# Effects of Socialization on Mental Health

## STA130 Course Project

Edie Chen Jason Li Zain Mahmoud Rana Nagash
TA Oliver Gatalo
Professor Scott Schwartz

STA130: An Introduction to Statistical Reasoning and Data Science
Department of Statistical Sciences
University of Toronto

# Introduction

Social interactions play a pivotal role in shaping individual mental health outcomes. It is becoming increasingly easier, especially for teenagers, to connect with their friends virtually from the comfort of their homes. One may argue that this is harmful for their mental health; but is this always the case?

Through this research, we aim to highlight the difference between physically interacting with community members as opposed to virtually connecting with them. We used the Canadian Social Connections Survey (CSCS) to investigate the relationship between various forms of social interactions (physical and non-physical) and how they affect the individuals' mental health states. This presentation outlines the variables we're using, our hypotheses, analyses, key findings, and the conclusions we've drawn from these findings.

# Our research questions

## Question 1

Is there an association between the frequency of days where an individual spends at least 5 minutes physically socializing and their level of depression?

## Question 2

Is there an association between playing video games and feeling depressed, and can going on walks counteract that?

## Question 3

How does the association between loneliness and video chatting compare to text messaging?

# Question 1: Variables

Independent variables:

`CONNECTION_social_days_family_p7d_grouped`:
days where individuals spent at least 5 minutes socializing with family.

`CONNECTION_social_days_friends_p7d_grouped`:
days where individuals spent at least 5 minutes socializing with friends.

`CONNECTION_social_days_coworkers_and_classmates_p7d_grouped`:
days where individuals spent at least 5 minutes socializing with co-workers or classmates.

`CONNECTION_social_days_neighbours_p7d_grouped`:
days where individuals spent at least 5 minutes socializing with neighbours.

Dependent variable:
`WELLNESS_phq_score`:
metric used to characterize an individual's level of depression on a scale of 0-6.

# Preliminary analysis

The independent variables were categorical with 4 categories each:

```
df_cleaned['CONNECTION_social_days_family_p7d_grouped'].unique()
```

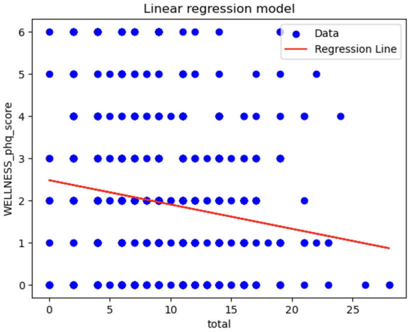
```
array(['None (0 Days)', 'Most days (4 - 6 days)',
       'Some days (1 - 3 days)', 'Every day (7 days)'], dtype=object)
```

Figure: Unique data entries in one of the columns

To better analyze the data, we gave each category a numeric value based on the midpoint of the interval. For example, the 'Most days (4-6)' category was assigned 5 (representing the midpoint of the number of days). Then, we added another column to represent the total number of days where each individual spent at least 5 minutes socializing with any one of the groups above using the numeric values we assigned to each category.

# Analysis

First, we examined the relationship between the total column and the numeric PHQ score column. We did this by fitting a simple linear regression through the data.




OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | WELLNESS_phq_score | **R-squared:** | 0.026 |
| **Model:** | OLS | **Adj. R-squared:** | 0.025 |
| **Method:** | Least Squares | **F-statistic:** | 15.49 |
| **Date:** | Sat, 23 Nov 2024 | **Prob (F-statistic):** | 9.30e-05 |
| **Time:** | 18:39:11 | **Log-Likelihood:** | -1153.2 |
| **No. Observations:** | 575 | **AIC:** | 2310. |
| **Df Residuals:** | 573 | **BIC:** | 2319. |
| **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 2.4779 | 0.158 | 15.675 | 0.000 | 2.167 | 2.788 |
| **total** | -0.0576 | 0.015 | -3.936 | 0.000 | -0.086 | -0.029 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 47.542 | **Durbin-Watson:** | 1.575 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 53.879 |
| **Skew:** | 0.722 | **Prob(JB):** | 2.00e-12 |
| **Kurtosis:** | 2.595 | **Cond. No.** | 22.9 |

# Analysis

We then created a bootstrapped distribution of model slope coefficients by repeatedly resampling from our original sample and refitting OLS models through the samples. Then, we created a 95% confidence interval of our bootstrapped coefficients for inference.

