Lab 1 - Self-Rated Health GSS Data Analysis

Christine Lucille Kuryla

2024-10-07

Contents

In this lab, I looked at the `health` variable in the GSS dataset. There will be an unpooled regression comparison across time periods done, with some age-period-cohort sprinkled in, and a final extra analysis on the change in the predictive power of age on SRH over time.

https://gssdataexplorer.norc.org/variables/437/vshow

Question on survey: "Would you say your own health, in general, is excellent, good, fair, or poor?" Recoded as: 4, 3, 2, 1, respectively

Variables of interest

The GSS variables chosen for this analysis were chosen to complement the main variable of interest: self-rated health (SRH) `health`, age/birthyear/year of survey for comparison reasons and a brief age-period-cohort exploration, as well as several variables that could potentially affect SRH, such as sex, self-rated happiness, self-assessed interest in their life, years of education, political views, self-assessed class, and financial satisfaction.

Variables

The variables chosen for this data set were:

- `year`: year of survey

- `cohort`: birth year of participant
- `age`: age of participant
- `health`: self-rated health
- `sex`: sex of participant
- `happy`: self-reported happiness
- `life`: self-assessed life interest rating
- `educ`: years of education
- `polviews`: political views
- `class`: self assessed class
- `satfin`: financial satisfaction

Variable details

"health"

https://gssdataexplorer.norc.org/variables/437/vshow?back=variableList

Variable: Health Module: Core Gss Tags: Health

"Would you say your own health, in general, is excellent, good, fair, or poor?"

-100 .i: Inapplicable
-99 .n: No answer
-98 .d: Do not Know/Cannot Choose
-97 .s: Skipped on Web
1 Excellent
2 Good
3 Fair
4 Poor

Note we will recode this data to reverse it to be more intuitive, such that for our analyses, higher numbers will indicate better self-rated health:

4 Excellent
3 Good
2 Fair
1 Poor

"year"

https://gssdataexplorer.norc.org/variables/1/vshow?back=variableList

Year of survey

Variable: Year

GSS year for this respondent

"cohort"

https://gssdataexplorer.norc.org/variables/5507/vshow?back=variableList

Year of birth

Birth cohort of respondent.

"happy"

general happiness

“Taken all together, how would you say things are these days–would you say that you are very happy, pretty happy, or not too happy?”

-100 .i: Inapplicable
-99 .n: No answer
-98 .d: Do not Know/Cannot Choose
-97 .s: Skipped on Web
1 Very happy
2 Pretty happy
3 Not too happy

Note we will recode this data to reverse it to be more intuitive, such that for our analyses, higher numbers will indicate higher happiness:

3 Very happy
2 Pretty happy
1 Not too happy

“sex”

https://gssdataexplorer.norc.org/variables/81/vshow?back=variableList

respondent’s sex

-100 .i: Inapplicable
-99 .n: No answer
-98 .d: Do not Know/Cannot Choose
-97 .s: Skipped on Web
1 MALE
2 FEMALE

“educ”

https://gssdataexplorer.norc.org/variables/55/vshow

highest year of school completed

Questions associated with this variable: ASK ALL PARTS OF QUESTION ABOUT RESPONDENT BEFORE GOING ON TO ASK ABOUT R’S FATHER; AND THEN R’S MOTHER; THEN R’S SPOUSE, IF R IS CURRENTLY MARRIED. A. What is the highest grade in elementary school or high school that (you/your father/ your mother/your [husband/wife]) finished and got credit for? CODE EXACT GRADE. B. IF FINISHED 9th-12th GRADE OR DK*: Did (you/he/she) ever get a high school diploma or a GED certificate? [SEE D BELOW.] [See REMARKS] C. Did (you/he/she) complete one or more years of college for credit–not including schooling such as business college, technical or vocational school? IF YES: How many years did (you/he/she) complete? Do you (Does [he/she]) have any college degrees? (IF YES: What degree or degrees?) CODE HIGHEST DEGREE EARNED.

-99 .n: No answer
-98 .d: Do not Know/Cannot Choose
0 No formal schooling 1 1st grade
2 2nd grade 13
. . . 19 7 or more years of college 20 8 or nore years of college

“life”

is life exciting or dull

https://gssdataexplorer.norc.org/variables/438/vshow

“In general, do you find life exciting, pretty routine, or dull?”

100 .i: Inapplicable
-99 .n: No answer
-98 .d: Do not Know/Cannot Choose
-97 .s: Skipped on Web
1 Exciting
2 Routine 3 Dull

Note we will recode this data to reverse it to be more intuitive, such that for our analyses, higher numbers will indicate a more exciting life:

3 Exciting
2 Routine 1 Dull

"satfin"

satisfaction with financial situation

https://gssdataexplorer.norc.org/variables/572/vshow

"We are interested in how people are getting along financially these days. So far as you and your family are concerned, would you say that you are pretty well satisfied with your present financial situation, more or less satisfied, or not satisfied at all?"

-100 .i: Inapplicable
-99 .n: No answer
-98 .d: Do not Know/Cannot Choose
-97 .s: Skipped on Web
1 Pretty well satisfied
2 More or less satisfied
3 Not satisfied at all

Note we will recode this data to reverse it to be more intuitive, such that for our analyses, higher numbers will indicate more financial satisfaction:

3 Pretty well satisfied
2 More or less satisfied
1 Not satisfied at all

"polviews"

think of self as liberal or conservative

https://gssdataexplorer.norc.org/variables/178/vshow

"We hear a lot of talk these days about liberals and conservatives. I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal–point 1–to extremely conservative–point 7. Where would you place yourself on this scale?"

-100 .i: Inapplicable
-99 .n: No answer
-98 .d: Do not Know/Cannot Choose
-97 .s: Skipped on Web
1 Extremely liberal
2 Liberal 3 Slightly liberal
4 Moderate, middle of the road
5 Slightly conservative
6 Conservative
7 Extremely conservative

Fetch, load, clean, explore data

```
# Feel free to modify to play with more covariates and variables.

#install.packages('gssr', repos =  c('https://kjhealy.r-universe.dev', 'https://cloud.r-project.org'))
# install.packages('gssrdoc', repos = c('https://kjhealy.r-universe.dev', 'https://cloud.r-project.org'

library(gssr)
library(gssrdoc)

data("gss_all") # this file is big!

# It's a bit excessive to download the entire GSS dataset every time we knit, so lets just save some va

data_gss <- as.data.frame(gss_all) %>%
  select("year",       # year of survey
         "cohort",     # birthyear
         "age",        # age at time of survey
         "health",     # self-rated health
         "sex",        # sex
         "happy",      # self-rated happiness
         "life",       # is life exciting or dull
         "educ",       # years of education
         "polviews",   # 1 extremely liberal, 4 moderate, 7 extremely conservative
         "class",     # 1 lower, 2 middle, 3 working, 4 upper, 5 no class
         "satfin"     # 1 pretty well satisfied, 2 more or less satisfied, 3 not satisfied at all
         )

write_csv(data_gss, "data/extracted_gss_variables.csv")
```

Load and clean data

Here we'll load our data, clean some unwanted values, and recode the unintuitive variables.

```
data_gss <- read_csv("data/extracted_gss_variables.csv") %>%
  filter(cohort != 9999) %>%
  filter(year > 0 , age > 0, health > 0, sex > 0, happy > 0, life > 0, educ > 0 , polviews > 0, class >
  na.omit() %>%
  mutate(health = 5 - health)  %>%  # reverse the coding so it's more intuitive (higher number for exce
  mutate(happy = 4 - happy) %>% # same
  mutate(life = 4 - life) %>% # reverse again, these variables tend to be unintuitively ordered!!!
  mutate(satfin = 4 - satfin) # %>%
```

```
## Rows: 72390 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (11): year, cohort, age, health, sex, happy, life, educ, polviews, class...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
  # mutate(polviews_extreme = case_match(polviews,
  #                                      7 ~ 4,
  #                                      6 ~ 3,
  #                                      5 ~ 2,
```

```
  #                                   4 ~ 1,
  #                                   3 ~ 2,
  #                                   2 ~ 3,
  #                                   1 ~ 4
  #                                   ))
```

Variable EDA

Here we will do a brief exploration of the data to get a sense of it.

