Tribal influence on governance of companies

Adeola Olawoyin (a.olawoyin.1@research.gla.ac.uk)

2023-06-21

Contents


```
# LOAD THE ORIGINAL DATASET
knitr::opts_chunk$set(echo = T)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(stats)
companies <- readxl::read_xlsx("data.xlsx", n_max = 129)
```

Feature Engineering

```
# DATA PREPARATION
companies_df <- companies %>%
  select(1:9, 29:37) %>%
  mutate(prop_HF_board = HausaFulani/TotalBoardMembers,
         prop_I_board = Igbo/TotalBoardMembers,
         prop_Y_board = Yoruba/TotalBoardMembers,
         prop_M_board = Minority/TotalBoardMembers,
         prop_NN_board = NonNative/TotalBoardMembers,
         homogeneity = pmax(prop_HF_board, prop_I_board, prop_Y_board, prop_M_board, prop_NN_board))
```

Here we derive a single measure of **homogeneity** as the maximum proportion of board members belonging to any single tribe. The range of this measure is between 0 to 1 and can be expressed in percentages. For example, a homogeneity score of 0.6 for Company A would mean that company A has 60% of its board members belonging to one tribe. The lower the score the more diverse the board of the company is and the higher the score the less diverse the company is.

Descriptive Statistics

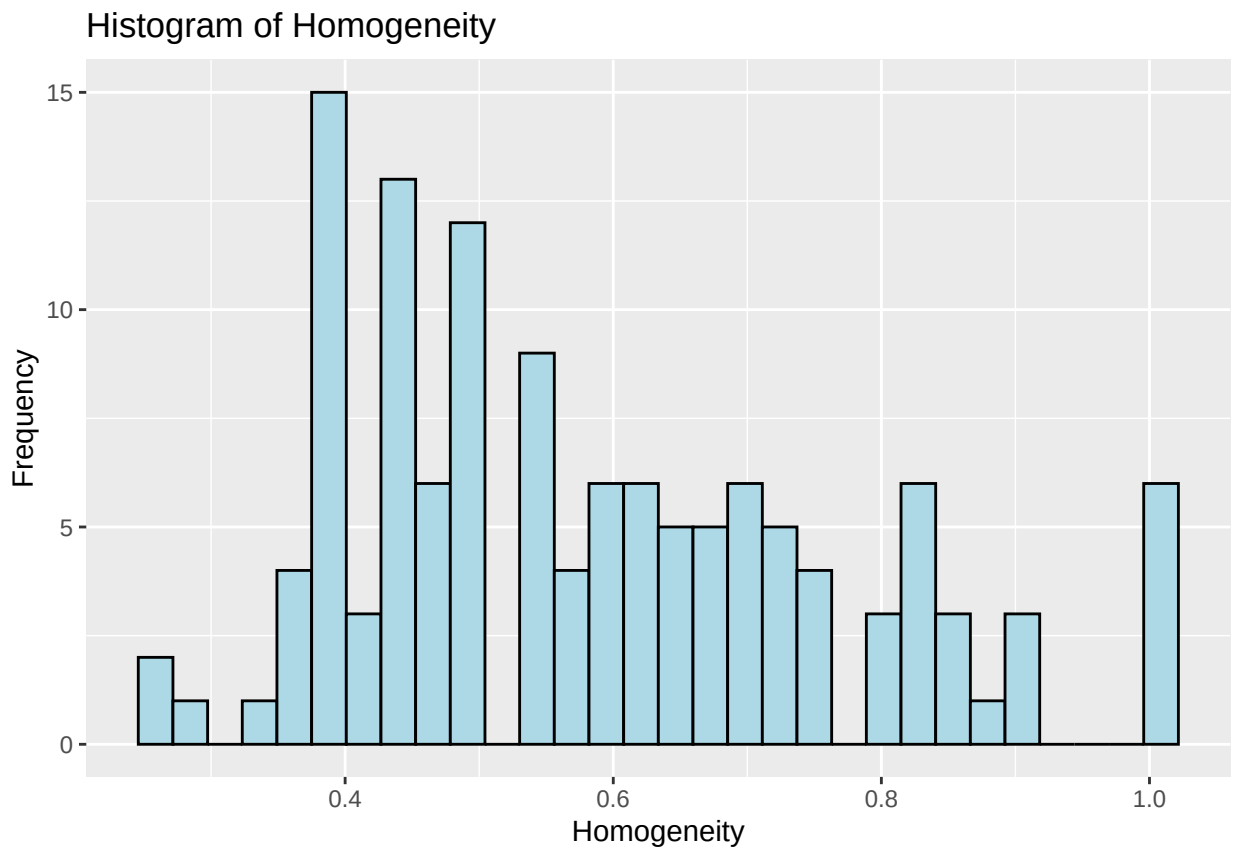
```
# DESCRIPTIVE STATISTICS
# Average homogeneity
print(summary(companies_df$homogeneity))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2500  0.4375  0.5556  0.5862  0.7000  1.0000
```