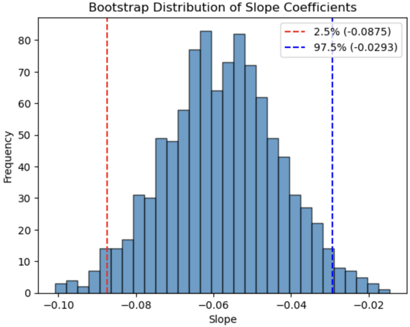



Figure: 95% confidence interval

# Summary and conclusion

The confidence interval we constructed only contained negative slopes between $-0.0875$ and $-0.0293$ and so we can conclude with 95% confidence that the true value of the slope coefficient lies in that interval. This means that as the number of days where an individual spends at least 5 minutes socializing increases, the average depression score decreases. However, the values of the slopes are very small and so the effect of socializing on depression scores is minuscule.

# Limitations and assumptions

For this analysis, we only considered days where individuals spent at least 5 minutes socializing. However, this threshold is vague and is not specific enough. One person could spend 5 minutes socializing and another could spend 5 hours socializing yet they would still be considered under the same category. I also made some assumptions to convert the categorical values into numeric values. However, given that the mapping of categories involved the midpoint of each interval, I believe this conversion was appropriate.

# Question 2: Variables

## Independent variables:

CONNECTION_activities_onlinegames_p3m:
how often an individual has played online games in the past 3 months

**Ordinal categorical outcomes:** Not in the past three months, Less than monthly, Monthly, A few times a month, Weekly, A few times a week, Daily or almost daily

CONNECTION_activities_walk_p3m:
how often an individual has gone on a walk with friends in the past 3 months

**Ordinal categorical outcomes:** Not in the past three months, Less than monthly, Monthly, A few times a month, Weekly, A few times a week, Daily or almost daily

## Dependent variable:

WELLNESS_malach_pines_burnout_measure_depressed:
how often an individual feels depressed **Ordinal categorical outcomes:** Never, Almost never, Rarely, Sometimes, Very Often, Always

# Limitations and assumptions

It is not possible to perfectly map the ordinal categories for how often an individual feels depressed Never, Almost never, Rarely, Sometimes, Very Often, Always numerically. It shouldn't be assumed that the "distance" between each category is "1", but I will map it to the numbers 0 through 5 for multiple linear regression.

It is not possible to measure level of depression in a binary variable, if I wanted to do something like a logistic regression.

I will keep the outcomes as a continuous variable (intended to be interpreted from a 0 through 5 scale), rather than converting it back to categorical.

# Cleaning data

First, I want to assign numbers to the ordinal categories of how often an individual feels depressed.
I will just use the consecutive numbers 0 (never) through 5 (always).

```python
# Mapping the variables to numeric values
mapping_dict = {
    'WELLNESS_malach_pines_burnout_measure_depressed': {
        'Never': 0,
        'Almost never': 1,
        'Rarely': 2,
        'Sometimes': 3,
        'Very Often': 4,
        'Always': 5
    }
}
```

After renaming variables, removing empty values, etc., this is what my DataFrame looks like.

| | OnlineGamesC | WalkWithSomeoneC | DepressionC | DepressionN |
|---|---|---|---|---|
| **0** | Not in the past three months | Daily or almost daily | Rarely | 2.0 |
| **1** | Not in the past three months | A few times a week | Almost never | 1.0 |
| **2** | Not in the past three months | A few times a month | Almost never | 1.0 |
| **3** | Weekly | Less than monthly | Rarely | 2.0 |
| **4** | Weekly | Monthly | Almost never | 1.0 |
| **...** | ... | ... | ... | ... |

# Multiple Linear Regression

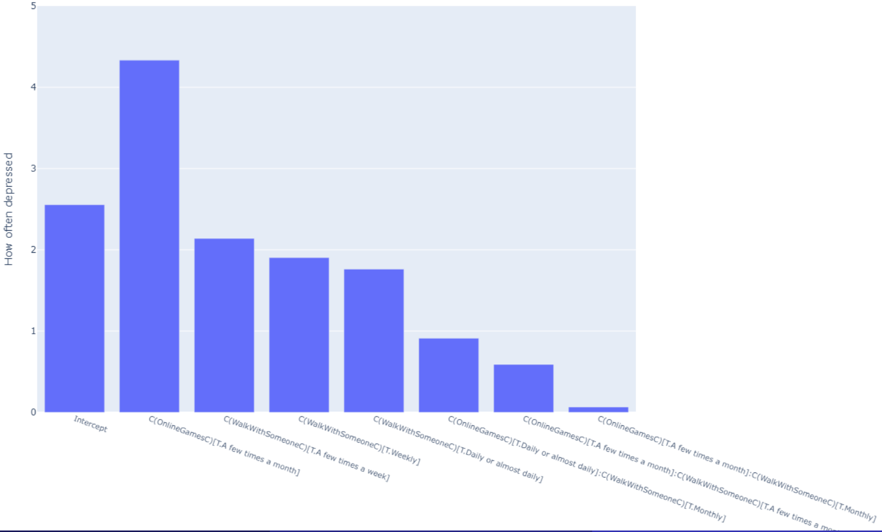
The table below retains only the significant (p-value $\leq 0.05$) outcomes (rest are omitted) from a multiple linear regression. I will create a new column that will show the predicted $\hat{y}$ values, and plot this on a bar plot.