Peek at GSS df

```
head(data_gss, n = 10)
```

```
## # A tibble: 10 x 11
##     year cohort   age health   sex happy  life  educ polviews class satfin
##    <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl> <dbl>  <dbl>
##  1  1974   1953    21      4     1     3     3    14        4     2      3
##  2  1974   1933    41      3     1     3     3    16        5     3      2
##  3  1974   1891    83      3     2     3     2    10        6     2      3
##  4  1974   1905    69      3     2     2     2    10        6     3      3
##  5  1974   1916    58      2     2     2     2    12        6     2      2
##  6  1974   1944    30      3     1     2     3    16        5     3      2
##  7  1974   1926    48      3     1     3     3    17        5     3      3
##  8  1974   1907    67      4     1     3     3    10        5     2      3
##  9  1974   1920    54      4     2     3     3    11        6     2      1
## 10  1974   1885    89      3     1     2     2     6        4     4      3
```

```
glimpse(data_gss)
```

```
## Rows: 39,907
## Columns: 11
## $ year     <dbl> 1974, 1974, 1974, 1974, 1974, 1974, 1974, 1974, 1974, 1974, 1~
## $ cohort   <dbl> 1953, 1933, 1891, 1905, 1916, 1944, 1926, 1907, 1920, 1885, 1~
## $ age      <dbl> 21, 41, 83, 69, 58, 30, 48, 67, 54, 89, 71, 27, 30, 23, 63, 7~
## $ health   <dbl> 4, 3, 3, 3, 2, 3, 3, 4, 4, 3, 2, 2, 2, 4, 1, 1, 1, 2, 1, 1, 4~
## $ sex      <dbl> 1, 1, 2, 2, 2, 1, 1, 1, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2~
## $ happy    <dbl> 3, 3, 3, 2, 2, 2, 3, 3, 3, 2, 3, 2, 2, 3, 3, 2, 3, 1, 2, 1, 3~
## $ life     <dbl> 3, 3, 2, 2, 2, 3, 3, 3, 3, 2, 2, 2, 2, 3, 1, 1, 3, 1, 2, 1, 3~
## $ educ     <dbl> 14, 16, 10, 10, 12, 16, 17, 10, 11, 6, 5, 11, 12, 12, 8, 8, 1~
## $ polviews <dbl> 4, 5, 6, 6, 6, 5, 5, 5, 6, 4, 2, 4, 4, 4, 4, 7, 2, 4, 6, 6, 3~
## $ class    <dbl> 2, 3, 2, 3, 2, 3, 3, 2, 2, 4, 3, 3, 3, 2, 1, 4, 3, 4, 3, 4, 1~
## $ satfin   <dbl> 3, 2, 3, 3, 2, 2, 3, 3, 1, 3, 3, 1, 1, 3, 1, 3, 1, 3, 3, 3, 1~
```

Histograms

Let's make a few quick histograms to get a sense of the survey response distributions.

```
# Tidyverse and flexible number of histograms

library(patchwork)

# Create list to store ggplot objects
plot_list <- list()
```

```r
# Loop through column names and create plots
for (var in colnames(data_gss)) {
  if (is.numeric(data_gss[[var]])) {  # Only create histograms for numeric columns
    p <- ggplot(data_gss, aes(x = .data[[var]])) +  # Use .data[[var]] for tidy evaluation
      geom_histogram(bins = 20, fill = "pink", color = "hotpink") +
      theme_minimal()

    plot_list[[var]] <- p
  }
}

# Combine all plots using patchwork and add a title
combined_plot <- wrap_plots(plot_list, ncol = 3) +
  plot_annotation(title = "Histograms of Survey Responses")

# Display the combined plot
print(combined_plot)
```

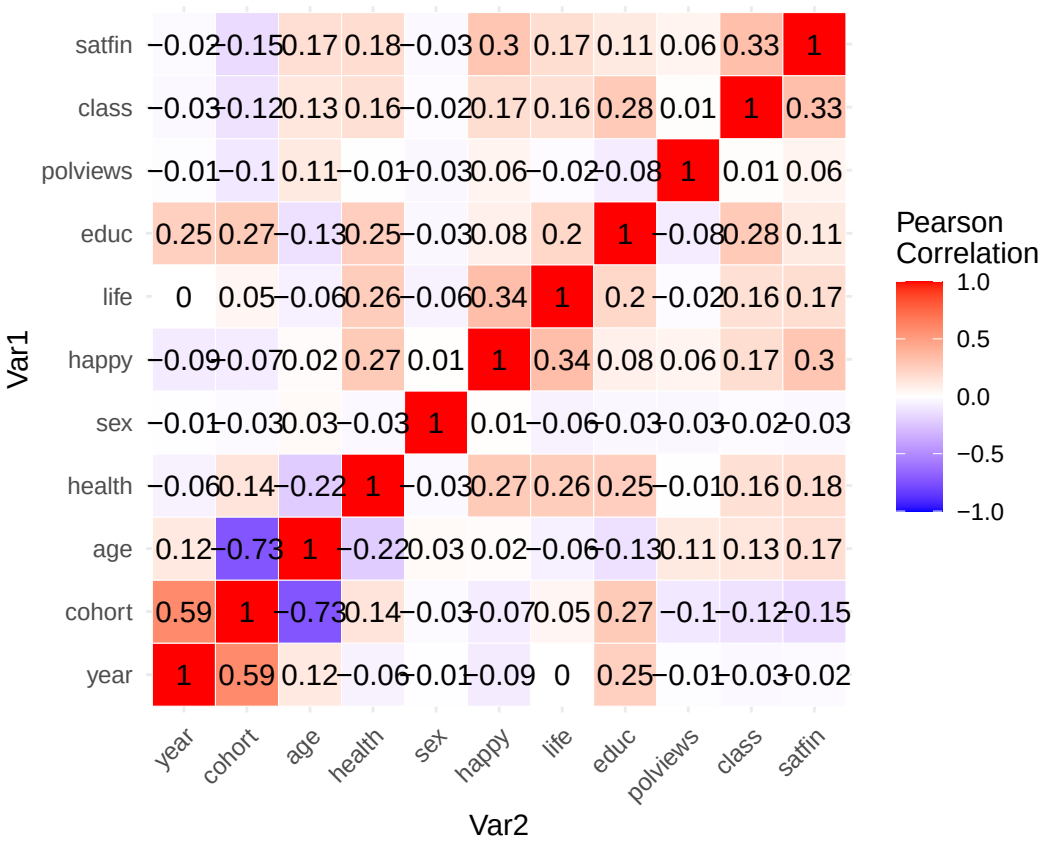
Correlation Heatmap

Just to get some idea and familiarize ourselves with the data.

```r
correlation_auto <- cor(data_gss)
knitr::kable(correlation_auto)
```