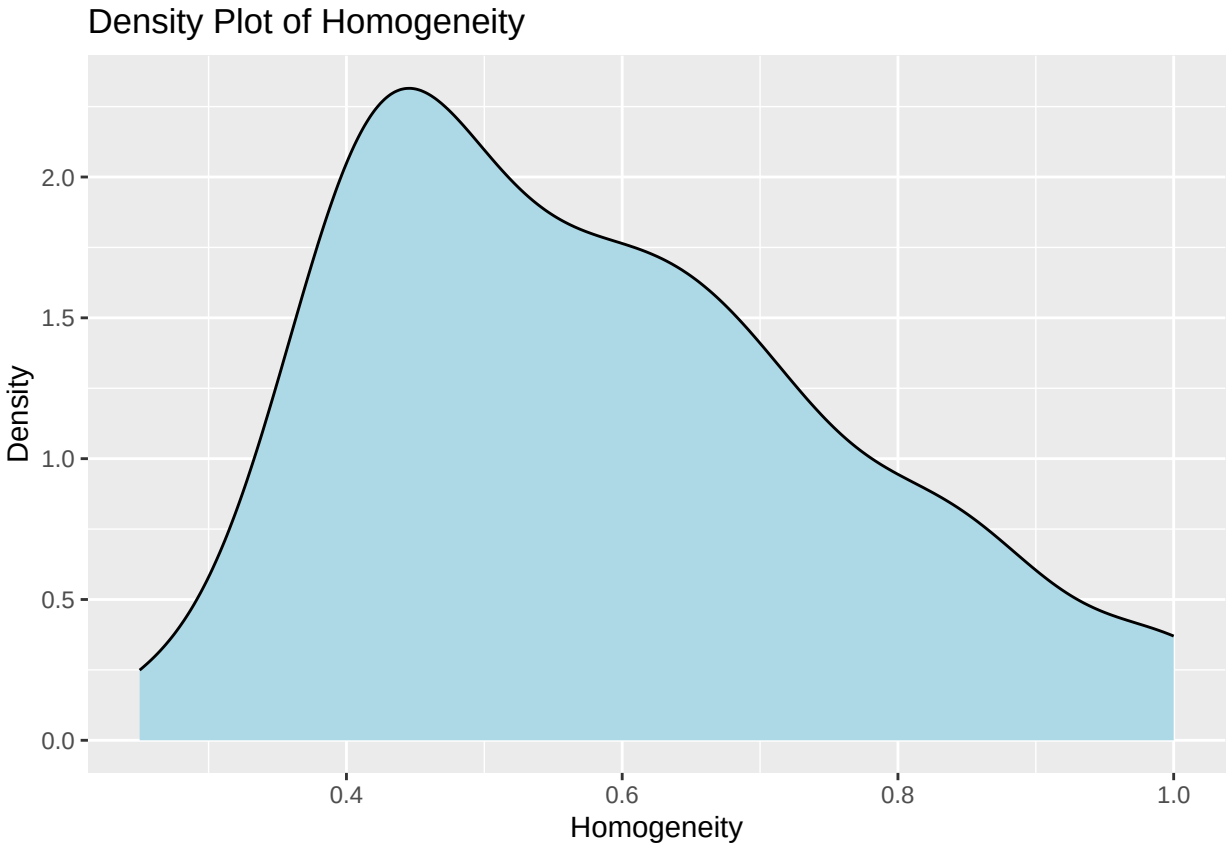
```
# Create a histogram
histogram <- ggplot(companies_df, aes(x = homogeneity)) +
  geom_histogram(fill = "lightblue", color = "black", bins = 30) +
  labs(x = "Homogeneity", y = "Frequency", title = "Histogram of Homogeneity")

# Create a density plot
density_plot <- ggplot(companies_df, aes(x = homogeneity)) +
  geom_density(fill = "lightblue", color = "black") +
  labs(x = "Homogeneity", y = "Density", title = "Density Plot of Homogeneity")

# Display the plots
print(histogram)
```




```
print(density_plot)
```



Here we explore the newly derived variable of interest (homogeneity) by generating summary statistics and visualisations (histogram and density plots).

