Mental Health Data Analysis Study

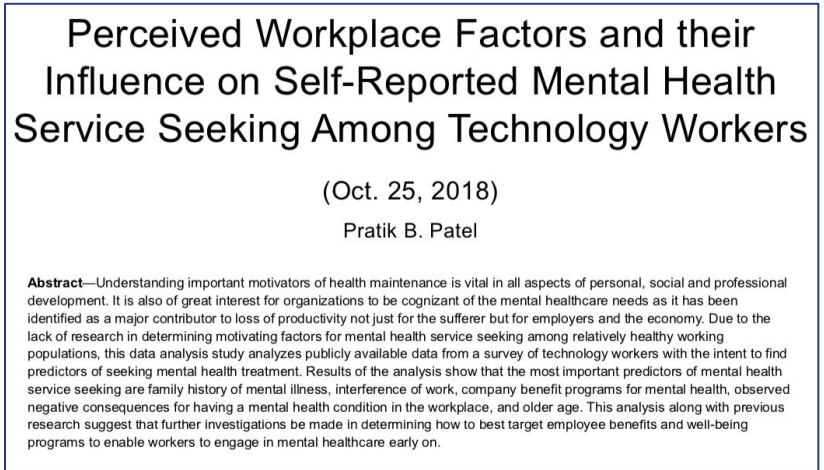
By: Lealena Befikadu, Aisha Koroma, Edelawit Sebesibe, Ayesha Shafquat

INST408F - Final Presentation

May 11, 2020

Data Description

- OSMI Mental Health in Tech Survey from **2018**
- Survey to **measure attitudes towards mental health in the tech workplace**
- OSMI publishes a link to a survey on their website and promotes it
- **Limitations:** survey is open to anyone = inaccurate data & few volunteers
- **Example:**



Perceived Workplace Factors and their Influence on Self-Reported Mental Health Service Seeking Among Technology Workers

(Oct. 25, 2018)

Pratik B. Patel

Abstract—Understanding important motivators of health maintenance is vital in all aspects of personal, social and professional development. It is also of great interest for organizations to be cognizant of the mental healthcare needs as it has been identified as a major contributor to loss of productivity not just for the sufferer but for employers and the economy. Due to the lack of research in determining motivating factors for mental health service seeking among relatively healthy working populations, this data analysis study analyzes publicly available data from a survey of technology workers with the intent to find predictors of seeking mental health treatment. Results of the analysis show that the most important predictors of mental health service seeking are family history of mental illness, interference of work, company benefit programs for mental health, observed negative consequences for having a mental health condition in the workplace, and older age. This analysis along with previous research suggest that further investigations be made in determining how to best target employee benefits and well-being programs to enable workers to engage in mental healthcare early on.

Data Cleaning and Preparation

- Variables that were recorded:
  - The **comfort** variable had 3 responses: No, Maybe, Yes. Because these are distinct categories we only considered No and Yes as valid values.
  - The **gender** variable had a text response option, so responses varied from binary to biological sexes. For this analysis only biological responses were considered i.e. female/male.
  - The **country of employment** variable had 40 countries as response options but only 5 of them were used in this analysis: “Canada”, “USA”, “United Kingdom”, “Germany” and “India”.
  - The **Healthcare** variable had 4 responses: No, Yes, IDK and NA. IDK and NA are not valid values so they were excluded.
- We cleaned the data by:
  - The comfort variable was subset to only No/Yes responses. All other values were coded to NA
  - We subset and re-coded gender to only include responses that meant ‘female’ and ‘male’. All other values were coded to NA
  - The country of employment variable was subset to only count countries with at least 5 responses. All other values were coded to NA
  - The healthcare variable was subset to only No/Yes responses. All other values were coded to NA