	year	cohort	age	health	sex	happy	life	educ	polviews	class	satfin
year	1.0000000	0.5892590	0.1168349	-0.0556566	-0.0064614	-0.0858322	-0.0006779	0.2493867	-0.0138046	-0.0269263	-0.0222855
cohort	0.5892590	1.0000000	-0.7335649	-0.1394829	-0.0263296	-0.0749725	-0.0515352	0.2729045	-0.0985640	-0.1225891	-0.1548524
age	0.1168349	-0.7335649	1.0000000	-0.2182736	-0.0269300	-0.0199591	-0.0639190	-0.1256863	0.1095461	0.1280408	0.1716037
health	-0.0556566	-0.1394829	-0.2182736	1.0000000	-0.0268431	-0.2724373	-0.2638685	-0.2537028	-0.0060678	0.1649492	0.1779453
sex	-0.0064614	-0.0263296	-0.0269300	-0.0268431	1.0000000	0.0069883	-0.0563487	-0.0316953	-0.0324606	-0.0160037	-0.0312474
happy	-0.0858322	-0.0749725	0.0199591	0.2724373	0.0069883	1.0000000	0.3378965	0.0763397	0.0560911	0.1747089	0.2958062
life	-0.0006779	0.0515352	-0.0639190	-0.2638685	-0.0563487	0.3378965	1.0000000	0.2040522	-0.0154497	0.1636502	0.1742316
educ	0.2493867	0.2729045	-0.1256863	-0.2537028	-0.0316953	0.0763397	0.2040522	1.0000000	-0.0755635	0.2778398	0.1112459
polviews	-0.0138046	-0.0985640	0.1095461	0.0060678	-0.0324606	0.0560911	-0.0154497	-0.0755635	1.0000000	0.0117992	0.0587300
class	-0.0269263	-0.1225891	0.1280408	0.1649492	-0.0160037	0.1747089	0.1636502	0.2778398	0.0117992	1.0000000	0.3346243
satfin	-0.0222855	-0.1548524	0.1716037	0.1779453	-0.0312474	0.2958062	0.1742316	0.1112459	0.0587300	0.3346243	1.0000000

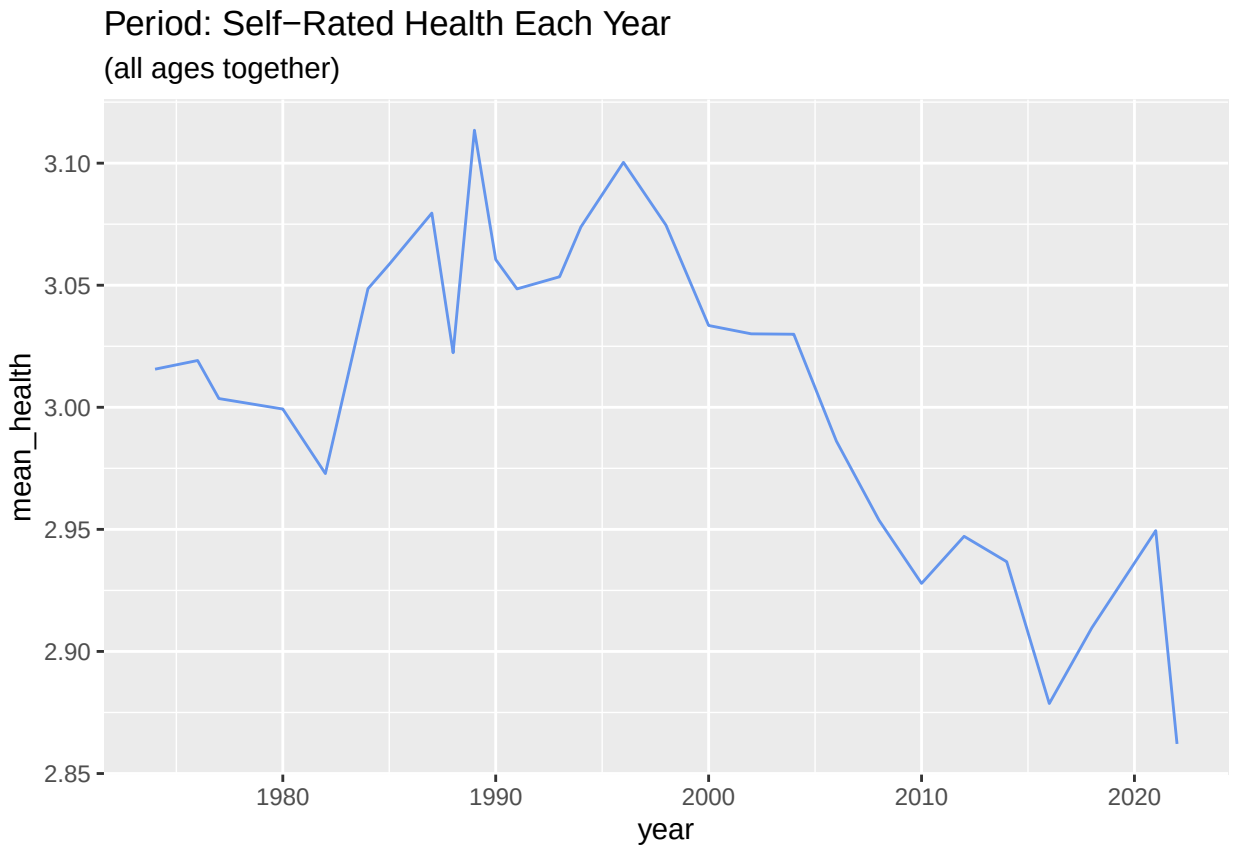
```
# Create a heatmap

#upper_tri <- matrixcalc::upper.triangle(correlation_auto)
#melted_cormat <- reshape2::melt(upper_tri, na.rm = TRUE)

melted_cormat <- reshape2::melt(cor(data_gss), na.rm = TRUE)

ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +
  theme_minimal()+
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
     hjust = 1))+
 coord_fixed() +
  geom_text(aes(Var2, Var1, label = if_else(value != 0, as.character(round(value, digits = 2)), " ")))
```



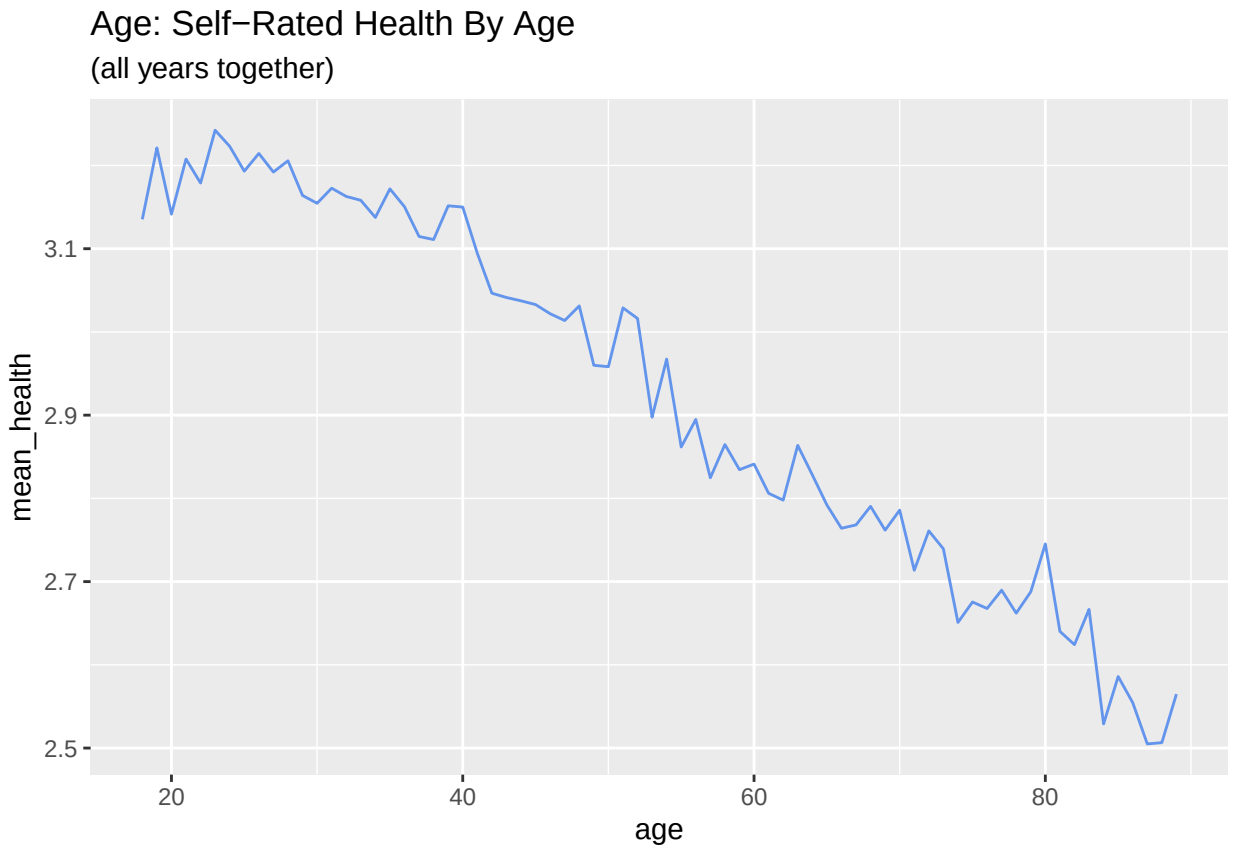
Plots of variables of interest

Now let's take a look at the trend of people's self-rated health over time, as well as by age, and by birthyear.

