The Effects of Inheritance, Industry, and Age on Billionaires' Net Worths

Jayden Lewis

Introduction and Data

On April 2nd, 2024, Forbes reported that there are 2,781 billionaires in the world. Among this small and unique population, there are a variety of net worths, making some billionaires wealthier than others. This project aims to analyze some of the variables that attribute to a billionaire's net worth and draw conclusions on how we may be able to predict who is the biggest bread winner.

To Explore this topic, we will be using a csv file from the "CORGIS Data Set Project," created by Ryan Whitcomb on May 17th, 2016 titled "billionaires." Scholars at Peterson Institute for International Economics have built off the Forbes World's Billionaires list from 1996-2014 and added more variables. The data set has 2614 observations and 22 columns. Each observation represents 1 of the 2,614 billionaires from Forbes's list and the 22 columns include numeric variables such as the billionaire's age, the year that that the company was founded, the billionaire's net worth, and the "Gross Domestic Product" of the country where the billionaire has citizenship. Some categorical variables include company name, the billionaire's relationship to the company,the country that the billionaire is a citizen of, the industry the billionaire is in, and the way that the money was inherited.

The general research question we will be answering is how do we expect inheritance status, company industry, and current age (as of May 17th, 2016) to impact a billionaire's net worth? To help us answer this question, we have created three sub questions that are relevant to the variables:

1. Is there is a relationship between billionaires who did or did not inherit their companies and how much they are worth?
2. Is there a relationship between the industry a billionaire is in and how much they are worth? these industries, does the same effect of inheritance remain constant?
3. Is there a relationship between a billionaire's age and how much they are worth?

Our response variable (numeric) is the number of billion of dollars that the billionaire is worth (3.5 for example), represented in the data as "wealth.worth.in.billions".

The first explanatory variable is inheritance status (categorical), originally represented in the data as wealth.how.inherited. Within this variable, the are the categories "not inherited," "spouse/widow," "father," 3rd generation," "4th generation," and "5th generation or longer." However, because our research questions is specifically looking at weather the billionaire inherited their wealth or not, we have created a new variable called "inheritance" with categories "Inherited" and "Not inherited." In our model, the reference variable will be "Inherited."

The second explanatory variable is the industry that the billionaire is in (categorical), originally represented in the data as wealth.how.industry. Within this variable, there are 18 different industries, some of which overlap. We also found that 16 observations had a "0" under this variable, representing that the industry was not reported/was unknown. To help efficiently answer our research question, we created a new variable called "industry" which grouped the industries into "Technology," "Finance," "Consumer/Retail," and "Other." We also renamed "Other" to "Aother" so that when analyzing the industries within our model, it is in reference to Other industries.

The final explanatory variable is the current age (numeric) of the billionaire as of May 17th, 2016 (when the data set was created). This variable is represented as demographics.age in the data set. Because 383 billionaire's ages were not available when the data was collected, and were represented with a "0", we filtered the variable to include all numbers greater than 0.

We put these new and filtered variables into a new data set titled "project_variables," which we will use to create all plots, models, and summary statistics. The data set we will be working with now has 2229 observations.

For each sub question, we have established a null and alternative hypothesis:

1. Null: the true slope for inheritance status is zero - there is no significant linear relationship between inheritance and net worth

   Alternative: the true slope for inheritance status is not zero - there is a significant linear relationship between inheritance and net worth

2. Null: the true slope for company industry is zero - there is no significant linear relationship between industry and net worth

   Alternative: the true slope for company industry is not zero - there is a significant linear relationship between industry and net worth

3. Null: the true slope for age is zero - there is no significant linear relationship between age and net worth

   Alternative: the true slope for age is not zero - there is a significant linear relationship between age and net worth

Here are the following relevant plots and summary statistics to asses the relationships between net worth and inheritance, industry, and age.