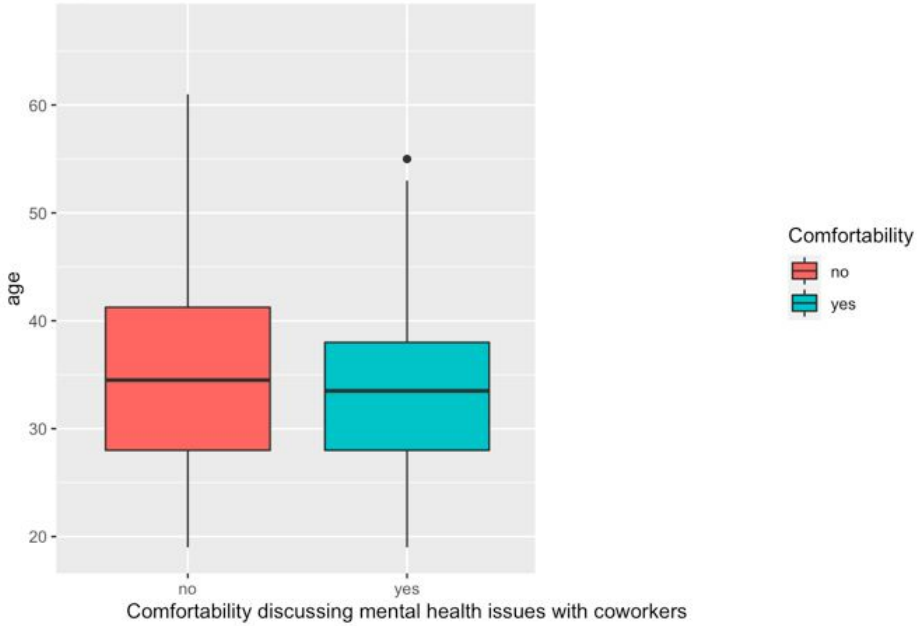
Hypothesis Test

- **Research question:**
  - Does someone's comfort level when talking about personal mental health issues with coworkers differ amongst age?
- **Variables:**
  - What is your age?
  - Would you feel comfortable discussing a mental health issue with your coworkers?
    - i.e. Yes and No
- **Null Hypothesis:** The mean age between people that are comfortable discussing personal mental health issues with coworkers and those who aren't, is the same.
- **Alternative Hypothesis:** The mean age differs between people that are comfortable discussing personal mental health issues with coworkers and those who aren't.
- **Significance Level:** Alpha = 0.05
- **P - Value:** 0.2806
- **Conclusion:** At a significance level of 0.05, we fail to reject the null hypothesis and conclude that there is NOT statistically significant evidence that the mean age differs between people that are comfortable discussing personal mental health issues with coworkers and those who aren't.

Hypothesis R output and Box Plot

```
        Wilcoxon rank sum test with continuity correction

data:  IV.age[DV.discuss == "no"] and IV.age[DV.discuss == "yes"]
W = 5366.5, p-value = 0.2806
alternative hypothesis: true location shift is not equal to 0
```

Figure 1: Comparison of Comfortability discussing mental health with coworkers



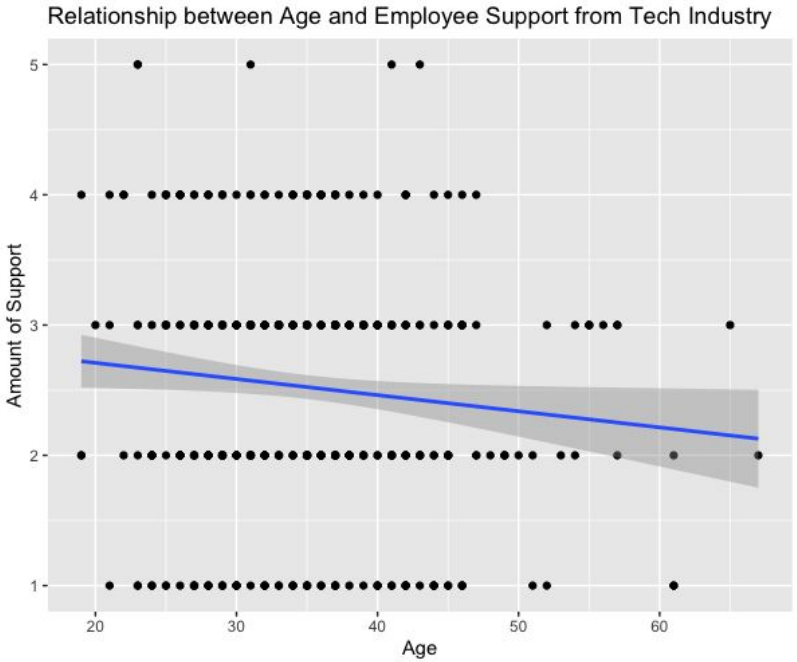
Interpretation: The Figure 1 boxplot shows the difference in spread/central tendency of the "Would you feel comfortable discussing a mental health issue with your coworkers?" variable, by Age. The uncomfortable group has a higher median age than those who are comfortable.The spread of the data for uncomfortable people is more than that of comfortable individuals. Only the comfortable group presents an outlier.