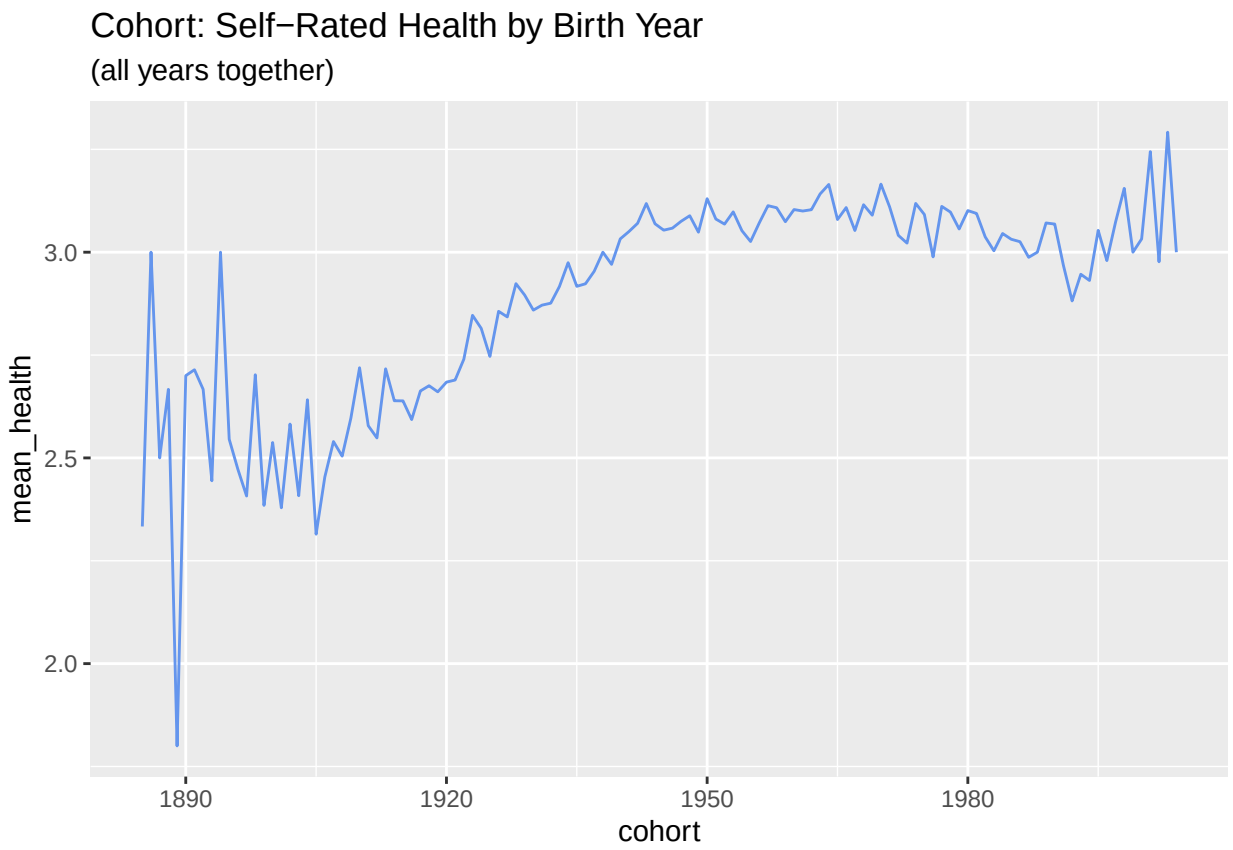
```
# Mean health over time, age, and birth year

data_gss %>%
  group_by(year) %>%
  summarize(mean_health = mean(health)) %>%
  ggplot(aes(x = year, y = mean_health)) +
  geom_line(color = "cornflowerblue") +
  labs(title = "Period: Self-Rated Health Each Year",
       subtitle = "(all ages together)" )
```



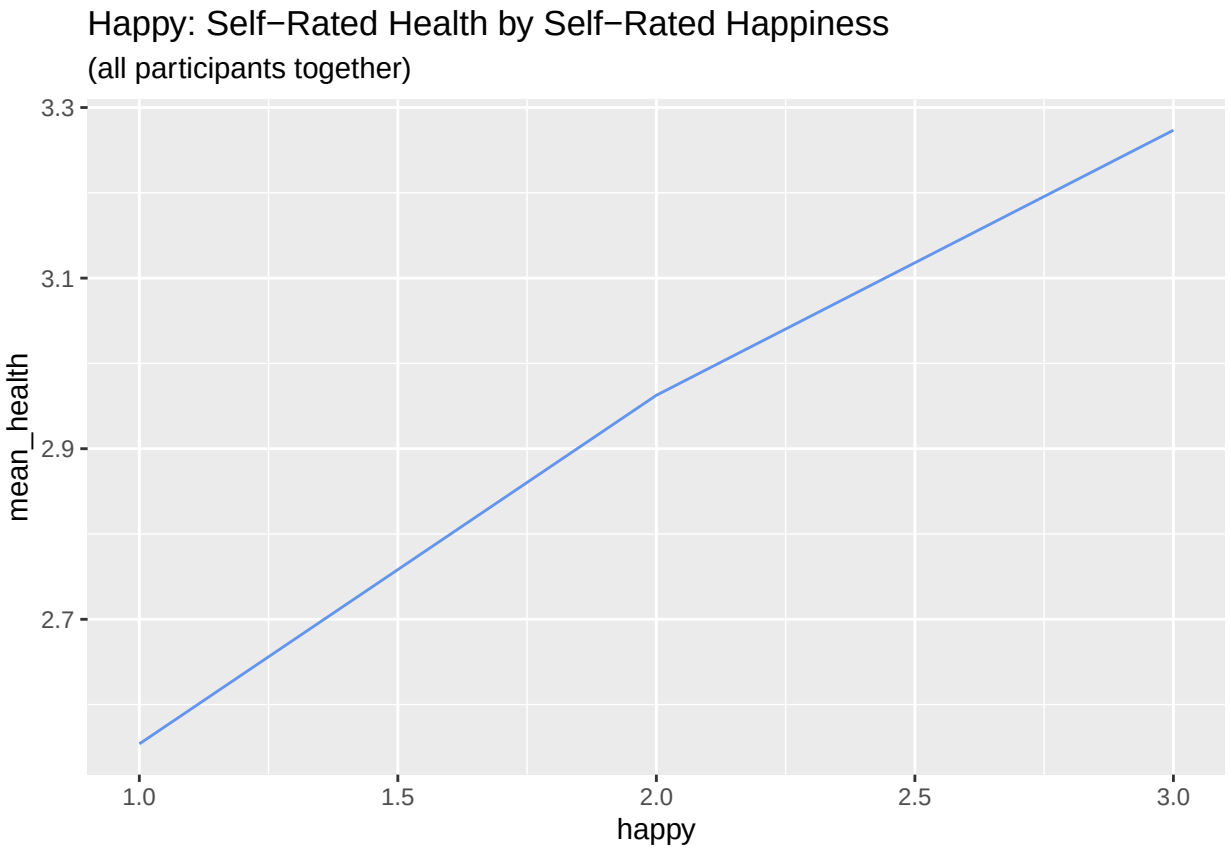

```
data_gss %>%
  group_by(age) %>%
  summarize(mean_health = mean(health)) %>%
  ggplot(aes(x = age, y = mean_health)) +
  geom_line(color = "cornflowerblue") +
  labs(title = "Age: Self-Rated Health By Age" ,
       subtitle = "(all years together)" )
```




```
data_gss %>%
  group_by(cohort) %>%
  summarize(mean_health = mean(health)) %>%
  ggplot(aes(x = cohort, y = mean_health)) +
  geom_line(color = "cornflowerblue") +
  labs(title = "Cohort: Self-Rated Health by Birth Year",
       subtitle = "(all years together)" )
```




```
data_gss %>%
  group_by(happy) %>%
  summarize(mean_health = mean(health)) %>%
  ggplot(aes(x = happy, y = mean_health)) +
  geom_line(color = "cornflowerblue") +
  labs(title = "Happy: Self-Rated Health by Self-Rated Happiness",
       subtitle = "(all participants together)" )
```