We can see from the **summary table** that the most diverse company in the data has a maximum of 25% board members from a single tribe and the least diverse company has all its board members (100%) from a single tribe.

The **histogram** reveals that 15 companies have around 40% of board members from a single tribe which is lower than the average homogeneity of 55%. There are 6 companies with 100% board members coming from a single tribe.

The **density plot** and histogram both show that we have a right-skewed distribution which suggests that most companies in the dataset are relatively diverse.

```
# Using the median as a threshold of diversity
sum(companies_df$homogeneity<0.55)
```

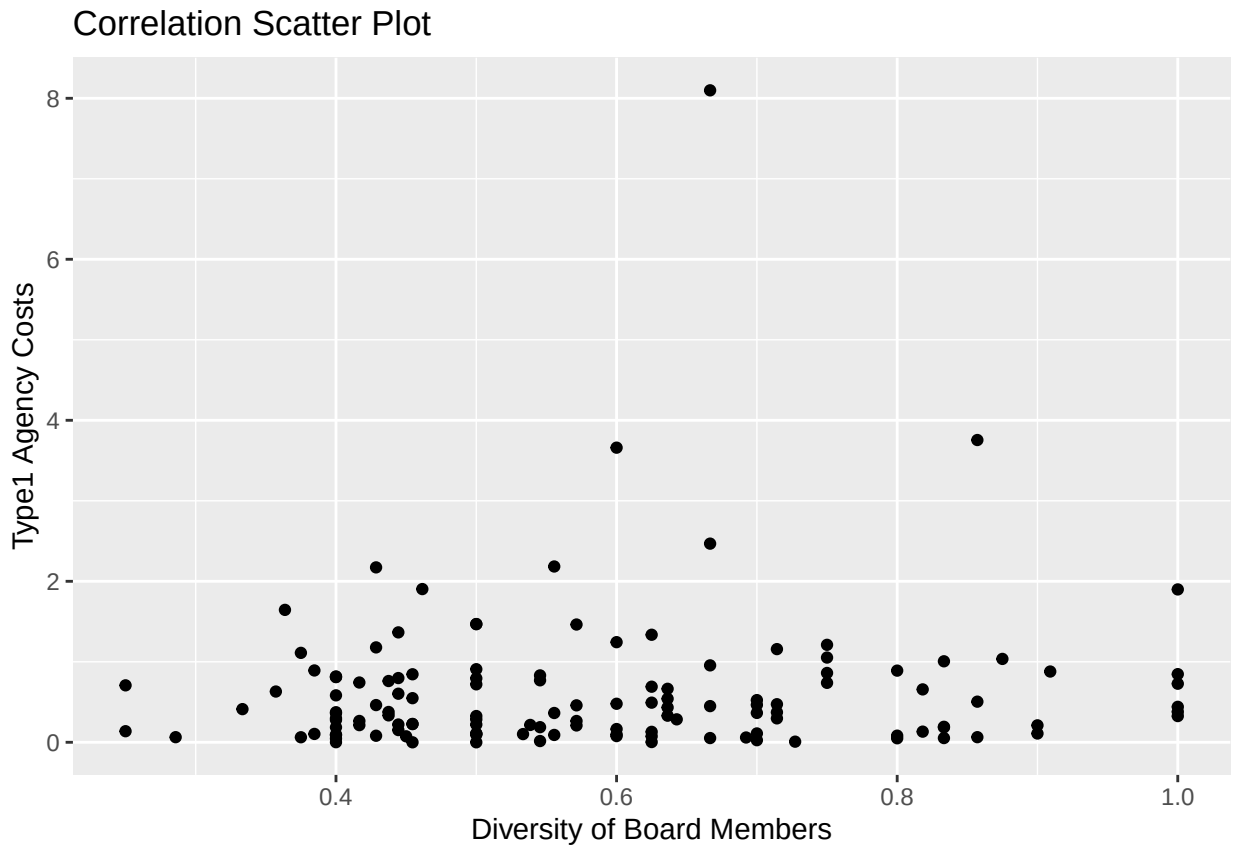
```
## [1] 63
```

```
sum(companies_df$homogeneity>0.55)
```

```
## [1] 66
```