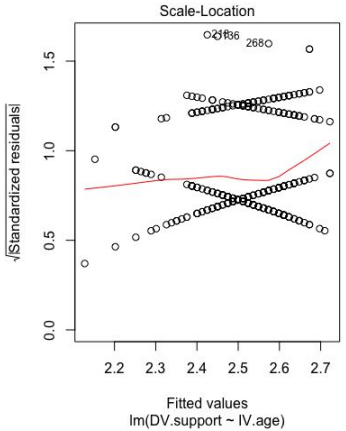
Simple Linear Regression

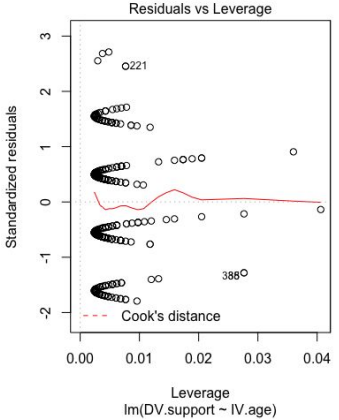
- **Research Question:**
  - Does the amount of support in the tech industry on mental health issues differ amongst the age of employees?
- **Variables:**
  - Independent Variable: Age (19-67)
  - Dependent Variable: Overall, how well do you think the tech industry supports employees with mental health issues? (1-5 Likert Scale)
- **Slope Coefficient:**
  - A 1-year increase in age causes the amount of support decreases by an average of 0.0124
- **Intercept coefficient:** When age equals zero, the average value of the amount of support is 2.96
- **P-value:** The p-value (0.03) is less than alpha, 0.05, so we can conclude that the regression model is statistically significant and that the amount of support in the tech industry on mental health issues is associated with the age of employees
- **R^2:** 0.85% of the variation in the amount of support in the tech industry on mental health issues is explained by the variation in age

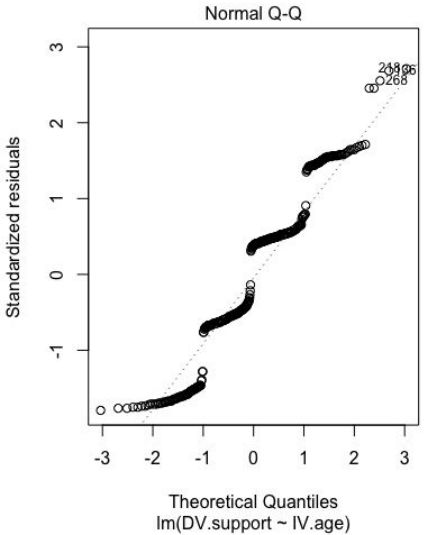
```
Call:
lm(formula = DV.support ~ IV.age, data = osmi)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6975 -0.5985  0.3520  0.5130  2.5749

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.957603   0.207749  14.236   <2e-16 ***
IV.age      -0.012385   0.005798  -2.136   0.0333 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9517 on 415 degrees of freedom
Multiple R-squared:  0.01087,   Adjusted R-squared:  0.00849
F-statistic: 4.562 on 1 and 415 DF,  p-value: 0.03327
```

