An Investigation of Factors Influencing Emergency Healthcare Expenditures

due November 16, 2021 by 11:59 PM

Maggie Lundberg, Riya Mohan, and Izzy Kjaerulff

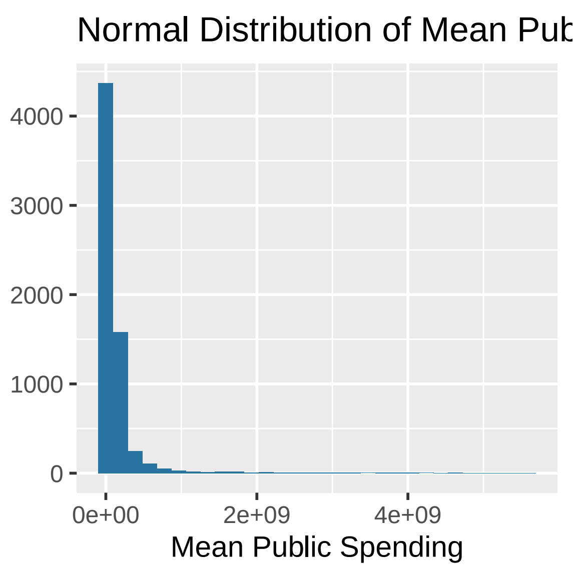
11/16/2021

Abstract

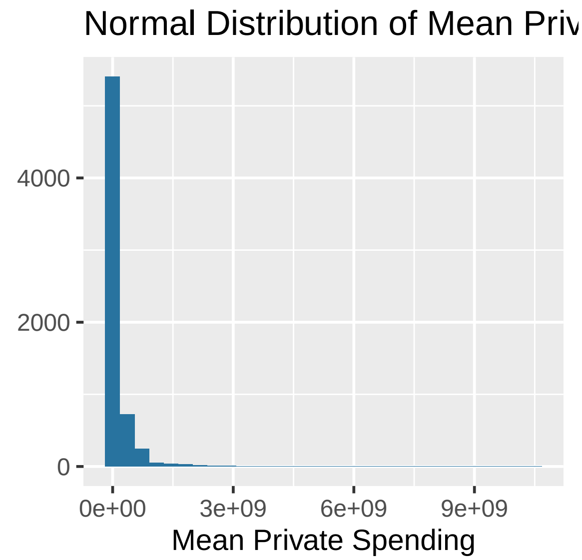
Nature of the Data

The data includes summary of both male and female spending as "both", so in order to perform an analysis on this data, we decided to exclude the both data points to avoid double counting? It is important to acknowledge that this data only included those who identified as either male or female, so this is not a complete representationof the population.

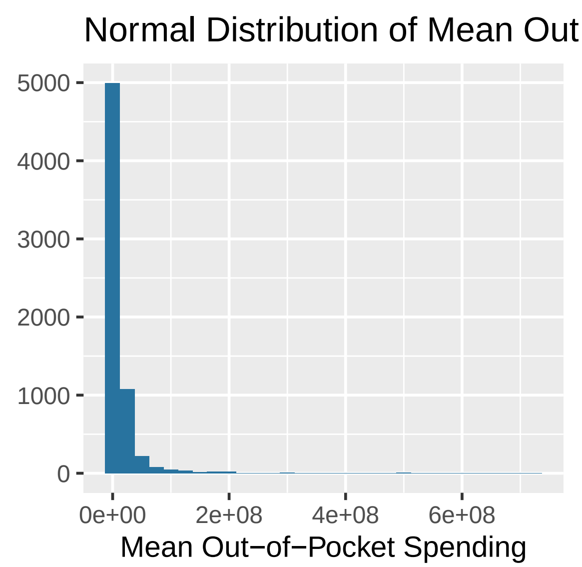
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

