coercion
```

```
## Warning: Unreplaced values treated as NA as `.x` is not compatible.
## Please specify replacements exhaustively or supply `.default`.
```

```
##    [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##   [25] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##   [49] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##   [73] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##   [97] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [121] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [145] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [169] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [193] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [217] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [241] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [265] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [289] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [313] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [337] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [361] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [385] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [409] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [433] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [457] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [481] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [505] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [529] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [553] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [577] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [601] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [625] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [649] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [673] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [697] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [721] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [745] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [769] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [793] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [817] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [841] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

```
##  [865] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [889] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [913] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [937] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [961] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [985] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

```
tally(Race ~ Sex, data = acsData)
```

```
##        Sex
## Race      0   1
##   asian  30  40
##   black  60  46
##   other  33  30
##   white 414 347
```

```
head(acsData, 5)
```

```
##   Sex Age Married Income  HoursWk  Race USCitizen HealthInsurance Language
## 1   0  31       0  60.00 40.00000 white         1               1        1
## 2   1  31       0   0.36 12.00000 black         1               1        0
## 3   1  75       0   0.00 39.99126 white         1               1        0
## 4   0  80       0   0.00 13.15004 white         1               1        0
## 5   1  64       1   0.00 32.71688 white         1               1        0
```

QUESTION #2 (10 Points)

1. Recode the income variable to value x 1,000
2. Calculate the minimum, maximum, mean, median, IQR, and range for income
3. Calculate the mean of income each for males and for females
4. Write one paragraph explaining and interpreting the descriptive statistics

The descriptive statistics for the Income variable depict a distribution with a significant range and variability. The minimum income reported in the dataset is $0. This suggests that there are individuals in the sample with no reported income. The first quartile, or the 25th percentile, is also $0. This indicates that at least 25% of the individuals in the dataset have no income, reinforcing the presence of a significant portion of the sample without reported earnings. The median income is $13,000. This means that half of the individuals earn less than $13,000, and the other half earn more, providing a central point of the income distribution. The third quartile, or the 75th percentile, is $31,841.09. This shows that 75% of the individuals earn less than this amount, indicating the upper limit of income for the majority of the sample. The maximum reported income is $563,000, which shows the highest income level within the dataset. The average income across all 1,000 observations is $22,785.13. The presence of individuals with no income and the widespread between the minimum and maximum values suggest diverse economic conditions among the respondents. The skewness towards higher incomes, as indicated by the difference between the mean and median, highlights the presence of economic disparities within the dataset. Additionally, the difference between the mean incomes for males and females underscores a gender-based disparity in earnings within the dataset. The average income for males is reported to be $28,573.61. The average income for females is significantly lower, at $17,794.32. Males, on average, earn significantly more than females.

```
acs_Data <- acsData
acs_Data$Income <- acsData$Income * 1000
head(acs_Data)
```

```
##   Sex Age Married   Income  HoursWk  Race USCitizen HealthInsurance Language
## 1   0  31       0 60000.00 40.00000 white         1               1        1
## 2   1  31       0   360.00 12.00000 black         1               1        0
## 3   1  75       0     0.00 39.99126 white         1               1        0
## 4   0  80       0     0.00 13.15004 white         1               1        0
## 5   1  64       1     0.00 32.71688 white         1               1        0
## 6   1  14       0 37215.49 26.64792 white         1               1        0
```