Using **55%** as threshold of diversity, with any company having a score below that as diverse and above as non-diverse — we can see from the values above that the companies in the dataset are roughly equality split between diverse (63) and non-diverse (66).

```
ggplot(companies_df, aes(x = homogeneity, y = AssetTurnover)) +
  geom_point() +
  labs(x = "Diversity of Board Members", y = "Type1 Agency Costs", title = "Correlation Scatter Plot")
```



The scatter plot indicates weak positive correlation between **AssetTurnover** and **Homogeneity**. We shall establish this formally with a linear model and take into account the confounding variables in our dataset. There is an indication of linearity which suggests that a linear model is suitable for examining the relationship between the two variables.

Hypothesis Testing (Regression)

We state our hypothesis as folllows:

H0: The diversity of board members have no association with Type1 agency costs.

H1: The highest proportion of board members affiliated to one tribe will be positively associated with Type1 agency costs.

Model without confounding variables

```
# REGRESSION MODEL (1)
model1 <- lm(homogeneity ~ AssetTurnover, data = companies_df)
summary(model1)
```

```
##
## Call:
## lm(formula = homogeneity ~ AssetTurnover, data = companies_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.33739 -0.14681 -0.03283  0.11901  0.41930
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.57499    0.01928  29.819   <2e-16 ***
## AssetTurnover  0.01752    0.01724   1.016    0.311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1795 on 127 degrees of freedom
## Multiple R-squared:  0.008067,   Adjusted R-squared:  0.0002566
## F-statistic: 1.033 on 1 and 127 DF,  p-value: 0.3114
```

Interpretation

The estimated intercept is 0.57499. It represents the expected value of the dependent variable (**homogeneity**) when the independent variable (**AssetTurnover**) is zero.

The estimated coefficient for **AssetTurnover** is 0.01752. It represents the expected change in the dependent variable for a one-unit increase in **AssetTurnover**, holding other variables constant. However, since the p-value for this coefficient is not statistically significant (p-value: 0.311), we fail to reject the null hypothesis that there is no association between **AssetTurnover** and **homogeneity**.

Multiple R-squared (0.008067) represents the proportion of variance in the dependent variable (**homogeneity**) explained by the independent variable (**AssetTurnover**). In this case, the value is quite low, suggesting that only a small fraction of the variability in **homogeneity** can be explained by **AssetTurnover**.

The F-statistic (1.033) tests the overall significance of the model. In this case, it has a value of approximately 1.033, with 1 and 127 degrees of freedom. The associated p-value (0.3114) is greater than the commonly used significance level of 0.05. Thus, we fail to reject the null hypothesis of no significant association between the independent variable (**AssetTurnover**) and the dependent variable (**homogeneity**).

In summary, based on the provided regression output, there is insufficient evidence to conclude that there is a significant association between **AssetTurnover** and **homogeneity**. The coefficient for **AssetTurnover** is not statistically significant, and the low R-squared values indicate that **AssetTurnover** on its own explain only a small portion of the variability in **homogeneity**.

Model with confounding variables

```
# REGRESSION MODEL (2)
model2 <- lm(homogeneity ~ Years + AssetTurnover + Employees +  as.factor(Sector) + as.factor(Categoriza
summary(model2)
```

```
##
## Call:
## lm(formula = homogeneity ~ Years + AssetTurnover + Employees +
##     as.factor(Sector) + as.factor(Categorization) + as.factor(GroupNotgroup),
##     data = companies_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32900 -0.13441 -0.02768  0.09827  0.45150
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                 5.925e-01  1.929e-01   3.072  0.00267 **
## Years                      -7.203e-04  7.962e-04  -0.905  0.36754
## AssetTurnover               2.537e-02  1.898e-02   1.337  0.18406
## Employees                   6.270e-06  2.233e-06   2.808  0.00587 **
## as.factor(Sector)2          5.737e-02  5.881e-02   0.976  0.33139
## as.factor(Sector)3          7.788e-02  5.547e-02   1.404  0.16303
## as.factor(Categorization)1 -2.107e-01  1.946e-01  -1.083  0.28124
## as.factor(Categorization)2 -8.057e-02  1.842e-01  -0.437  0.66272
## as.factor(Categorization)3  9.011e-04  2.001e-01   0.005  0.99642
## as.factor(Categorization)4 -8.881e-02  2.232e-01  -0.398  0.69141
## as.factor(GroupNotgroup)2   3.462e-02  3.486e-02   0.993  0.32286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1756 on 113 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.1166, Adjusted R-squared:  0.03844
## F-statistic: 1.492 on 10 and 113 DF,  p-value: 0.1515
```