Figure 1.1

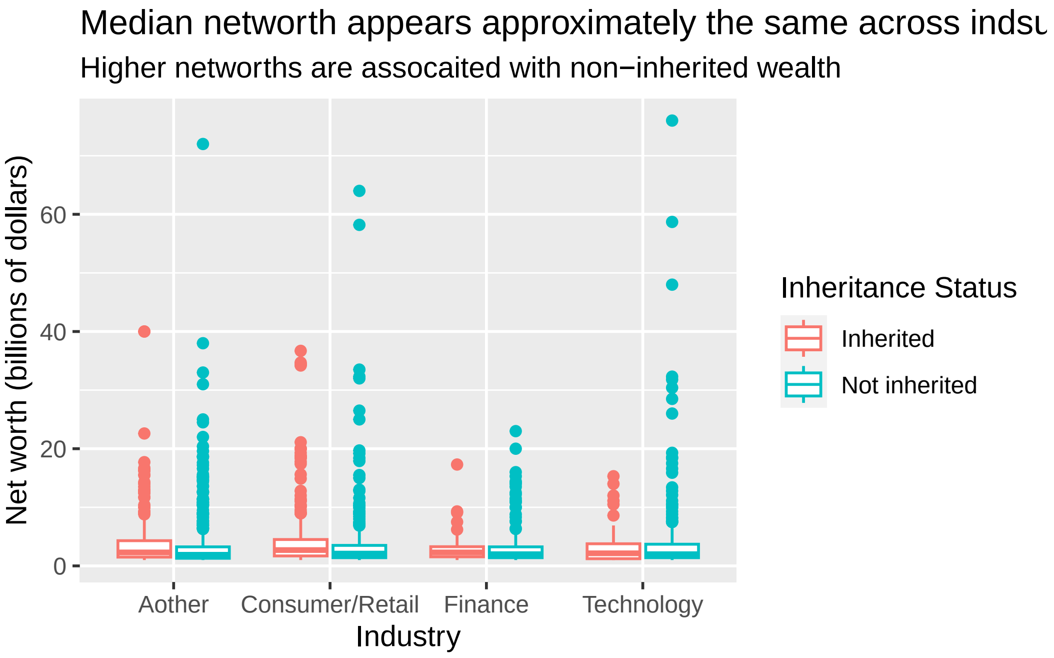


Figure 1.2

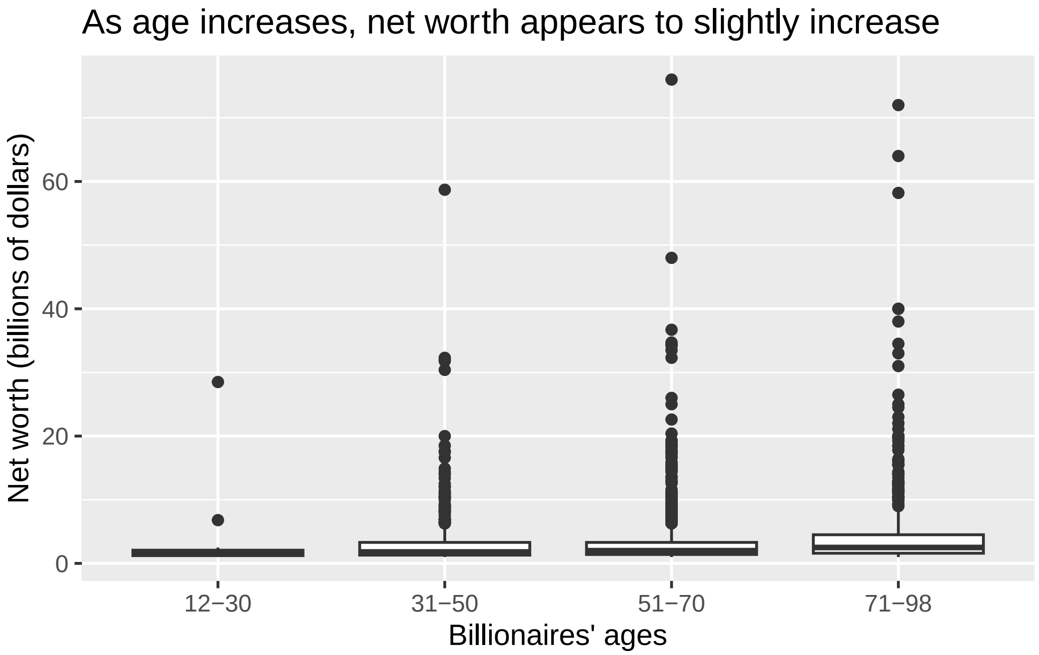


Figure 1.3

```
  total_median total_mean
1          2.1   3.723329
```

Figure 1.4

```
# A tibble: 2 x 3
  inheritance   meadian  mean
  <chr>           <dbl> <dbl>
1 Inherited         2.4  3.96
2 Not inherited     2    3.60
```