```
fav_stats(acs_Data$Income)
```

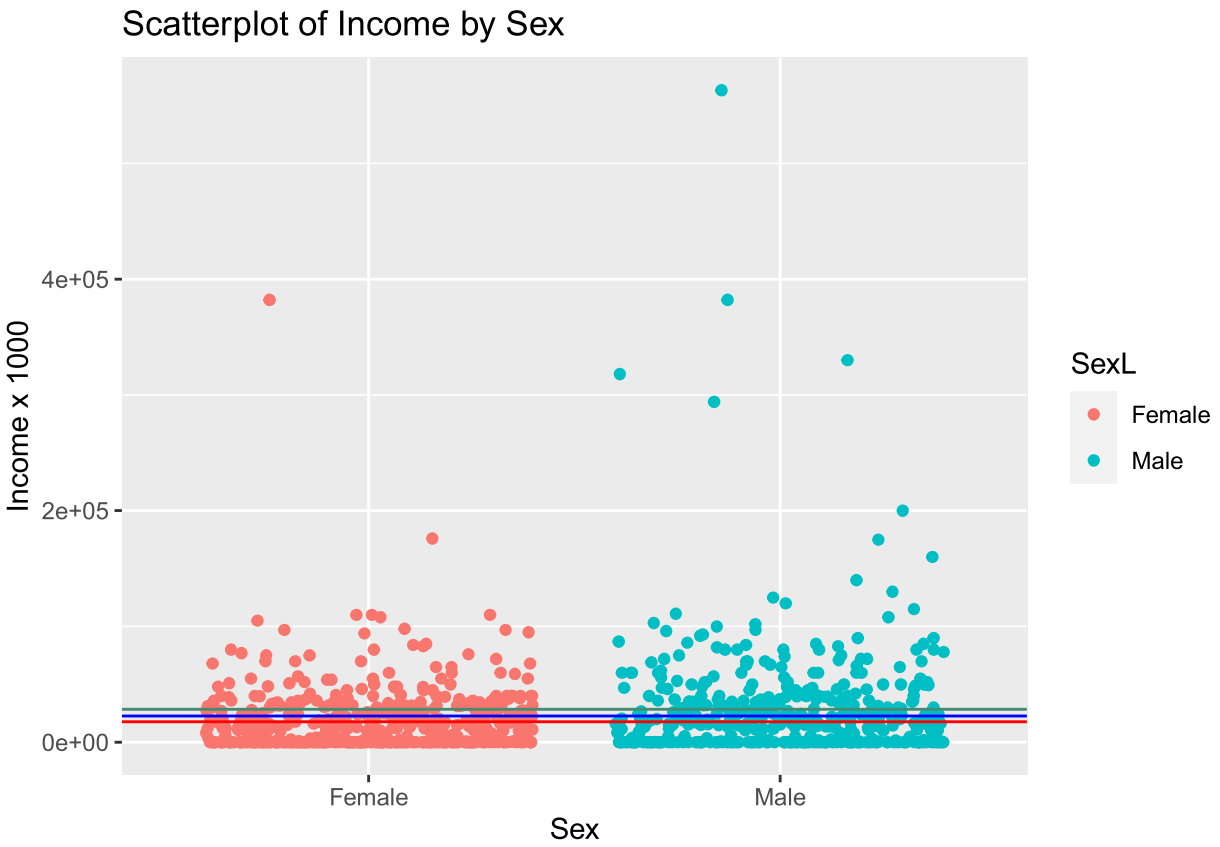
```
##  min Q1 median       Q3    max     mean       sd    n missing
##    0  0  13000 31841.09 563000 22785.13 39141.88 1000       0
```

```
acs_DataFemales <- filter(acs_Data, Sex == 0)

 acsData_Female_mean <- mean(acs_DataFemales$Income)

 acsData_Males <- filter(acs_Data, Sex == 1)