| | | coef | std err | t | P>\|t\| | [0.025 | 0.975] | yhat |
|---|---|---|---|---|---|---|---|---|
| **0** | Intercept | 2.5546 | 0.122 | 21.001 | 0.000 | 2.316 | 2.793 | 2.5546 |
| **3** | C(OnlineGamesC)[T.A few times a month] | 1.7787 | 0.776 | 2.293 | 0.022 | 0.255 | 3.302 | 4.3333 |
| **10** | C(WalkWithSomeoneC)[T.Weekly] | -0.6510 | 0.190 | -3.430 | 0.001 | -1.024 | -0.278 | 1.9036 |
| **11** | C(WalkWithSomeoneC)[T.A few times a week] | -0.4154 | 0.193 | -2.157 | 0.031 | -0.794 | -0.037 | 2.1392 |
| **12** | C(WalkWithSomeoneC)[T.Daily or almost daily] | -0.7927 | 0.238 | -3.328 | 0.001 | -1.260 | -0.325 | 1.7619 |
| **21** | C(OnlineGamesC)[T.A few times a month]:C(WalkWithSomeoneC)[T.Monthly] | -2.4898 | 0.997 | -2.498 | 0.013 | -4.447 | -0.533 | 0.0648 |
| **24** | C(OnlineGamesC)[T.Daily or almost daily]:C(WalkWithSomeoneC)[T.Monthly] | -1.6440 | 0.777 | -2.115 | 0.035 | -3.170 | -0.118 | 0.9106 |
| **27** | C(OnlineGamesC)[T.A few times a month]:C(WalkWithSomeoneC)[T.A few times a month] | -1.9651 | 0.987 | -1.991 | 0.047 | -3.903 | -0.027 | 0.5895 |

UNIVERSITY OF TORONTO

# Bar plot

Sorted in descending order, with intercept first.



# Findings

Recall that the values on the bar chart is how level of depression changes as the variables in the x-axis change.

We find that the intercept (don't play online games and don't go on walks with friends) is 2.5 on this 0-5 scale. It is interesting that the intercept is at the halfway point.

We find that even just playing online games a few times a month can significantly increase one's level.

As an individual's frequency of walks with friends increases, their level of depression decreases.

This change is apparent in both individuals who play games and individuals who don't.

In conclusion, we find that only playing online games correlates positively with one's level of depression, while going on walks with friends significantly lowers that.

UNIVERSITY OF TORONTO

# Question 3 Variables

How does the association between loneliness and video chatting compare to text messaging?

**Independent Variables:**

Video chatted with friends/family in the past 3 months:
Texted or messaged someone in the past 3 months to check in:

7 Options:

"Not in the past three months" ... "Daily or almost daily",

**Dependent Variables:**

How many days felt lonely in the past week:

5 Options:

'None of the time (e.g., 0 days)': 0 ... 'All of the time (e.g. 5-7 days)': 6

```
CONNECTION_activities_text_or_messaged_p3m_combined
Few times a month or weekly      266
Few times a week or daily        226
Less than monthly or monthly     209
Not in the past three months     106
Name: count, dtype: int64
```

```
CONNECTION_activities_video_chat_p3m_combined
Not in the past three months     254
Few times a month or weekly      234
Less than monthly or monthly     202
Few times a week or daily        117
Name: count, dtype: int64
```

We combined the independent variables into categories of two options to reduce the number of options from 7 to 4

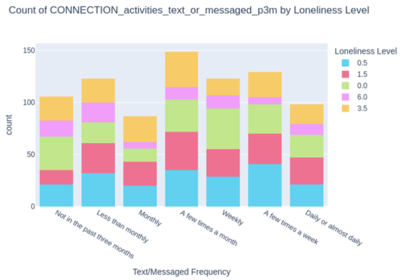
More narrowed down results

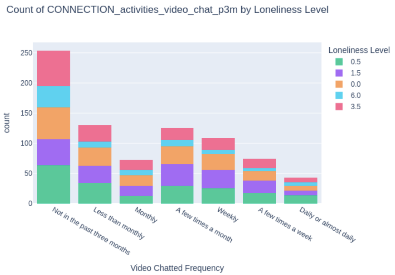
We selected "Few times a month or weekly" as the baseline (since it is a moderate level of frequency)