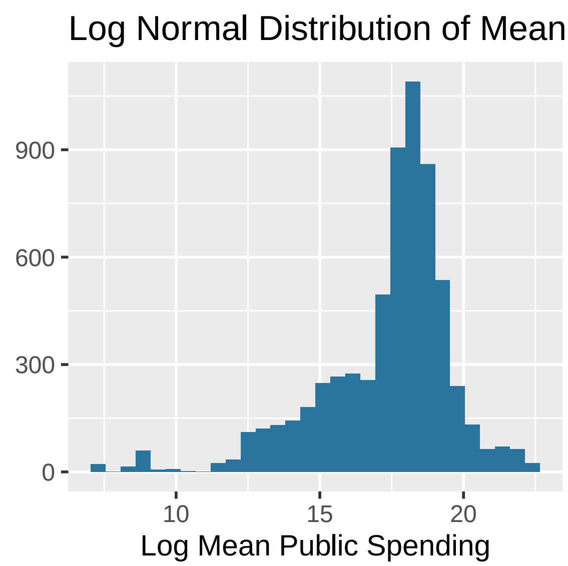



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

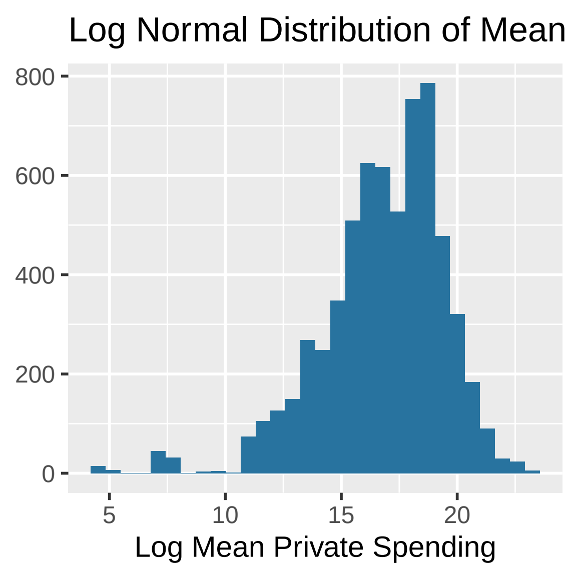


The normal distribution for public spending, private spending, and out-of-pocket pending all show a severe right skew in the data. Therefore, all three variables do not meet the normal distribution assumption needed for many tests, such as ANOVA; however, this can easily be resolved by applying a log transformation to the data to give a fairly normal distribution of the data.

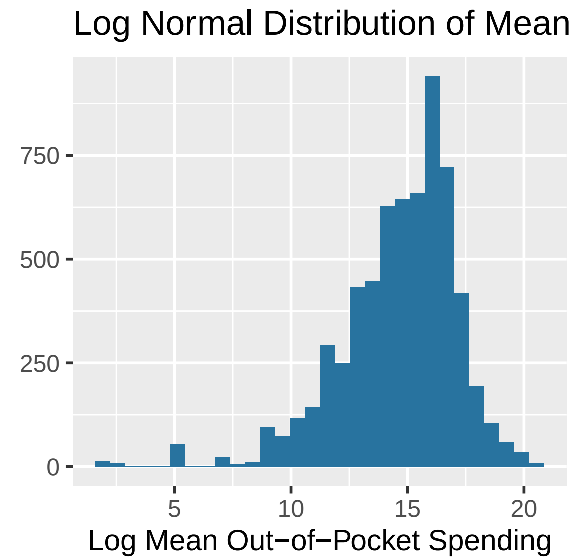
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



These graphs of the log distribution of the various spending means appear to be fairly normal in distribution, which means they meet the requirements to be used in various analyses. In order to convert to log scale, those with mean_all, mean_pub, mean_pri, and mean_oop equal to zero must be excluded.

Gender

Our first question in this analysis is if males and females spend a different amount of money on emergency services.

First this t-test looks at overall differences in log mean emergency department spending between males and females

```
##
##  Welch Two Sample t-test
##
## data:  spending_malefemale$lmean_all by spending_malefemale$sex
## t = 1.4247, df = 6219.5, p-value = 0.1543
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -0.0315862  0.1996079
## sample estimates:
## mean in group Female   mean in group Male
##             18.06275             17.97874
```

This t-test shows that for mean spending of all emergency services payment types, the p value of 0.1543 (95% CI -0.0315862, 0.1996079) indicates there is not a significant difference between male and female spending.