Interpretation

The estimated intercept is 0.5970. It represents the expected value of the dependent variable (**homogeneity**) when all independent variables are zero.

The estimated coefficient for **Years** is -0.0007203. It represents the expected change in the dependent variable for a one-unit increase in **Years**, holding other variables constant. However, since the p-value for this coefficient is not statistically significant (p-value: 0.36), we fail to reject the null hypothesis that there is no association between **Years** and **homogeneity**.

The estimated coefficient for **AssetTurnover** is 0.02537. It represents the expected change in the dependent variable for a one-unit increase in **AssetTurnover**, holding other variables constant. However, since the p-value for this coefficient is not statistically significant (p-value: 0.18), we fail to reject the null hypothesis that there is no association between **AssetTurnover** and **homogeneity**.

The estimated coefficient for **Employees** is 6.270e-06. It represents the expected change in the dependent variable for a one-unit increase in **Employees**, holding other variables constant. The p-value (p-value: 0.005) indicates that this coefficient is statistically significant, suggesting a potential association between **Employees** and **homogeneity**.

For the categorical confounding variables (Sector, Categorization and Business Model) — none of the categorical variables have statistically significant coefficients, as indicated by their respective p-values. The output includes coefficients for different levels of categorical variables. Each level has its own coefficient compared to the reference level, which is typically represented as **(variable)1**. The coefficients for these categorical variables represent the expected change in the dependent variable compared to the reference level, while holding other variables constant. For example, the coefficient for **(Sector)2** is 0.05737, indicating the expected change in **homogeneity** when comparing level 2 of **Sector** to the reference level (level 1).

Logistic regression model with a binary outcome variable and confounding variables

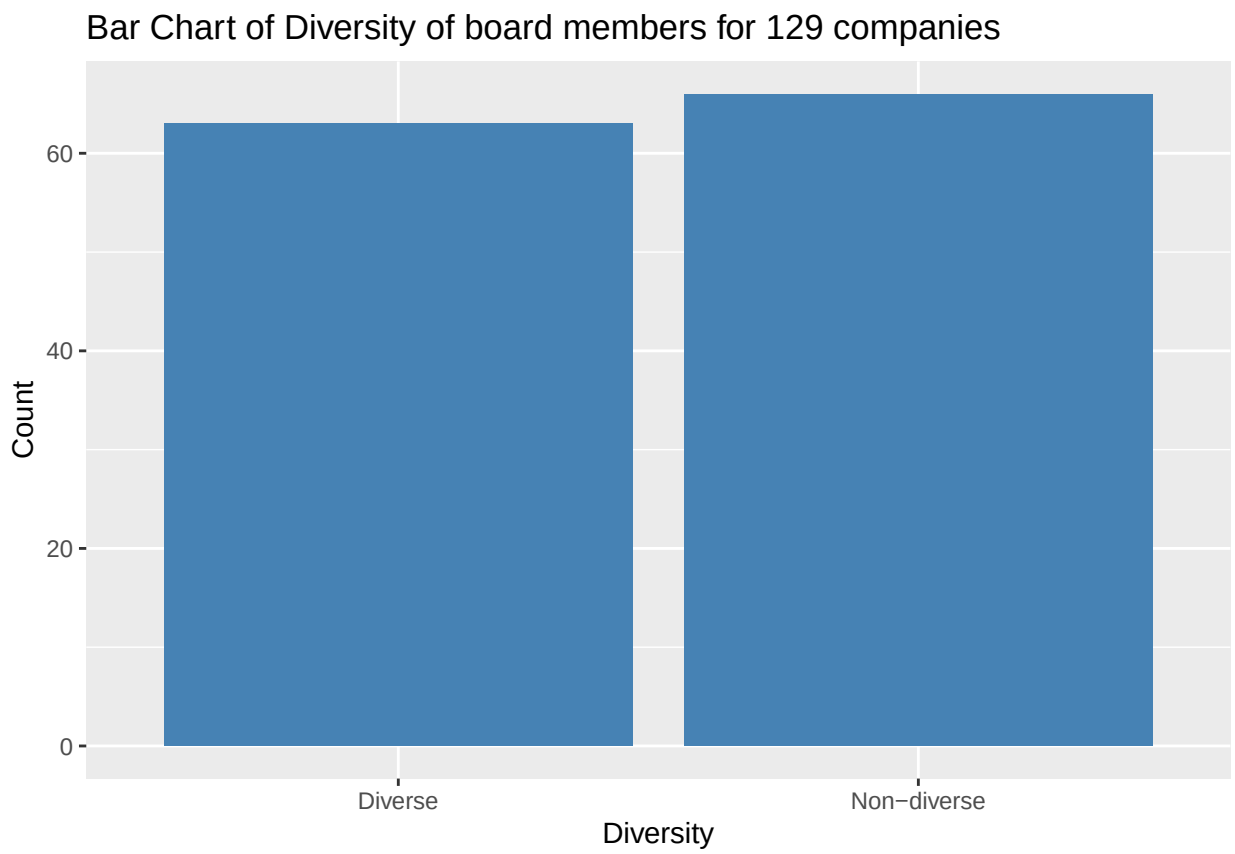
```r
companies_df$diversity <- ifelse(companies_df$homogeneity < 0.55, "Diverse", "Non-diverse")


count <- table(companies_df$diversity)

# Convert the count to a data frame
```

```r
bar_df <- data.frame(diversity = names(count), count = as.numeric(count))

# Generate the bar chart using ggplot
ggplot(bar_df, aes(x = diversity, y = count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  xlab("Diversity") +
  ylab("Count") +
  ggtitle("Bar Chart of Diversity of board members for 129 companies")
```