Comparing SRH predictors in the 1980s vs 2010s

We're going to explore which factors influence self-rated health in different time periods. We will start looking at the 1980s and compare it to the 2010s. Then we will look at it in a more fine-grained manner.

Regression for 1980s

First we will run a regression on the 1980s by subsetting the data.

As we can see below, all factors seem to statistically significantly predict self-rated health, except for self-rated class and political views. Collectively, these variables explain about 21% of the variance. Age is negatively associated with SRH, which is reasonable, as older people tend to have worse health. Being female is associated with higher SRH, as are self-rated happiness, life satisfaction, higher education, and higher financial satisfaction. These are intuitively reasonable, as healthier people are probably happier (or at least, happy people rate their health higher), similarly with life satisfaction. Higher education is often associated with higher objective and subjective health. Additionally, higher financial satisfaction is reasonably associated with higher SRH because people can afford to take care of their health. The F-statistic on this regression is 304.9 with a p-value less than 10^-16, which is highly significant and implies that this model is statistically significantly different from no effects.

```
# Run a regression on the variables of interest, subsetted into the 1980-1989

lm(health ~ age + as.factor(sex) + happy + life + educ + polviews + class + satfin,
   data = (data_gss %>% filter(year >= 1980 & year <= 1989))
   ) %>%
  summary()
```

```
##
## Call:
## lm(formula = health ~ age + as.factor(sex) + happy + life + educ +
##     polviews + class + satfin, data = (data_gss %>% filter(year >=
##     1980 & year <= 1989)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54819 -0.46496  0.01872  0.57542  2.13072
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.7861187  0.0609334  29.313   <2e-16 ***
## age             -0.0120483  0.0004893 -24.623   <2e-16 ***
## as.factor(sex)2 -0.0402847  0.0159712  -2.522   0.0117 *
## happy            0.2216346  0.0138326  16.023   <2e-16 ***
## life             0.1702987  0.0149377  11.401   <2e-16 ***
## educ             0.0491659  0.0028809  17.066   <2e-16 ***
## polviews         0.0033600  0.0059686   0.563   0.5735
## class            0.0218612  0.0136300   1.604   0.1088
## satfin           0.1149901  0.0116907   9.836   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7595 on 9291 degrees of freedom
## Multiple R-squared:  0.2069, Adjusted R-squared:  0.2062
## F-statistic: 302.9 on 8 and 9291 DF,  p-value: < 2.2e-16
```

Regression for 2010s

Now, we will run a regression on the 2000s/2010s by subsetting the data again.

As we can see below, all factors seem to statistically significantly predict self-rated health, except for sex. Collectively, these variables explain about 19% of the variance. Age is sill negatively associated with SRH. Variables associated with higher SRH include self-rated happiness, life satisfaction, more conservative political views, higher education, higher self-rated class, and higher financial satisfaction. The F-statistic on this regression is 425.7 with a p-value less than 10^-16, which is highly significant and implies that this model is statistically significantly different from no effects.

```
# Run a regression on the variables of interest, subsetted into the 2000s/2010s years

lm(health ~ age + as.factor(sex) + happy + life + educ + polviews + class + satfin,
   data = (data_gss %>% filter(year >= 2010 & year <= 2019))
   ) %>%
  summary()
```

```
##
## Call:
## lm(formula = health ~ age + as.factor(sex) + happy + life + educ +
##     polviews + class + satfin, data = (data_gss %>% filter(year >=
##     2010 & year <= 2019)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.57157 -0.47474  0.02747  0.53103  2.31362
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.0495256  0.0665043  15.781  < 2e-16 ***
## age             -0.0070728  0.0005152 -13.728  < 2e-16 ***
## as.factor(sex)2  0.0287843  0.0176578   1.630  0.10312
## happy            0.2183796  0.0147618  14.794  < 2e-16 ***
## life             0.2046324  0.0161011  12.709  < 2e-16 ***
## educ             0.0460597  0.0031864  14.455  < 2e-16 ***
## polviews         0.0195194  0.0060545   3.224  0.00127 **
## class            0.1077572  0.0146582   7.351 2.18e-13 ***
## satfin           0.1244617  0.0134911   9.225  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7469 on 7266 degrees of freedom
## Multiple R-squared:  0.1938, Adjusted R-squared:  0.1929
## F-statistic: 218.3 on 8 and 7266 DF,  p-value: < 2.2e-16
```

Interpretation of the two regressions

We see similar qualitative results for the effect of age, self-rated happiness, life satisfaction, higher education, and higher financial satisfaction on SRH, all of which make intuitive sense, as discussed above. However, in more recent years (2010s), sex no longer has a significant association with SRH, and political views (being more conservative) as well as higher self-rated class, have a statistically significant association with SRH, while in the 1980s, they did not.

Formally compare the 1980s with the 2010s

First let's look at the coefficients in each time period

Coefficients on SRH prediction for 1980s

```
lm(health ~ age + as.factor(sex) + happy + life + educ + polviews + class + satfin,
   data = (data_gss %>% filter(year >= 1980 & year <= 1989))) %>%
  tidy(conf.int = TRUE) %>%
  knitr::kable()
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.7861187	0.0609334	29.3126624	0.0000000	1.6666760	1.9055615
age	-0.0120483	0.0004893	-24.6230482	0.0000000	-0.0130074	-0.0110891
as.factor(sex)2	-0.0402847	0.0159712	-2.5223323	0.0116745	-0.0715917	-0.0089776
happy	0.2216346	0.0138326	16.0226331	0.0000000	0.1945197	0.2487496
life	0.1702987	0.0149377	11.4006003	0.0000000	0.1410175	0.1995798
educ	0.0491659	0.0028809	17.0663005	0.0000000	0.0435187	0.0548130
polviews	0.0033600	0.0059686	0.5629352	0.5734926	-0.0083399	0.0150598
class	0.0218612	0.0136300	1.6039080	0.1087683	-0.0048565	0.0485790
satfin	0.1149901	0.0116907	9.8360328	0.0000000	0.0920738	0.1379064

Coefficients on SRH prediction for 2010s

```
lm(health ~ age + as.factor(sex) + happy + life + educ + polviews + class + satfin,
   data = (data_gss %>% filter(year >= 2010 & year <= 2019))) %>%
  tidy(conf.int = TRUE) %>%
  knitr::kable()
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.0495256	0.0665043	15.781315	0.0000000	0.9191578	1.1798934
age	-0.0070728	0.0005152	-13.728383	0.0000000	-0.0080827	-0.0060629
as.factor(sex)2	0.0287843	0.0176578	1.630117	0.1031201	-0.0058301	0.0633987
happy	0.2183796	0.0147618	14.793610	0.0000000	0.1894423	0.2473169
life	0.2046324	0.0161011	12.709182	0.0000000	0.1730695	0.2361953
educ	0.0460597	0.0031864	14.455098	0.0000000	0.0398134	0.0523060
polviews	0.0195194	0.0060545	3.223937	0.0012700	0.0076508	0.0313880
class	0.1077572	0.0146582	7.351333	0.0000000	0.0790229	0.1364915
satfin	0.1244617	0.0134911	9.225443	0.0000000	0.0980152	0.1509083

Coefficient comparison for the 1980s vs 2010s

When we compare the coefficients, we see that they all remain in the same direction, except for sex. However, the p-value for sex, though significant for the 1980s, is not significant for the 2010s (if we use a threshold of $p < 0.05$), so we should use caution in interpreting that as a change from females having higher SRH to males having higher SRH.