Next, we perform a t-test on each type of insurance to see if there is a difference in spending between males and females:

```
##
##  Welch Two Sample t-test
##
## data:  spending_malefemale$lmean_pub by spending_malefemale$sex
## t = 1.8142, df = 6201, p-value = 0.0697
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -0.00833746  0.21532602
```

```
## sample estimates:
## mean in group Female   mean in group Male
##             17.40512             17.30162
```

The t-test on emergency services spending for people who have public insurance indicates there is not a significant difference between male and female spending, with p value of 0.0697 (95% CI -0.00833746, 0.21532602).

```
##
##  Welch Two Sample t-test
##
## data:  spending_malefemale$lmean_pri by spending_malefemale$sex
## t = 0.70583, df = 6254.9, p-value = 0.4803
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -0.08283085  0.17603825
## sample estimates:
## mean in group Female   mean in group Male
##             16.82891             16.78231
```

The t-test on emergency services spending for people who have private insurance indicates there is not a significant difference between male and female spending, with p value of 0.4803 (95% CI -0.08283085, 0.17603825).

```
##
##  Welch Two Sample t-test
##
## data:  spending_malefemale$lmean_oop by spending_malefemale$sex
## t = 0.9799, df = 6230.6, p-value = 0.3272
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -0.0615859  0.1846904
## sample estimates:
## mean in group Female   mean in group Male
##             14.66032             14.59877
```

The t-test on emergency services spending for people who pay out of pocket indicates there is not a significant difference between male and female spending, with p value of 0.3272 (95% CI -0.0615859, 0.1846904).

The t-tests for each type of insurance indicate that there is not enough evidence to reject the null hypothesis that emergency department spending is the same for males and females who have public insurance, private insurance, or pay out of pocket, leading us to the conclusion that gender does not influence emergency spending in the forms of payment studied here.

Disease category and Emergency Spending

In order to determine emergency department spending based on disease type, an ANOVA test is performed due to the data for spending on the log scale being normally distributed, relatively similar variance, and independent.

The null hypothesis for this ANOVA test is that the overall mean of spending are the same for each disease category

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## agg_cause     14  19152  1368.0   521.9 <2e-16 ***
## Residuals   6365  16685     2.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the p-value here of <2e-16, these data or more extreme data it is highly unlikely the null hypothesis is true. Therefore, we perform step-down tests using a Holm correction for multiple comparisons.

```
## [1] 92
```

The step-down t tests indicate 92 disease category pairs are different out of 105, indicating most disease categories do differ in the amount of government spending by the emergency department. There is lots of variation!!

```
## # A tibble: 15 x 5
##    term                                   estimate std.error statistic   p.value
##    <chr>                                     <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                             17.7      0.0772    229.    0
##  2 agg_causeCancers                        -2.52     0.109     -23.1  2.71e-113
##  3 agg_causeCardiovascular diseases         1.29     0.109      11.8  7.26e- 32
##  4 agg_causeChronic respiratory diseases    1.40     0.111      12.7  1.65e- 36
##  5 agg_causeCommunicable and nutrition d~   1.73     0.109      15.9  1.05e- 55
##  6 agg_causeDiabetes and kidney diseases   -0.110    0.109      -1.01 3.13e-  1
##  7 agg_causeDigestive diseases              1.96     0.109      18.0  1.85e- 70
##  8 agg_causeEndocrine disorders            -1.15     0.109     -10.5  1.32e- 25
##  9 agg_causeInjuries                        2.68     0.109      24.5  3.15e-127
## 10 agg_causeMaternal and neonatal condit~  -3.12     0.130     -24.1  1.14e-122
## 11 agg_causeMusculoskeletal conditions      1.05     0.109       9.62 9.27e- 22
## 12 agg_causeNeurological disorders          0.970    0.109       8.88 8.41e- 19
## 13 agg_causeOther non-communicable disea~   1.60     0.109      14.6  1.33e- 47
## 14 agg_causePrevention and coordination    -3.00     0.109     -27.5  1.30e-157
## 15 agg_causeSkin and other sense organ d~   1.09     0.109       9.98 2.71e- 23
```

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

```
## [1] 0.5344174
```

Age

!! had to take out the observations with "All Ages" because I think it will just mess up the pairs but let me know what you think or whether you think there's anything we can do with that group

We wonder whether there is a correlation between government healthcare expenditures in the emergency department and age. The age variable is categorical, split into 19 groups that generally include 5 years each, apart from the first (<1 year) and last (85 plus) groups.

To address this question, we began by using an overall test with ANOVA.

Below is an overall test of the null hypothesis that all of the means for age groups across the years are equal, as opposed to the alternative that at least one mean is different.

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## age_group_name   18   4239  235.50   52.17 <2e-16 ***
## Residuals      6031  27226    4.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this F-test (ndf = 18, ddf = 6229), a significant difference among age groups was identified. Our p-value tells us that this data (or data more extreme) would be very unlikely if the null hypothesis were true because it shows statistical significance at an alpha well below 0.05. Therefore, we reject the null hypothesis that the mean expenditures for all age groups are equal.