Figure 1.5

```
# A tibble: 4 x 3
  industry        median  mean
  <chr>            <dbl> <dbl>
1 Aother             2.1  3.47
2 Consumer/Retail    2.3  4.15
3 Finance            2    3.13
4 Technology         2    4.32
```

Methodology

Our research questions aims to answer whether or not there is a linear relationship between our predictors and net worth. Therefore, we will be using a linear regression model. Net worth is our continuous response variable and we are predicting that there are linear relationships between inheritance, industry, and age. Since we are not categorizing net worth into "high" and "low" or into "low," "medium," "high," and "extreme" to predict the probability of a billionaires falling into these levels, a logistic or ordinal regression model is not appropriate. Within inheritance, our model's reference variable will be billionaires who inherited their wealth. Therefore our model will predict the relationship between non-inherited wealth and net worth in comparison to inherited wealth based on the slope coefficient associated with "inheritance". Within industry, our model's reference variable will be billionaires in industries other than tech, finance, and consumer/retail. Therefore our model will be predicting the relationships between Technology and net worth; Finance and net worth and Consumer/Retail and net worth in comparison to other industries based on the slope coefficients associated with "industry". Within age, our model will predict relationship between age and net worth based on the slope coefficient associated with "demographic.age."

We also decided to add an interaction term between inheritance and industry (inheritance*industry). The interaction term allows us to investigate whether the effect of inheriting a company on a billionaire's net worth varies across different industries. For example, the impact of inheritance might be more pronounced in industries like finance or technology compared to consumer/retail. By including an interaction term, our model can capture these nuances instead of assuming that the inheritance effect is uniform across all industries. This will lead to more accurate predictions of net worth.

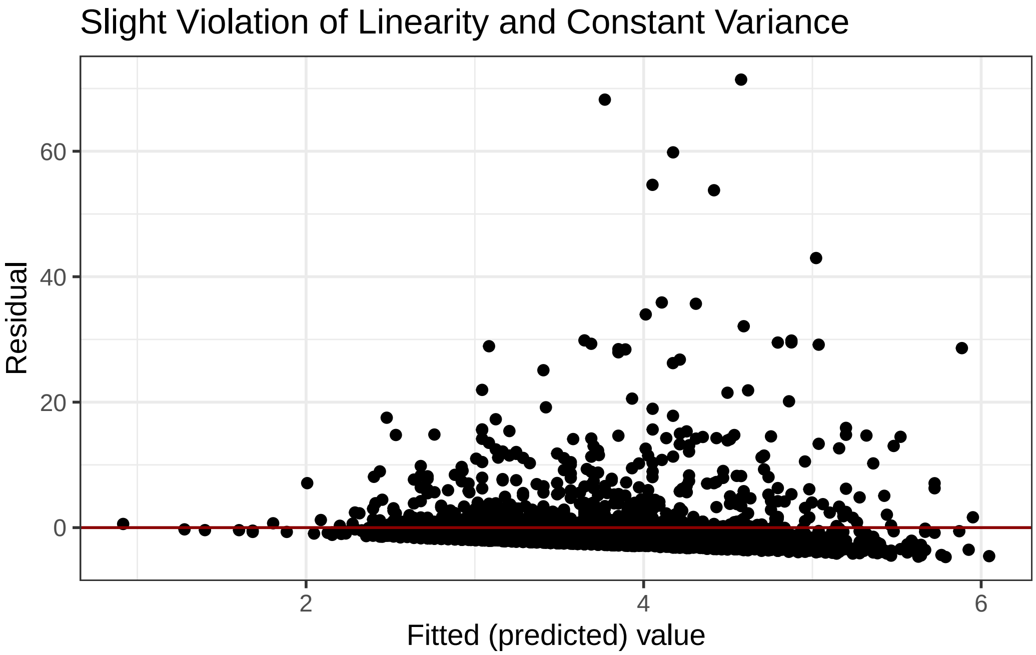
Furthermore, we considered whether or not Independence, Linearity, Constant Variance, and Normality are violated.