As noted before, age, education, happiness, life satisfaction, and financial satisfaction are always statistically significantly associated with SRH. However, self-rated class and political views are not associated with SRH in the 1980s, but they are associated in the 2010s (higher class and more conservative political views are associated with higher SRH).

```
estimates_80s <- lm(health ~ age + as.factor(sex) + happy + life + educ + polviews + class + satfin,
   data = (data_gss %>% filter(year >= 1980 & year <= 1989))) %>%
  tidy(conf.int = TRUE) %>%
  select(term,
         coef_80s = estimate,
         se_80s = std.error,
         p_80s = p.value
         )

estimates_10s <- lm(health ~ age + as.factor(sex) + happy + life + educ + polviews + class + satfin,
   data = (data_gss %>% filter(year >= 2010 & year <= 2019))) %>%
  tidy(conf.int = TRUE) %>%
  select(term,
         coef_10s = estimate,
         se_10s = std.error,
         p_10s = p.value
         )

compare_80s_10s <- merge(estimates_80s, estimates_10s, by = "term") %>%
  select(term,
         coef_80s, coef_10s,
         se_80s, se_10s,
         p_80s, p_10s)

knitr::kable(compare_80s_10s)
```

term	coef_80s	coef_10s	se_80s	se_10s	p_80s	p_10s
(Intercept)	1.7861187	1.0495256	0.0609334	0.0665043	0.0000000	0.0000000
age	-0.0120483	-0.0070728	0.0004893	0.0005152	0.0000000	0.0000000
as.factor(sex)2	-0.0402847	0.0287843	0.0159712	0.0176578	0.0116745	0.1031201

term	coef_80s	coef_10s	se_80s	se_10s	p_80s	p_10s
class	0.0218612	0.1077572	0.0136300	0.0146582	0.1087683	0.0000000
educ	0.0491659	0.0460597	0.0028809	0.0031864	0.0000000	0.0000000
happy	0.2216346	0.2183796	0.0138326	0.0147618	0.0000000	0.0000000
life	0.1702987	0.2046324	0.0149377	0.0161011	0.0000000	0.0000000
polviews	0.0033600	0.0195194	0.0059686	0.0060545	0.5734926	0.0012700
satfin	0.1149901	0.1244617	0.0116907	0.0134911	0.0000000	0.0000000

Z-test

Now let's use a Z-test to compare the coefficients.

```
z_80s_10s <- compare_80s_10s %>%
  mutate(b1minusb2 = coef_80s - coef_10s,
         denom = sqrt( (se_80s^2) + (se_10s^2) )) %>%
  mutate(z = b1minusb2 / denom) %>%
  mutate(p_value = 2 * (1 - pnorm(abs(z)))) %>%
  select(term, coef_80s, coef_10s, z, p_value)

knitr::kable(z_80s_10s)
```

term	coef_80s	coef_10s	z	p_value
(Intercept)	1.7861187	1.0495256	8.1663922	0.0000000
age	-0.0120483	-0.0070728	-7.0024728	0.0000000
as.factor(sex)2	-0.0402847	0.0287843	-2.9009380	0.0037205
class	0.0218612	0.1077572	-4.2913735	0.0000178
educ	0.0491659	0.0460597	0.7230989	0.4696191
happy	0.2216346	0.2183796	0.1609027	0.8721701
life	0.1702987	0.2046324	-1.5632392	0.1179963
polviews	0.0033600	0.0195194	-1.9006874	0.0573430
satfin	0.1149901	0.1244617	-0.5305740	0.5957140

Interpretation of Results

We can see that there is an (extremely) statistically significant difference in the coefficient of age on SRH. Although it is in the same direction (older people have lower SRH) in both decades, the magnitude is markedly lower in the 2010s compared to the 1980s. We will explore this further later in the document.

The p-values for the z-tests for education, happiness, life satisfaction, and financial satisfaction are not significantly different for the 1980s vs the 2010s. All of these variables are significantly, positively associated with SRH in both decades (higher education, happiness, life satisfaction, and financial satisfaction are associated with higher self-rated health). The non-significant z-test means that they tend to have the same weight in the prediction of SRH in both time periods.

The p-value for the z-test for class ($p = 0.0000178$) is significant, which means that in the 2010s, people's self-rated class impacts their self-rated health (higher self-rated class is associated with higher SRH), while in the 1980s, it did not (no significant association seen, although the coefficient was also positive). This is interesting and could reflect the changing class dynamics that have occurred over the past decades.

The p-value for the z-test for political views ($p = 0.0573$) is somewhat significant (the threshold people often use of $p < 0.05$ is somewhat arbitrary and interpretation is usually context dependent). This suggests that it is possible that although political views in the 1980s had no effect on SRH, in the 2010s, people with more conservative views had higher SRH.

We also see a significant z-value for the predictive power of sex on SRH, as well as a reverse in direction (females have higher SRH in the 1980s and males have higher SRH in the 2010s), however, recall from earlier that the the p-value for sex, though significant for the 1980s, is not significant for the 2010s (if we use a threshold of $p < 0.05$), so we should use caution in interpreting that as a change from females having higher SRH to males having higher SRH, although there may well be an effect there, it deserves further exploration.

Bonus - Self rated health as predicted by age over the years

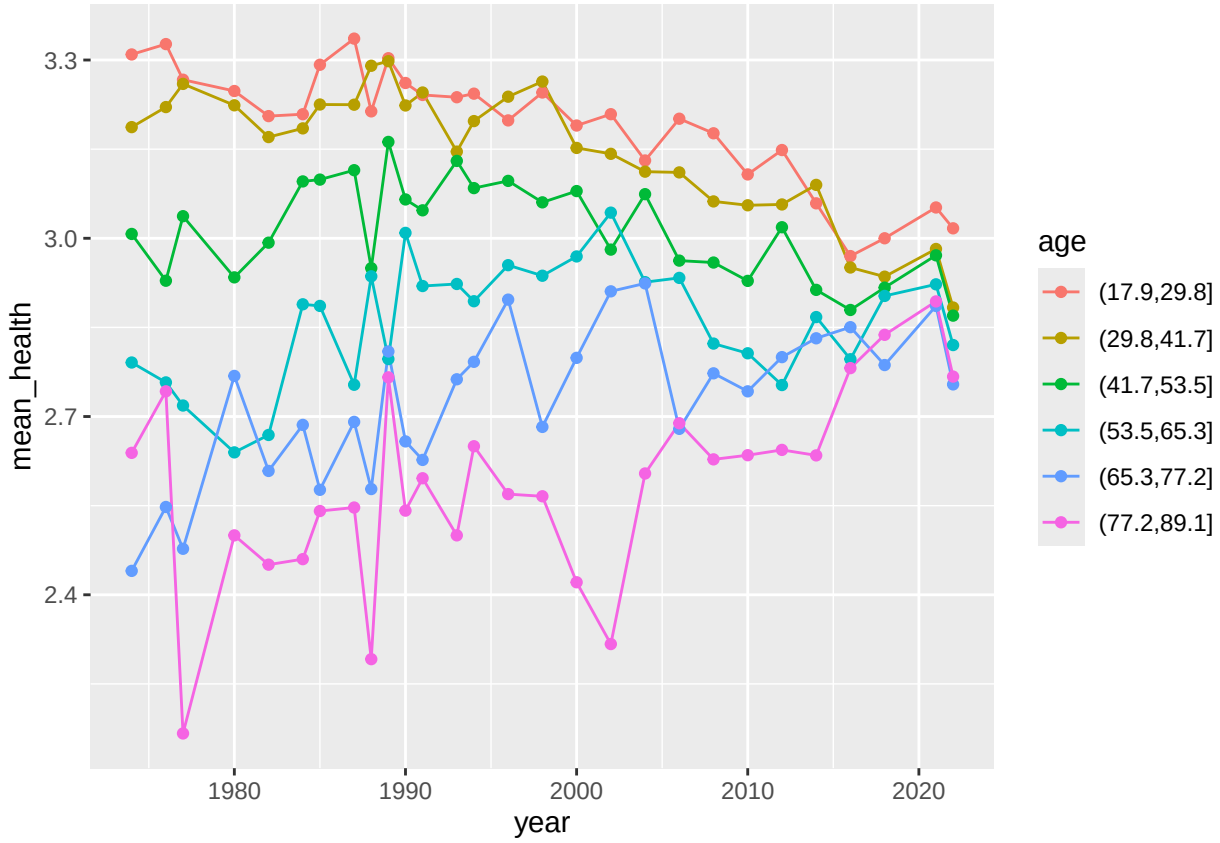
Let's explore the effect of different cohorts on SRH at certain ages.