To see which specific means may be different from one another, we used planned step-down tests with a Holm correction to minimize Type I errors.

```
## [1] 99
```

The pairwise t-tests used for our ANOVA step-down tests suggest that there are 99 different age pairs out of the 171 possible combinations. This tells us that more age pairs are different than are similar and that therefore the majority of age group pairs differ in terms of mean expenditures.

```
agefit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(lmean_all ~ age_group_name, data = spending_malefemale)
  tidy(agefit)
```

```
## # A tibble: 20 x 5
##    term                     estimate std.error statistic   p.value
##    <chr>                       <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                 15.4      0.120    128.   0
##  2 age_group_name1 to 4         1.20     0.168      7.13 1.13e- 12
##  3 age_group_name10 to 14       1.92     0.168     11.5  3.74e- 30
##  4 age_group_name15 to 19       2.59     0.168     15.4  1.04e- 52
##  5 age_group_name20 to 24       2.96     0.168     17.7  3.38e- 68
##  6 age_group_name25 to 29       3.09     0.168     18.4  5.11e- 74
##  7 age_group_name30 to 34       3.16     0.168     18.8  7.10e- 77
##  8 age_group_name35 to 39       3.23     0.168     19.2  2.76e- 80
##  9 age_group_name40 to 44       3.30     0.168     19.7  1.34e- 83
## 10 age_group_name45 to 49       3.32     0.168     19.8  9.53e- 85
## 11 age_group_name5 to 9         1.27     0.168      7.58 4.07e- 14
## 12 age_group_name50 to 54       3.27     0.168     19.5  4.33e- 82
## 13 age_group_name55 to 59       2.89     0.168     17.2  5.13e- 65
## 14 age_group_name60 to 64       2.70     0.168     16.1  5.96e- 57
## 15 age_group_name65 to 69       2.63     0.168     15.7  1.90e- 54
## 16 age_group_name70 to 74       2.48     0.168     14.8  1.01e- 48
## 17 age_group_name75 to 79       2.37     0.168     14.1  1.58e- 44
## 18 age_group_name80 to 84       2.32     0.168     13.8  7.53e- 43
## 19 age_group_name85 plus        2.39     0.168     14.2  3.11e- 45
## 20 age_group_nameAll Ages       5.84     0.166     35.1  5.41e-247
```

```
![](finalreport_template_files/figure-latex/age-expenditure-barplot-1.pdf)<!-- -->

![](finalreport_template_files/figure-latex/age-pub-barplot-1.pdf)<!-- -->

![](finalreport_template_files/figure-latex/age-pri-barplot-1.pdf)<!-- -->

![](finalreport_template_files/figure-latex/age-oop-barplot-1.pdf)<!-- -->


## Gender and Age Interaction
```

 # A tibble: 3 x 5

term estimate std.error statistic p.value

1 (Intercept) 17.3 0.0445 390. 0

2 sexMale -0.104 0.0574 -1.80 0.0712

3 age_group_id 0.00328 0.000887 3.70 0.000216

[1] 0.002338971

 # A tibble: 4 x 5

term estimate std.error statistic p.value

1 (Intercept) 17.3 0.0483 359. 0

2 sexMale -0.0803 0.0694 -1.16 0.247

3 age_group_id 0.00379 0.00123 3.08 0.00211

4 sexMale:age_group_id -0.00106 0.00177 -0.595 0.552

[1] 0.002237904

 # A tibble: 3 x 5

term estimate std.error statistic p.value

1 (Intercept) 17.1 0.0508 337. 0

2 sexMale -0.0463 0.0656 -0.707 4.80e- 1

3 age_group_id -0.0132 0.00101 -13.0 3.18e-38

[1] 0.02563492

A tibble: 4 x 5

term estimate std.error statistic p.value

1 (Intercept) 17.1 0.0552 310. 0

2 sexMale 0.00254 0.0793 0.0320 9.74e- 1

3 age_group_id -0.0121 0.00141 -8.60 1.02e-17

4 sexMale:age_group_id -0.00222 0.00203 -1.10 2.73e- 1

[1] 0.02566573

A tibble: 3 x 5

term estimate std.error statistic p.value

1 (Intercept) 14.9 0.0485 308. 0

2 sexMale -0.0613 0.0626 -0.980 3.27e- 1

3 age_group_id -0.0113 0.000967 -11.7 3.10e-31

[1] 0.02080634

A tibble: 4 x 5

term estimate std.error statistic p.value

1 (Intercept) 14.9 0.0526 283. 0

2 sexMale -0.0181 0.0757 -0.239 8.11e- 1

3 age_group_id -0.0103 0.00134 -7.70 1.59e-14

4 sexMale:age_group_id -0.00197 0.00193 -1.02 3.09e- 1

[1] 0.02081155

```
In order to test the possibility that there is a joint interaction of gender and age, a main effects and
```