acsData_Male_mean <- mean(acsData_Males$Income)
```

QUESTION 3 (10 Points)

1. Create a new variable SexL such that Sex = 0 is "Female", Sex = 1 is "Male", and else is "Undefined"
2. Create an appropriate visualization for income and the new SexL variable
3. Write one paragraph explaining and interpreting the visualization

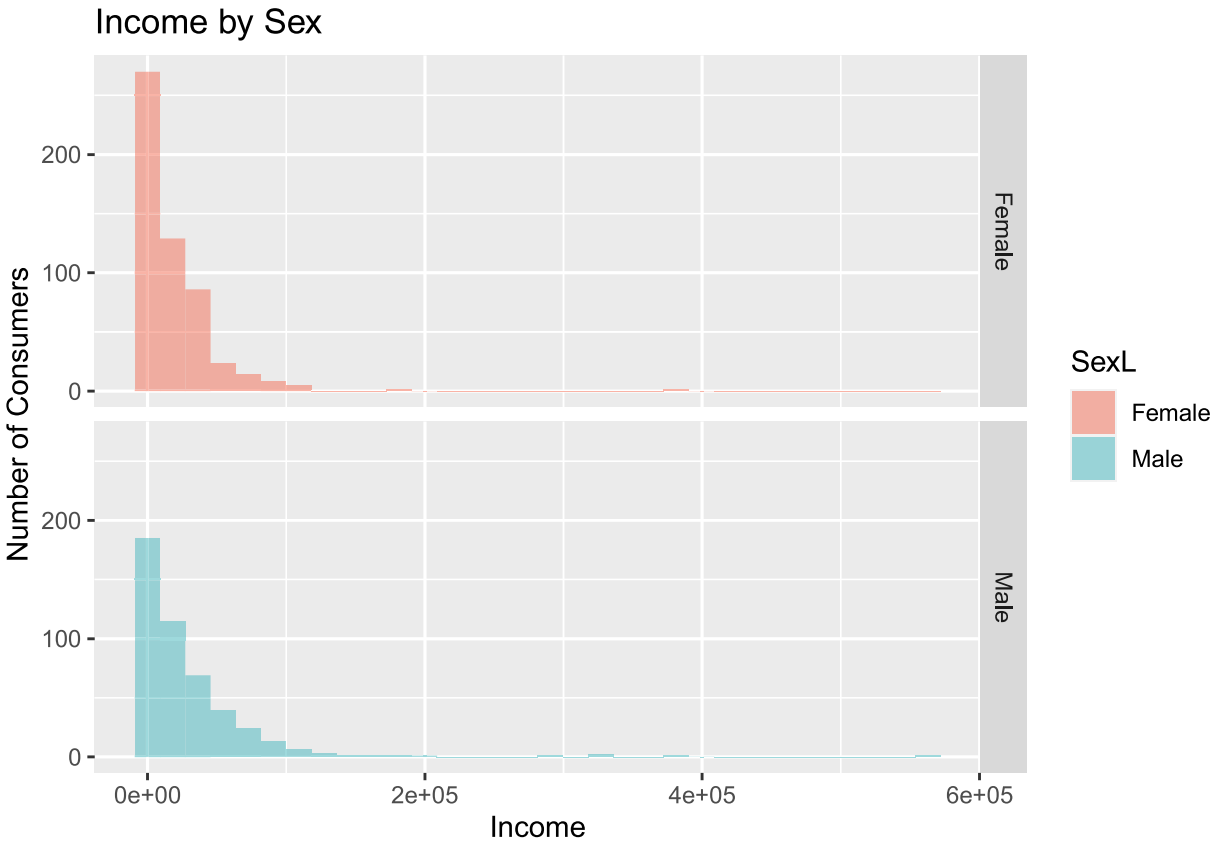
The visualization presents a scatter plot that contrasts income distribution between females and males. It is immediately apparent that income points for females, represented in red, are generally lower on the scale compared to those for males, indicated in aquamarine. The red line illustrates the mean income for females, while the aquamarine line represents the mean income for males, clearly showing the income disparity where males, on average, earn more than females. The blue line represents the grand mean income for the entire population under study, regardless of sex. This line falls closer to the mean income for females, suggesting that the overall population mean is more influenced by the female income data, possibly due to a larger number of female observations or the distribution of income among females. The scatter plot also shows a wider spread of income values for males, as indicated by the greater vertical dispersion of aquamarine points, suggesting more variability in male income. Additionally, there are outliers visible for both sexes, with extreme values that deviate substantially from the respective group means, indicating the presence of individuals with incomes significantly higher than the average for their sex.



QUESTION 4 (10 Points)

1. Create a histogram of income by sex with facet grid
2. Write one paragraph explaining and interpreting the visualization

The visualization is a faceted histogram showing the distribution of income segmented by sex. The two facets allow for a direct visual comparison between the income distributions for females (in pink) and males (in blue). In both histograms, the x-axis represents income, and the y-axis represents the number of consumers. For females, the distribution appears to be right-skewed, with a concentration of observations in the lower income brackets and a tail stretching towards higher incomes. Most female consumers fall within the lower income range, as indicated by the height of the bars on the left side of the female histogram. The male income distribution displays a similar right-skewed pattern but with a slightly wider spread across income ranges. The bars in the male histogram suggest that males are more evenly distributed across different income levels, including higher income brackets, compared to females. Both distributions taper off as income increases, with fewer consumers at the higher income levels, which is typical of income data. However, the peak of the male distribution is less pronounced than that of the female distribution, suggesting that income is more varied among males.



QUESTION 5 (10 Points) 1. Create a model of income for females 3. Write one paragraph explaining and interpreting the model

The female model represents a linear regression analysis where the income of females is modeled without any predictors—essentially, it is an intercept-only model. In this case, the formula Income ~ NULL indicates that no variables are being used to predict income; instead, the model only calculates the mean income for females in the dataset acs_DataFemales. The coefficient for the intercept is given as $17,794, which can be interpreted as the average (mean) income for females within this particular dataset. This figure serves as a baseline against which the impact of additional variables on female income could be measured if further predictors were to be included in the model. Essentially, the model suggests that if we do not take any other variables into account, a typical female from this dataset can be expected to have an income of $17,794. This basic model sets the groundwork for more complex modeling that could incorporate other factors to explain variations in income among females.