SRH vs year of survey for different ages

In the following figure, I cut the age of participants into 6 groups and plotted the mean of the group's self-rated health for each year (that's what each dot is). As you can qualitatively see, the spread seems to narrow.

```
data_gss %>%
  mutate(age = cut(age, breaks = 6)) %>% # Create cohorts with 6 breaks
  group_by(age, year) %>%
  summarize(mean_health = mean(health)) %>%
  ggplot(aes(x = year, y = mean_health, color = age)) +
  geom_line() +
  geom_point()
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```



This qualitative result is robust to the size of the categorical variables I split "age" into.

```r
par(mfrow = c(2, 2))

p1 <- data_gss %>%
  mutate(age = cut(age, breaks = 10)) %>% # Create cohorts with 6 breaks
  group_by(age, year) %>%
  summarize(mean_health = mean(health)) %>%
  ggplot(aes(x = year, y = mean_health, color = age)) +
  labs(title = "SRH for Different Ages over the Years") +
  geom_line()
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

```r
p2 <- data_gss %>%
  mutate(age = cut(age, breaks = 7)) %>% # Create cohorts with 6 breaks
  group_by(age, year) %>%
  summarize(mean_health = mean(health)) %>%
  ggplot(aes(x = year, y = mean_health, color = age)) +
  labs(title = "SRH for Different Ages over the Years") +
  geom_line()
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

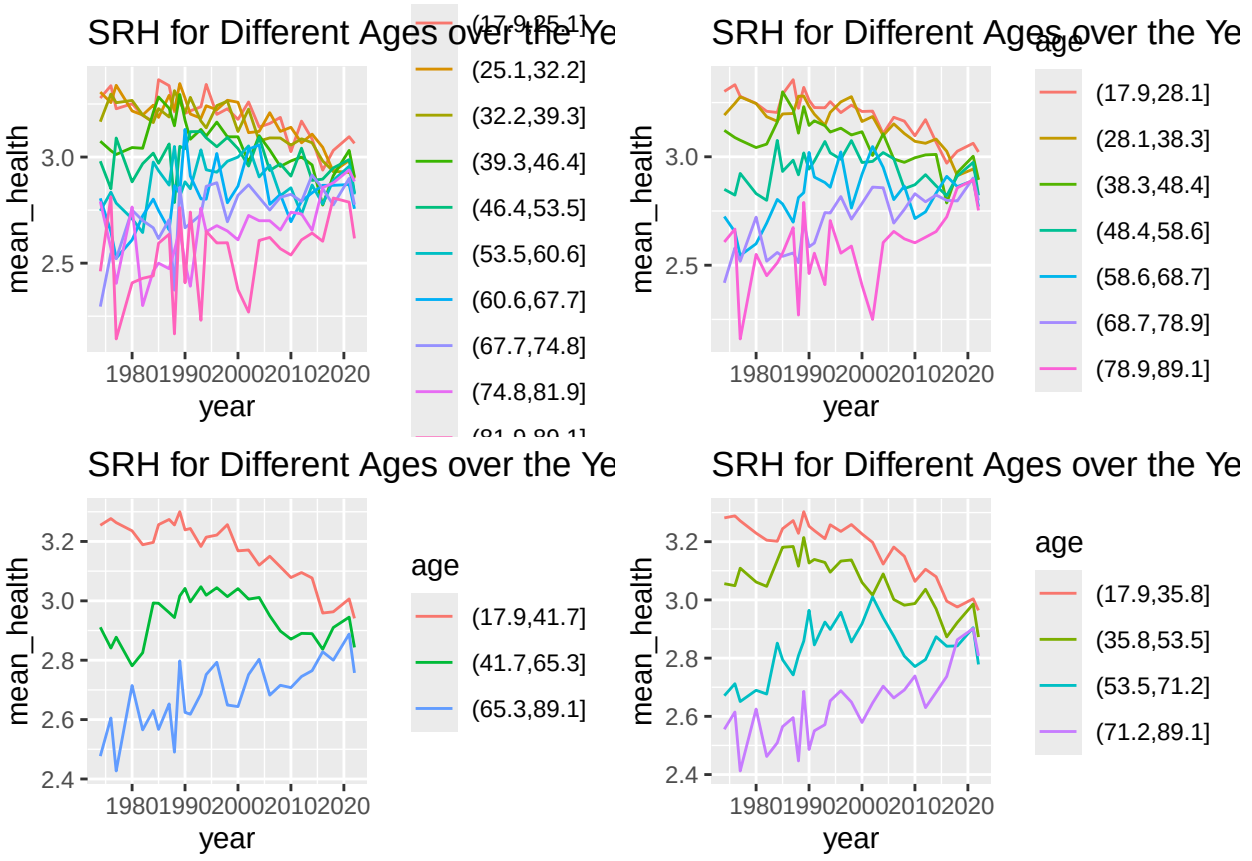
```r
p3 <- data_gss %>%
  mutate(age = cut(age, breaks = 3)) %>% # Create cohorts with 6 breaks
  group_by(age, year) %>%
  summarize(mean_health = mean(health)) %>%
  ggplot(aes(x = year, y = mean_health, color = age)) +
  labs(title = "SRH for Different Ages over the Years") +
  geom_line()
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

```r
p4 <- data_gss %>%
  mutate(age = cut(age, breaks = 4)) %>% # Create cohorts with 6 breaks
  group_by(age, year) %>%
  summarize(mean_health = mean(health)) %>%
  ggplot(aes(x = year, y = mean_health, color = age)) +
  labs(title = "SRH for Different Ages over the Years") +
  geom_line()
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

