# sevenr: the billionaire project

Koji Bilbao, Maya Markus-Malone, Jacob Rosenzweig, Scott Tremblay, Michael Wang

# The dataset

Organization: Peterson Institute for International Economics

Authors: Caroline Freund and Sarah Oliver

Date created: 2016

How: publicly available information, company websites, and Forbes lists

# The dataset

22 variables, 2614 observations

Each observation is a billionaire

Variables include identifying variables (name, year, company name, etc.)

Other variables are more interesting (wealth, industry, inheritance, etc.)

| Variable name (unmodified) | Meaning |
| --- | --- |
| company.relationship | relationship to the company (founder, etc.) |
| company.sector | the sector of the economy to which the business belongs |
| company.type | type of business of the company |
| demographics.age | age of the individual |
| location.citizenship | name of the country for which the individual has citizenship |
| location.country.code | 3-letter country code for which the individual has citizenship |
| location.gdp | gross domestic product of the country for which the individual has citizenship |
| location.region | part of the world where the individual lives |
| *wealth* | *net worth of the individual in billions of USD* |
| wealth.how.category | where the individual's money came from |
| wealth.how.industry | the specific industry the individual profited from |
| wealth.how.inherited | whether the individual's wealth was inherited or not |
| wealth.how.was.founder | whether the individual founded their company or not |
| wealth.type | the type of billionaire that they are |

# The question and hypothesis

We want to investigate the relationship between a billionaire's absolute wealth and the other variables.

What is the best model for predicting wealth?

Our hypothesis is that our model will include industry type, GDP of location, inheritance, and age as variables.

# Data cleaning

Before using the dataset to construct a model, many of the variables and their values had to be cleaned

For example, company.relation (CEO, founder, etc.) had many values with varying capitalizations, iterations of different positions, etc. that had to be fixed

For other variables, NA values and 0s that didn't make sense had to be removed

# Data cleaning

Another problem was that data points were collected at 3 different years (1996, 2001, 2014).

We chose to filter for the year 2001 as it is the most recent year for which GDP information is available.

The end result of cleaning is a smaller dataset with 413 observations.

```
Rows: 413
Columns: 24
$ name                      <chr> "Bill Gates", "Warren Buffett", "Paul Allen",…
$ rank                      <int> 1, 2, 3, 4, 6, 7, 8, 9, 10, 10, 12, 13, 14, 1…
$ year                      <int> 2001, 2001, 2001, 2001, 2001, 2001, 2001, 200…
$ company.founded           <int> 1975, 1962, 1975, 1977, 1980, 1962, 1962, 196…
$ company.name              <chr> "Microsoft", "Berkshire Hathaway", "Microsoft…
$ company.relation          <chr> "founder", "founder", "founder", "founder", "…
$ company.sector            <chr> " Software", " Finance", "technology", " soft…
$ company.type              <chr> "new", "new", "new", "new", "new", "new", "ne…
$ age                       <int> 45, 70, 48, 56, 44, 53, 55, 57, 52, 81, 73, 4…
$ demographics.gender       <chr> "male", "male", "male", "male", "male", "male…
$ location.citizenship      <chr> "United States", "United States", "United Sta…
$ location.country.code     <chr> "USA", "USA", "USA", "USA", "SAU", "USA", "US…
$ location.gdp              <dbl> 1.06e+13, 1.06e+13, 1.06e+13, 1.06e+13, 1.83e…
$ location.region           <chr> "North America", "North America", "North Amer…
$ wealth.type               <chr> "founder non-finance", "founder non-finance",…
$ wealth                    <dbl> 58.7, 32.3, 30.4, 26.0, 20.0, 18.8, 18.7, 18.…
$ wealth.how.category       <chr> "New Sectors", "Traded Sectors", "New Sectors…
$ wealth.how.from.emerging  <chr> "True", "True", "True", "True", "True", "True…
$ industry                  <chr> "Technology", "Consumer goods, retail, restau…
$ wealth.how.inherited      <chr> "not inherited", "not inherited", "not inheri…
$ wealth.how.was.founder    <chr> "True", "True", "True", "True", "True", "True…
$ wealth.how.was.political  <chr> "True", "True", "True", "True", "True", "True…
$ company.age               <dbl> 26, 39, 26, 24, 21, 39, 39, 39, 39, 39, 85, 2…
$ inherited                 <fct> Not inherited, Not inherited, Not inherited, …
```

glimpse(billion2001)