Age and Disease Type Interaction

A tibble: 16 x 5

term estimate std.error statistic p.value

1 (Intercept) 17.1 0.0780 219. 0

2 agg_causeCancers -2.68 0.109 -24.6 3.75e-128

3 agg_causeCardiovascular diseases 1.16 0.109 10.7 1.92e- 26

4 agg_causeChronic respiratory diseases 1.27 0.110 11.5 2.51e- 30

5 agg_causeCommunicable and nutrition d~ 1.59 0.109 14.6 1.68e- 47

6 agg_causeDiabetes and kidney diseases 0.0502 0.109 0.462 6.44e- 1

7 agg_causeDigestive diseases 1.64 0.109 15.1 1.30e- 50

8 agg_causeEndocrine disorders -1.27 0.109 -11.7 2.94e- 31

9 agg_causeInjuries 1.97 0.109 18.1 1.68e- 71

10 agg_causeMaternal and neonatal condit~ -3.14 0.129 -24.4 1.03e-125

11 agg_causeMusculoskeletal conditions 0.806 0.109 7.42 1.32e- 13

12 agg_causeNeurological disorders 0.760 0.109 7.00 2.77e- 12

13 agg_causeOther non-communicable disea~ 1.23 0.109 11.3 1.65e- 29

14 agg_causePrevention and coordination -3.06 0.109 -28.2 3.71e-165

15 agg_causeSkin and other sense organ d~ 0.863 0.109 7.95 2.22e- 15

16 age_group_id 0.00334 0.000623 5.35 9.00e- 8

[1] 0.506887

```
# A tibble: 30 x 5
term estimate std.error statistic p.value

1 (Intercept) 17.3 0.0907 191. 0
2 agg_causeCancers -3.14 0.128 -24.5 1.27e-126
3 agg_causeCardiovascular diseases 0.630 0.128 4.91 9.52e- 7
4 agg_causeChronic respiratory diseases 1.06 0.129 8.22 2.35e- 16
5 agg_causeCommunicable and nutrition d~ 1.42 0.128 11.0 4.05e- 28
6 agg_causeDiabetes and kidney diseases -0.244 0.128 -1.90 5.70e- 2
7 agg_causeDigestive diseases 1.48 0.128 11.6 1.19e- 30
8 agg_causeEndocrine disorders -1.44 0.128 -11.2 6.78e- 29
9 agg_causeInjuries 1.72 0.128 13.4 2.25e- 40
10 agg_causeMaternal and neonatal condit~ -2.34 0.154 -15.2 4.33e- 51
# ... with 20 more rows

[1] 0.529351
```