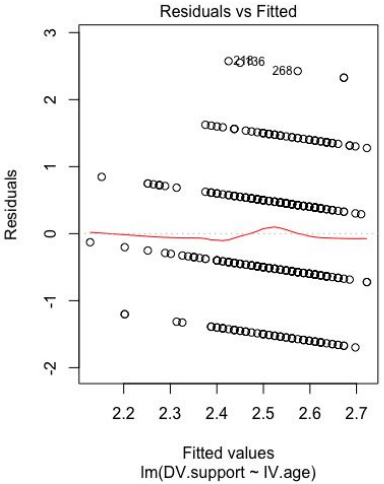
Simple Linear Regression Plots











All assumptions were violated,

meaning our data could be biased

and inefficient

Multiple Linear Regression

- **Research question:**
  - Does the age of employees, their gender, and country of employment share a relationship with the amount of importance employers place on mental health?
- **Variables:**
  - Independent variable:
    - What country do you work in? (i.e. United States of America, United Kingdom, Canada, Germany, India)
    - What is your gender? (i.e. Male, Female)
    - What is your age?
  - Dependent Variable:
    - Overall, how much importance does your employer place on mental health? (i.e. 0-10 Likert Scale)
- **P - Value:**
  - The p-value is 0.1181. The p-value is more than alpha, 0.05, so we can conclude that the regression model is NOT statistically significant and that age, gender and country of employment are not associated with the amount of importance employers place on mental health.
- **R^2:**
  - The adjusted R^2 is 0.01554. 1.55% of the variation in the amount of importance employers place on mental health is explained by the variation in age, gender and country of employment.

Multiple Linear Regression

Interpret the Slope and Intercept Coefficient

- **Slope:**
  - **Age:** A 1-year increase in age causes the amount of importance employers place on mental health to decrease by an average of 0.00771, holding all other variables in the model constant.
  - **Gender:** The amount of importance employers place on mental health is lower by 0.46532 for males than for females, holding all other variables in the model constant.
  - **Country**: The amount of importance employers place on mental health is lower by 0.58481 for Germany than for Canada, holding all other variables in the model constant.
  - The amount of importance employers place on mental health **is higher** by 2.05017 for India than for Canada, holding all other variables in the model constant.
  - The amount of importance employers place on mental health **is higher** by 2.03821 for the United Kingdom than for Canada, holding all other variables in the model constant.
  - The amount of importance employers place on mental health **is higher** by 0.57287 for the United States than for Canada, holding all other variables in the model constant.
- **Intercept:**
  - When all of the predictor variables equal zero, the average value of the amount of importance employers place on mental health is 4.81097

Multiple Linear Regression R results and Scatter plot

```
Call:
lm(formula = DV.importance ~ IV.age + factor(IV.gender) + factor(IV.country),
    data = osmi)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1680 -1.6892  0.3282  1.8706  5.3667

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.81097    1.32319   3.636 0.000333 ***
IV.age                -0.00771    0.01948  -0.396 0.692554
factor(IV.gender)male -0.46532    0.33038  -1.408 0.160181
factor(IV.country)DE  -0.58481    1.46312  -0.400 0.689700
factor(IV.country)IN   2.05017    1.56535   1.310 0.191425
factor(IV.country)UK   2.03821    1.23571   1.649 0.100246
factor(IV.country)USA  0.57287    1.04862   0.546 0.585313
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.507 on 265 degrees of freedom
  (145 observations deleted due to missingness)
Multiple R-squared:  0.03734,   Adjusted R-squared:  0.01554
F-statistic: 1.713 on 6 and 265 DF,  p-value: 0.1181
```

Interpretation:

Figure 3: Scatterplot shows no linear relationship between Age and The amount of importance employers place on mental health.