# The method

The primary way to investigate the relationship between a quantitative response variable and many explanatory variables is a multiple linear regression

To select the best model, we decided to perform a backwards elimination on a pool of 8 variables using AIC as the selection criterion

# The method

| variable_list | aic_values |
|---|---|
| all included | 2483.899 |
| -age | 2482.594 |
| -company.age | 2482.317 |
| -company.relation | 2477.996 |
| -company.type | 2480.748 |
| -industry | 2487.738 |
| -inherited | 2482.089 |
| -location.gdp | 2484.224 |
| -wealth.type | 2480.405 |

# Results

-After several rounds of backwards elimination, only one variable remained: industry (previously wealth.how.industry)

-Industry describes the type of industry the billionaire’s associated company belongs to

```r
best_fit <- linear_reg() |>
  set_engine("lm") |>
  fit(wealth ~ industry, data = billion2001)

best_fit |>
  tidy() |>
  select(term, estimate)
```

```
# A tibble: 8 × 2
  term                                estimate
  <chr>                                  <dbl>
1 (Intercept)                            3.97
2 industryEnergy, mining, and metals    -1.89
3 industryMedia                          0.140
4 industryMoney management              -1.21
5 industryNon-consumer industrial       -1.62
6 industryOther                         -2.11
7 industryReal estate, construction     -1.38
8 industryTechnology                     0.831
```

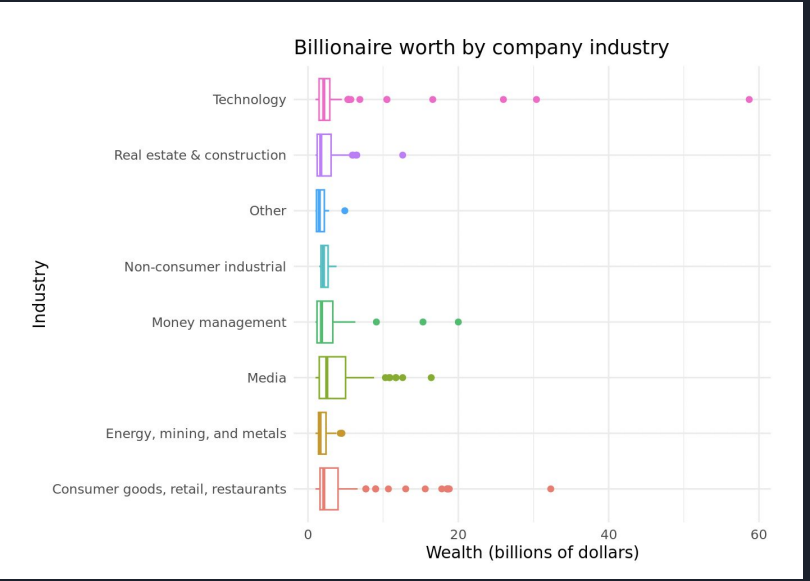
```r
glance(best_fit)$r.squared
```

```
[1] 0.03561891
```

# Conclusion

Our answer to our research question is that the best model for predicting wealth is one with industry as the explanatory variable

Caveat: the $R^2$ value is only 0.036, which is very low. This means that albeit being the best model after backwards elimination with the available variables of the dataset, it is still a very weak fit



# Conclusion

We think the variables in the dataset are still important to the extent of wealth accumulation; however, the process of becoming a billionaire is probably more complicated than just what industry you join, where you live, etc.

Future research on this topic should involve discovering and qualifying/quantifying more possible variables related to wealth accumulation