Using a threshold of 0.55 and categorizing an homogeneity score below that as diverse and equal or above 0.55 as non-diverse, we get a roughly equal split between both groups.

```r
# REGRESSION MODEL (3)
companies_df <- companies_df %>%
  mutate(diversity = fct_recode(as.factor(diversity), `1` = "Diverse", `0` = "Non-diverse"))

model3 <- glm(as.numeric(diversity) ~ Years + AssetTurnover + Employees +  as.factor(Sector) + as.factor
summary(model3)
```

```
##
## Call:
## glm(formula = as.numeric(diversity) ~ Years + AssetTurnover +
##     Employees + as.factor(Sector) + as.factor(Categorization) +
##     as.factor(GroupNotgroup), data = companies_df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
```

```
## -0.73830  -0.43715  -0.02188   0.40552   0.82442
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.688e+00  5.201e-01   3.246  0.00154 **
## Years                        -5.843e-03  2.147e-03  -2.721  0.00753 **
## AssetTurnover                 8.337e-02  5.118e-02   1.629  0.10611
## Employees                     1.024e-05  6.021e-06   1.701  0.09173 .
## as.factor(Sector)2            6.001e-02  1.586e-01   0.378  0.70583
## as.factor(Sector)3            2.098e-01  1.496e-01   1.403  0.16344
## as.factor(Categorization)1 -6.078e-01  5.248e-01  -1.158  0.24927
## as.factor(Categorization)2 -2.008e-01  4.968e-01  -0.404  0.68682
## as.factor(Categorization)3 -3.138e-02  5.397e-01  -0.058  0.95374
## as.factor(Categorization)4 -3.376e-01  6.018e-01  -0.561  0.57587
## as.factor(GroupNotgroup)2   1.546e-01  9.402e-02   1.645  0.10284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2243022)
##
##     Null deviance: 30.992  on 123  degrees of freedom
## Residual deviance: 25.346  on 113  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 179.03
##
## Number of Fisher Scoring iterations: 2
```

Interpretation

Again, no evidence of an association between **AssetTurnover** and **Diversity, t**aking all counfounding variables into account. However, interestingly in this model, there is a statistically significant association between that **Years** and **AssetTurnover**, holding other variables constant.

Notes

1. I have used AssetTurnover as the independent variable, it may be worth standardizing cash flow and using it instead to see if there's any difference.
2. It may also be worth dropping one or more of the confounding variables to observe any change.