A tibble: 16 x 5
term estimate std.error statistic p.value

1 (Intercept) 16.4 0.0903 182. 0

2 agg_causeCancers -2.24 0.126 -17.8 1.48e- 69

3 agg_causeCardiovascular diseases 1.62 0.126 12.9 1.75e- 37

4 agg_causeChronic respiratory diseases 1.71 0.127 13.4 1.13e- 40

5 agg_causeCommunicable and nutrition d~ 2.05 0.126 16.3 2.02e- 58

6 agg_causeDiabetes and kidney diseases -0.245 0.126 -1.94 5.19e- 2

7 agg_causeDigestive diseases 2.43 0.126 19.3 6.39e- 81

8 agg_causeEndocrine disorders -0.793 0.126 -6.31 2.99e- 10

9 agg_causeInjuries 3.44 0.126 27.3 9.02e-156

10 agg_causeMaternal and neonatal condit~ -3.16 0.149 -21.2 2.27e- 96

11 agg_causeMusculoskeletal conditions 1.50 0.126 11.9 2.36e- 32

12 agg_causeNeurological disorders 1.32 0.126 10.5 1.29e- 25

13 agg_causeOther non-communicable disea~ 2.03 0.126 16.2 1.27e- 57

14 agg_causePrevention and coordination -2.64 0.126 -21.0 2.18e- 94

15 agg_causeSkin and other sense organ d~ 1.41 0.126 11.2 8.06e- 29

16 age_group_id -0.0131 0.000722 -18.2 4.45e- 72

[1] 0.5054947

```
# A tibble: 30 x 5
term estimate std.error statistic p.value

1 (Intercept) 16.7 0.106 157. 0
2 agg_causeCancers -2.77 0.151 -18.4 1.21e-73
3 agg_causeCardiovascular diseases 1.05 0.151 6.96 3.67e-12
4 agg_causeChronic respiratory diseases 1.50 0.152 9.90 6.18e-23
5 agg_causeCommunicable and nutrition di~ 1.82 0.151 12.1 2.51e-33
6 agg_causeDiabetes and kidney diseases -0.647 0.151 -4.30 1.74e- 5
7 agg_causeDigestive diseases 2.31 0.151 15.3 5.24e-52
8 agg_causeEndocrine disorders -1.04 0.151 -6.94 4.46e-12
9 agg_causeInjuries 3.23 0.151 21.4 1.91e-98
10 agg_causeMaternal and neonatal conditi~ -2.78 0.181 -15.4 2.21e-52
# ... with 20 more rows

[1] 0.5149051
```

 # A tibble: 16 x 5

term estimate std.error statistic p.value

1 (Intercept) 14.7 0.0850 173. 0

2 agg_causeCancers -2.75 0.118 -23.3 6.48e-115

3 agg_causeCardiovascular diseases 0.975 0.118 8.23 2.17e- 16

4 agg_causeChronic respiratory diseases 1.11 0.120 9.26 2.68e- 20

5 agg_causeCommunicable and nutrition d~ 1.67 0.118 14.1 1.02e- 44

6 agg_causeDiabetes and kidney diseases -0.428 0.118 -3.62 3.03e- 4

7 agg_causeDigestive diseases 1.83 0.118 15.5 5.56e- 53

8 agg_causeEndocrine disorders -0.917 0.118 -7.75 1.08e- 14

9 agg_causeInjuries 2.65 0.118 22.4 9.40e-107

10 agg_causeMaternal and neonatal condit~ -3.52 0.141 -25.1 1.97e-132

11 agg_causeMusculoskeletal conditions 0.888 0.118 7.51 6.94e- 14

12 agg_causeNeurological disorders 0.736 0.118 6.22 5.30e- 10

13 agg_causeOther non-communicable disea~ 1.57 0.118 13.3 1.30e- 39

14 agg_causePrevention and coordination -3.34 0.118 -28.2 1.73e-165

15 agg_causeSkin and other sense organ d~ 0.841 0.118 7.10 1.37e- 12

16 age_group_id -0.0112 0.000680 -16.5 3.10e- 60

[1] 0.5161683

 # A tibble: 30 x 5

term estimate std.error statistic p.value

1 (Intercept) 14.9 0.100 149. 0

2 agg_causeCancers -3.18 0.142 -22.4 2.10e-107

3 agg_causeCardiovascular diseases 0.404 0.142 2.85 4.34e- 3

4 agg_causeChronic respiratory diseases 0.893 0.143 6.26 4.13e- 10

5 agg_causeCommunicable and nutrition d~ 1.50 0.142 10.6 5.92e- 26

6 agg_causeDiabetes and kidney diseases -0.797 0.142 -5.63 1.91e- 8

7 agg_causeDigestive diseases 1.68 0.142 11.9 3.63e- 32

8 agg_causeEndocrine disorders -1.14 0.142 -8.08 7.94e- 16

9 agg_causeInjuries 2.39 0.142 16.9 1.79e- 62

10 agg_causeMaternal and neonatal condit~ -3.09 0.170 -18.2 7.93e- 72

 # ... with 20 more rows

[1] 0.5261024

```
##Spending Over Time

!! I kinda like this but idk if it adds anything but it is fun, need to make the words smaller so you ca
![](finalreport_template_files/figure-latex/spending_time_graph-1.pdf)<!-- -->

!! can we divide this to have a predictor for each year?
```

 # A tibble: 2 x 5

term estimate std.error statistic p.value

1 (Intercept) -92.1 18.8 -4.89 0.00000102

2 year_id 0.0548 0.00936 5.85 0.00000000516

“‘