Multiple Linear Regression Boxplots

Figure 4: Comparison of The amount of importance employers place on mental

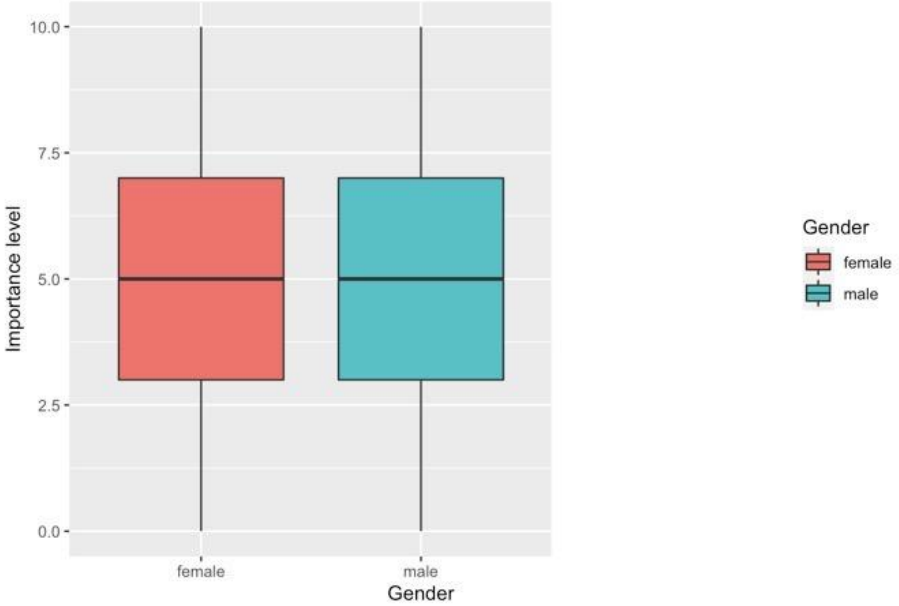
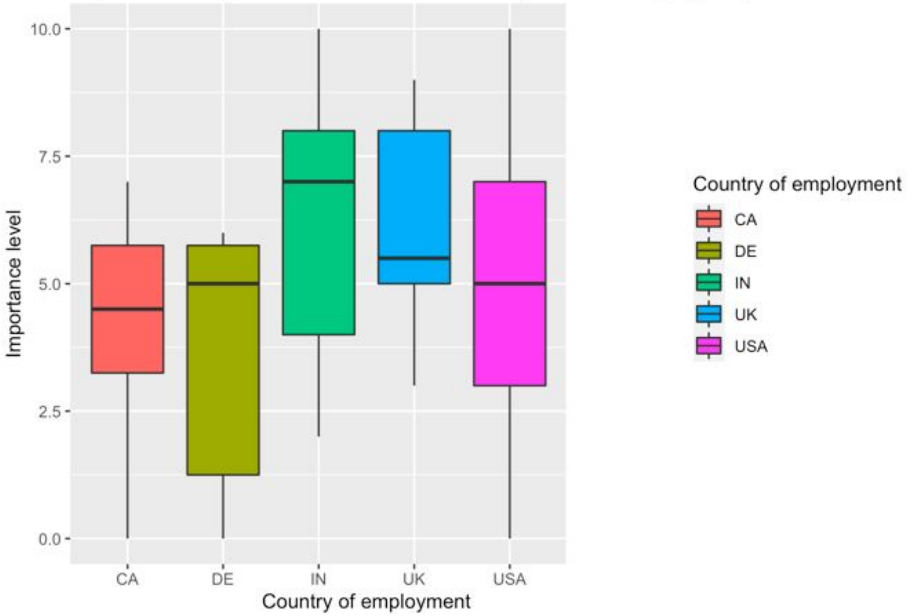