QUESTION 6 (20 Points)

1. Create a model of income by sex
2. Write one paragraph explaining and interpreting the model

The model is a linear regression model that aims to predict income based on the variable SexL. In this model, the Income ~ SexL formula indicates that income is being analyzed in relation to the sex of individuals in the acs_Data dataset. The intercept, which is $17,794, represents the average income for the baseline category, which, by default, is the category not represented by the dummy variable—in this case, females. This means that when the sex of an individual is female (when SexLMale is 0), the predicted income is $17,794. The coefficient for SexLMale is $10,779, which represents the average difference in income between males and females. More specifically, this positive coefficient suggests that males earn, on average, $10,779 more than females in this dataset.

QUESTION 7 (20 Points)

1. Calculate the predicted value and residual value for each observation using the income by sex model
2. Calculate the sum of squared deviations and sum of absolute deviations
3. Write one paragraph explaining and interpreting the results

The predicted value for each observation is the income estimated by the model based on the individual's sex. In this case, the model includes an intercept and a coefficient for sex. The predicted value for females would be the intercept ($17,794), and for males, it would be the intercept plus the coefficient for sex ($17,794 + $10, 779 = $28,573). In this scenario, the intercept represents the average income for females, and the coefficient for sex quantifies the average additional income for males compared to females. Residuals are the differences between the actual observed incomes and the incomes predicted by the model for each individual. A residual is positive if the actual income is higher than the predicted income and negative if it is lower. The residuals can thus be interpreted as the individual deviations from the average income pattern described by the model. For example, a residual of -28,573 in the context of this regression model represents the difference between the observed income for a particular data point and the value predicted by the model. This negative residual indicates that the model's prediction was higher than the actual observed value by $28,573. The sum of absolute deviation (SAD) and sum of squared deviations (SSD) are measures of variability or dispersion in a dataset. The SAD of 22,270,304 represents the cumulative absolute difference between each observed value and the mean. This value indicates the total "error" or deviation from a central point. On the other hand, the SSD, presented in scientific notation as 1.501666e+12, equates to 1,501,666,000,000 when written out fully. This value signifies the aggregate of each observation's squared difference from the mean, amplifying the impact of outliers due to the squaring of each deviation. A large SSD, such as this, suggests considerable variability in the dataset and possibly the presence of extreme values that diverge significantly from the mean.