Independence

The independence condition is not violated. Although these billionaires are different ages and either started or inherited their companies at different times, all of their net worth were collected at the same time, and all of their companies were making money at the time the data was collected. Time would not confound net worth in this scenario. At the time the net worths were collected, all companies were experiencing the same economic factors and factors relevant to advancements or regressions in society/economy.

Linearity & Constant Variance

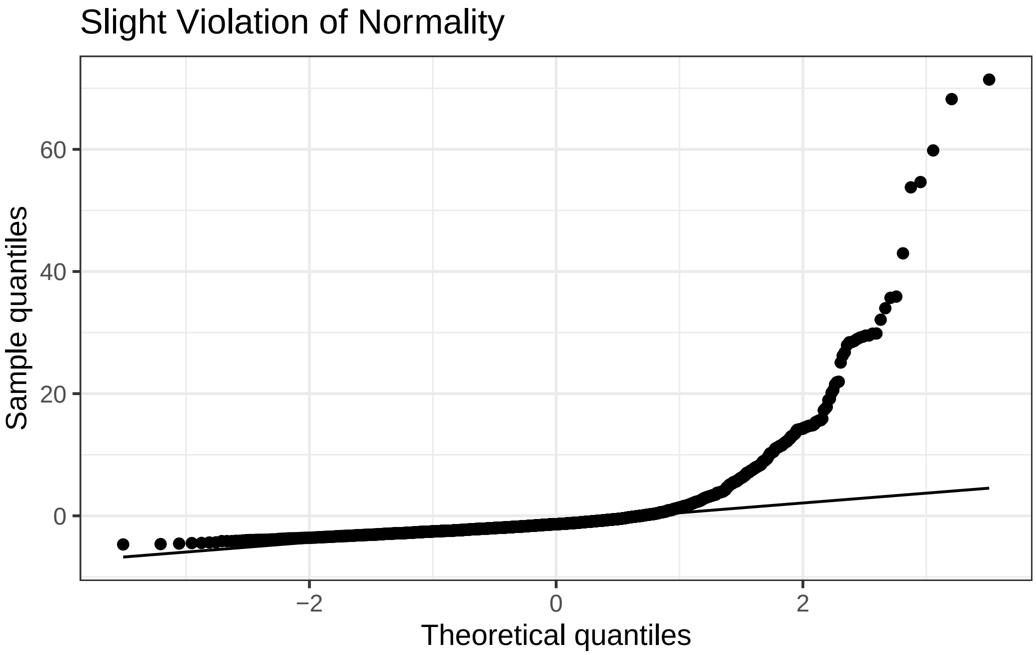
Figure 2.1



In the above residual plot, we see that the majority of the points are clustered right around the horizontal axis, suggesting that the linearity is achieved. The plot is a little misleading because there 2,229 points, and although some are not symmetrically distributed with the main cluster, this is a small number of outliers compared to all of the points. We believe that enough of the points are symetrical distributed to deem that linearity is achieved. Again with constant variance, a majority of the points are right on top of each other and therefore evenly space in the vertical sense. Some of the points are further spaced apart, but this is small number in comparison to all of the 2,299 points.

Normality

Figure 2.2



According to the above Q-Q plot, up until the last third of the plot, normality was achieved. However, it appears that a little less than 1/3 of the points clearly deviate from the linear line, suggesting that normality may be violated. Because these points deviated points fall closer to the tail than the majority of the linear line, it is difficult to determine whether or not linearity is fully achieved. However, we do not see enough evidence to deem that normality has been completely violated.