Figure 5: Comparison of The amount of importance employers place on mental



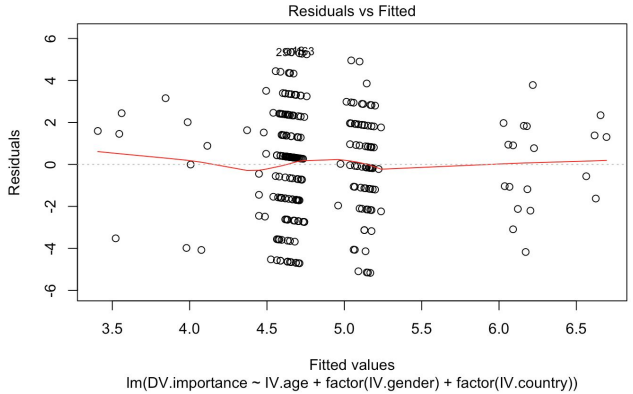
Interpretation:

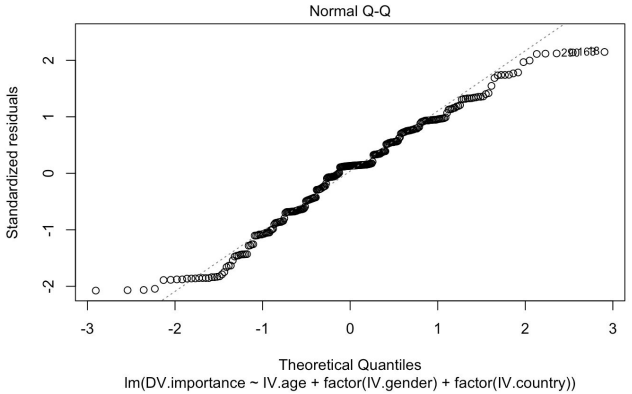
- Figure 4: Boxplot shows the difference in spread/central tendency of the "Overall, how much importance does your employer place on mental health?" variable, by Gender. The female and male group seem to have equal median ages.The spread of the data for both groups are the same and neither group presents an outlier.
- Figure 5: Boxplot shows the difference in spread/central tendency of the "Overall, how much importance does your employer place on mental health?" variable, by Country of employment. India has the highest median amongst the 5 countries. Then follows the United Kingdom as second. The United States and Denmark have the same median.Canada comes in last with the lowest median and smallest spread of data. No countries present an outlier.

Multiple Linear diagnostic plots

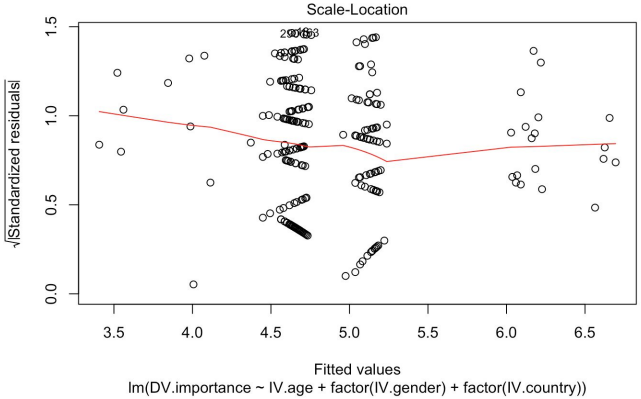
Assumptions:

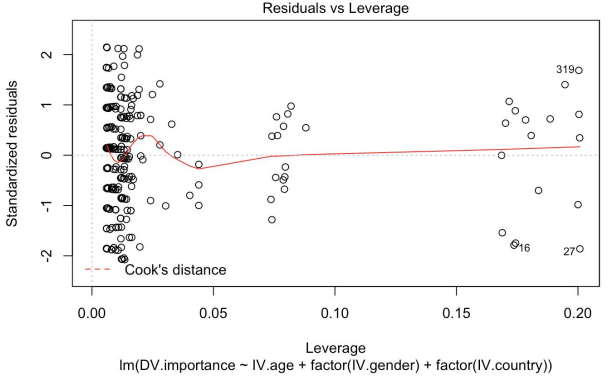
- **Residuals vs Fitted:** Our data appears to violate this assumption because the relationship between the independent variables and the amount of importance employers place on mental health is not linear. And patterns are present.
- **Normal Q-Q:** Our data appears to MOSTLY meet this assumption, but the ends aren't quite perfectly fitted on the dashed line. We can assume the residuals are normally distributed.
- **Scale-Location:** Our data appears to violate this assumption because our points are not randomly spread out and the red line has a slightly horizontal positive slope.
- **Residuals vs Leverage:** Our data appears to violate this assumption because we have influential data points that are within the red Cook's distance curved line.