```
predict_sex <- predict(sex_model)
resid_sex <- resid(sex_model)

SAD_Sex_model <- sum(abs(resid(sex_model)))
SS_sex_model <- sum(resid(sex_model)^2)
```

QUESTION 8 (20 Points)

1. Run an analysis of variance on the model of income by sex
2. Write one paragraph explaining and interpreting the results Hint:

- H0: bi = 0
- Ha: bi <> 0

The ANOVA table presents the results of a linear regression analysis examining the effect of sex (denoted as SexL) on income. The 1 degree of freedom (df) indicates that sex explains a portion of the variance in income. The F statistic is 19.2, which is a measure of the ratio of variance explained by the model to the variance within the residuals. The p-value associated with the F statistic is .0000, indicating that the results are statistically significant at conventional significance levels (typically $p < .05$). This leads to the rejection of the null hypothesis, which posits that sex has no effect on income, in favor of the alternative hypothesis, which suggests that sex does have an effect on income. In essence, these results suggest that there is a statistically significant difference in income based on sex.

```
supernova(sex_model)
```

```
##  Analysis of Variance Table (Type III SS)
##  Model: Income ~ SexL
##
##                                  SS  df              MS      F   PRE     p
##  ----- --------------- | ------------ --- --------------- ------ ----- -----
##  Model (error reduced) | 2.888920e+10   1 28889204142.860 19.200 .0189 .0000
##  Error (from model)    | 1.501666e+12 998  1504675056.121
##  ----- --------------- | ------------ --- --------------- ------ ----- -----
##  Total (empty model)   | 1.530555e+12 999  1532086997.149
```

QUESTION 9 (10 Points)

1. Calculate the proportional reduction in error
2. Write one paragraph explaining and interpreting the improvement of the linear model by adding sex Hint:

- H0: bi = 0
- Ha: bi <> 0

The inclusion of sex as a predictor in the linear model has led to a slight improvement in the model's ability to explain variability in income, as indicated by the Proportion Reduction in Error (PRE) of 0.0189 or 1.89%. This metric means that adding sex as a variable reduces the error in predicting income by nearly 2% compared to a model without any predictors. In terms of hypothesis testing, the null hypothesis would assert that sex has no effect on income (the coefficient for sex is equal to zero), and the alternative hypothesis would contend that sex does have an effect on income (the coefficient for sex is not equal to zero). The PRE value, although modest, signifies that sex is indeed a relevant factor in determining income, providing evidence to reject the null hypothesis in favor of the alternative. However, the relatively small PRE value also indicates that while sex is statistically significant, it accounts for only a small portion of the total variation in income, suggesting that there are other variables not included in the model that could further explain the differences in income.

```r
sse_sex <- sum(sex_model$residuals^2)
sst_sex <- sum((acs_Data$Income - mean(acs_Data$Income))^2)
ssm_sex <- sst_sex - sse_sex
pre <- ssm_sex / sst_sex
print(pre)
```

```
## [1] 0.01887499
```

QUESTION 10 (20 Points)

Write one paragraph interpreting and concluding the results of your analysis.

The analysis conducted on the relationship between sex and income through linear regression and ANOVA has yielded significant insights. The statistical tests have confirmed that sex is a statistically significant predictor of income, as evidenced by the F statistic and the associated p-value, indicating a clear rejection of the null hypothesis that sex has no effect on income. However, the Proportion Reduction in Error (PRE) value of 1.89% suggests that while sex contributes to the variability in income, it accounts for a relatively small fraction of the total variance. This implies that other factors, beyond sex, play a more substantial role in determining income levels. The modest PRE value also underscores the complexity of income determination. However, the presence of a statistically significant difference in income between sexes highlights the need for further exploration into the causes of this disparity.

END OF FINAL EXAM