# Assumptions

- Converting the categorical variable to numerical values makes it not completely accurate since it assumes an exact number of days that they feel lonely (it is not a continuous variable but it being treated as one)
- How many days people feel lonely in a week is not reflective of how lonely they feel on average
- Converting independent variables into 4 categories of options instead of 7 is not reflective of how the surveyors actually responded to the question and narrows down the frequency of their activities

# Visualization of the Raw Data




Graph 1 - Text/Messaged

Graph 2 - Video Chatting

# Simple Linear Regression & Data Wrangling

We first tried to analyze the two factors separately with simple linear regression.

| | | | |
|---|---|---|---|
| **Dep. Variable:** | LONELY_direct_continuous | **R-squared:** | 0.009 |
| **Model:** | OLS | **Adj. R-squared:** | 0.005 |
| **Method:** | Least Squares | **F-statistic:** | 2.406 |
| **Date:** | Sun, 01 Dec 2024 | **Prob (F-statistic):** | 0.0660 |
| **Time:** | 16:53:34 | **Log-Likelihood:** | -1621.2 |
| **No. Observations:** | 807 | **AIC:** | 3250. |
| **Df Residuals:** | 803 | **BIC:** | 3269. |
| **Df Model:** | 3 | | |
| **Covariance Type:** | nonrobust | | |

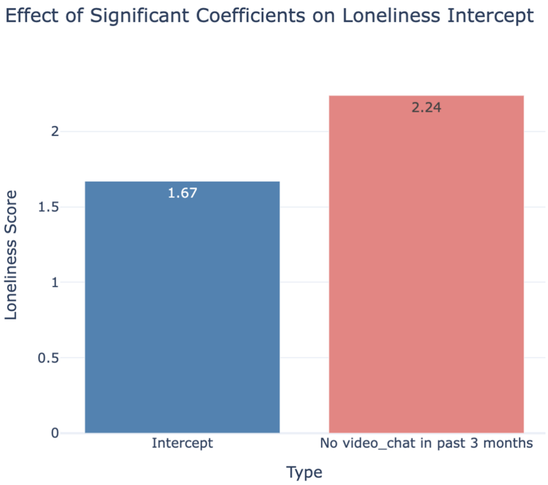
Loneliness vs. Video Chatting Frequency

| | | | |
|---|---|---|---|
| **Dep. Variable:** | LONELY_direct_continuous | **R-squared:** | 0.008 |
| **Model:** | OLS | **Adj. R-squared:** | 0.005 |
| **Method:** | Least Squares | **F-statistic:** | 2.230 |
| **Date:** | Sun, 01 Dec 2024 | **Prob (F-statistic):** | 0.0834 |
| **Time:** | 16:53:33 | **Log-Likelihood:** | -1621.5 |
| **No. Observations:** | 807 | **AIC:** | 3251. |
| **Df Residuals:** | 803 | **BIC:** | 3270. |
| **Df Model:** | 3 | | |
| **Covariance Type:** | nonrobust | | |

Loneliness vs. Text Messaging Frequency

The fitting of the model was poor, so we decided to analyze with a multilinear regression model with interactions.

# Final Result & Analysis




R-squared value = 0.032

The condition number = 26.4

Residual analysis suggests minimal bias

**Conclusions:**

- Treating texting and video chatting a few times a month or weekly as the baseline
- Comparing that intercept to no video chat in the past 3 months
- Final Result - Not video chatting in the past 3 months increases the days one feels lonely from 1.67 (when texting and video chatting a few times a month or weekly) to 2.24 days
- So, video chatting could be beneficial to reducing loneliness

*Reminder that this is not completely accurate due to assumptions mentioned previously

*Reminder that correlation $\neq$ causation

# Conclusion

**Connections**

- Aim to analyze how people's mental health is associated with different socializing forms

**Findings**

- Socializing and Depression:
    - The average depression score decreases when physical socialization $\geq 5$ min/day
    - Depression levels also drop when taking a walk with others.
    - Playing video games, however, may have an opposing effect on depression levels.
- Socializing and Loneliness:
    - Not video chatting in the past 3 months increases the days one feels lonely