The violation of these assumptions means, the forecasts, confidence intervals, and scientific insights yielded the regression model may be (at best) inefficient or (at worst) seriously biased or misleading





Logistic Regression

Based on their sex, age and country they live in, can that be used to predict whether or not employers provide mental health benefits as part of healthcare coverage?

Independent variables:

- What country do you live in?
  - United States of America, United Kingdom, Canada, Germany, India
- i.e.What is your gender
  - Male, Female
- What is your age?
  - 19-67

Dependent variable:

- Does your employer provide mental health benefits as part of healthcare coverage?
  - No, Yes

```
Call:
glm(formula = factor(Does.your.employer.provide.mental.health.benefits.as.part.of.healthcare.coverage.) ~
    What.is.your.age. + factor(What.is.your.gender.) + factor(What.country.do.you..strong.live..strong..in.),
    family = "binomial", data = mental_health_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5478   0.2849   0.4263   0.4712   0.8235

Coefficients:
                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       20.15867 2852.83351   0.007    0.994
What.is.your.age.                                 -0.02218    0.03207  -0.692    0.489
factor(What.is.your.gender.)male                  -0.88652    0.68686  -1.291    0.197
factor(What.country.do.you..strong.live..strong..in.)DE  -37.38202 4292.63232  -0.009    0.993
factor(What.country.do.you..strong.live..strong..in.)IN  -37.27409 4332.09134  -0.009    0.993
factor(What.country.do.you..strong.live..strong..in.)UK  -17.34448 2852.83334  -0.006    0.995
factor(What.country.do.you..strong.live..strong..in.)USA -16.30934 2852.83312  -0.006    0.995

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 145.74  on 195  degrees of freedom
Residual deviance: 105.53  on 189  degrees of freedom
  (221 observations deleted due to missingness)
AIC: 119.53

Number of Fisher Scoring iterations: 17
```

Logistic Regression

Interpret the slope coefficient

- **What is your age:**
  - A 1 year increase in the average age causes the log odds of having a mental health benefits as part of healthcare coverage to decrease by -0.02218, holding the other independent variables constant.
- **What country do you live in:**
  - The log odds of having a mental health benefits as part of healthcare coverage between someone who lives in Germany and someone who lives in Canada is -37.38202,holding the other independent variables constant.
  - The log odds of having a mental health benefits as part of healthcare coverage between someone who lives in India and someone who lives in Canada is -37.27409,holding the other independent variables constant.
  - The log odds of having a mental health benefits as part of healthcare coverage between someone who lives in the United Kingdom and someone who lives in Canada is -17.34448,holding the other independent variables constant.
  - The log odds of having a mental health benefits as part of healthcare coverage between someone who lives in the United States of America,and someone who lives in Canada is -16.30934,holding the other independent variables constant.
- **What is your gender:**
  - The log odds of having a mental health benefits as part of healthcare coverage between someone who male and someone who female is -0.88652,holding the other independent variables constant.

Logistic Regression

P-value:

- Since the p-value for age is 0.489 which is more than alpha, 0.05, we can conclude that age is not associated with whether or not an individual has a mental health benefits as part of healthcare coverage

- Since the p-value for sex is 0.197 which is more than alpha, 0.05, we can conclude that sex is not associated with whether or not an individual has a mental health benefits as part of healthcare coverage

- Since the p-value for the country they live in is 0.993(Germany and India) and 0.995(United Kingdom and United States of America) which are both more than alpha, 0.05, we can conclude that all of the countries tested is not associated with whether or not an individual has a mental health benefits as part of healthcare coverage

Thank you for Watching!