```r
gridExtra::grid.arrange(p1, p2, p3, p4, nrow=2)
```



Relationship of self-rated health to age, separated out by years

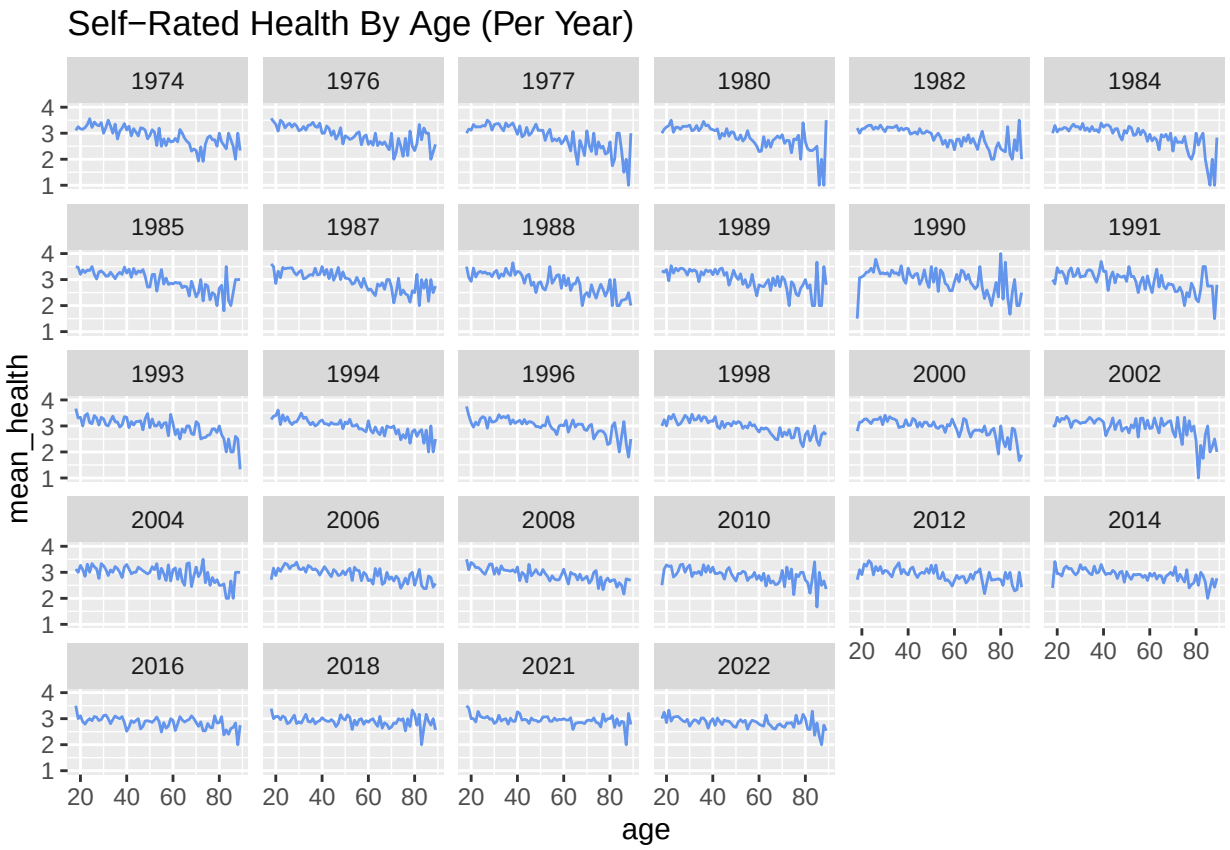
Well, it seems like the spread of self-rated health among ages decreases as time goes on (later years). Let's look at that by faceting mean self-rated health vs age by year.

Note that intuitively, we'd expect it to be a negative slope because older people intuitively should have worse health.

Notice *the slopes seem to flatten over time.*

```
# health vs age per year
data_gss %>%
  group_by(age, year) %>%
  summarize(mean_health = mean(health)) %>%
  ggplot(aes(x = age, y = mean_health)) +
  geom_line(color = "cornflowerblue") +
  facet_wrap(~ year) +
  labs(title = "Self-Rated Health By Age (Per Year)" )
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```



Regress self-rated health on age, for each year

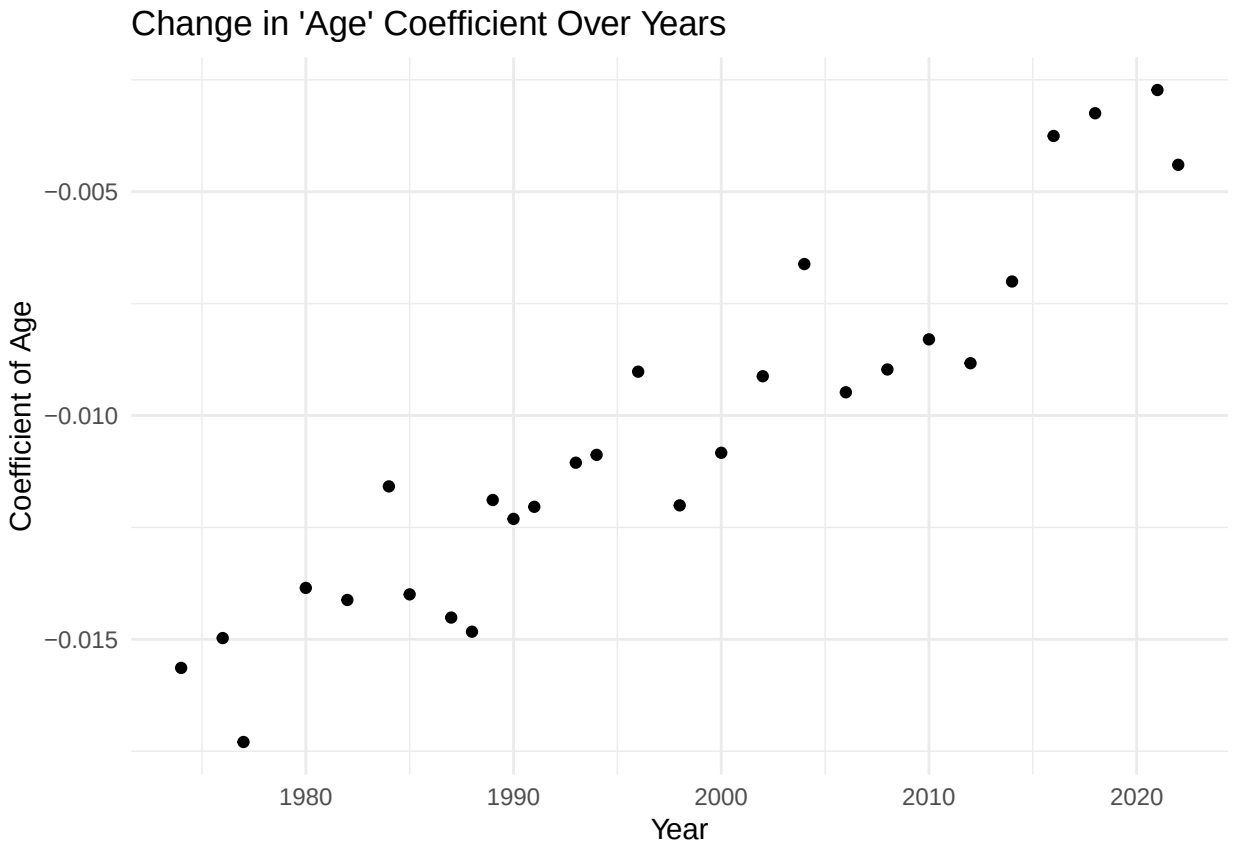
Let's do a simple linear regression on each self-rated-health vs age, subsetted for each year (the plots on the faceted figure), look at the significance, and plot the coefficients for age with 95% CIs:

```
library(broom)

# Aggregate slopes

# years_of_gss <- c(data_gss %>% select(year) %>% unique() )
# lm_health_v_age_0 <- data_gss %>%
#   group_by(year) %>%
#   summarize(coef = coef(lm(health ~ age, data = cur_data()))["age"])

# Perform linear regression for each year and extract the coefficient of 'age' with confidence interval
lm_health_v_age_0 <- data_gss %>%
  group_by(year) %>%
  do(tidy(lm(health ~ age, data = .), conf.int = TRUE)) %>%  # Add conf.int = TRUE for CIs
  filter(term == "age") %>%
  select(year, coef = estimate, conf.low, conf.high, se = std.error, t_statistic = statistic,  p_value =

# View the results with confidence intervals, se, t statistic, and p value
# print(lm_health_v_age_0)
knitr::kable(lm_health_v_age_0)
```

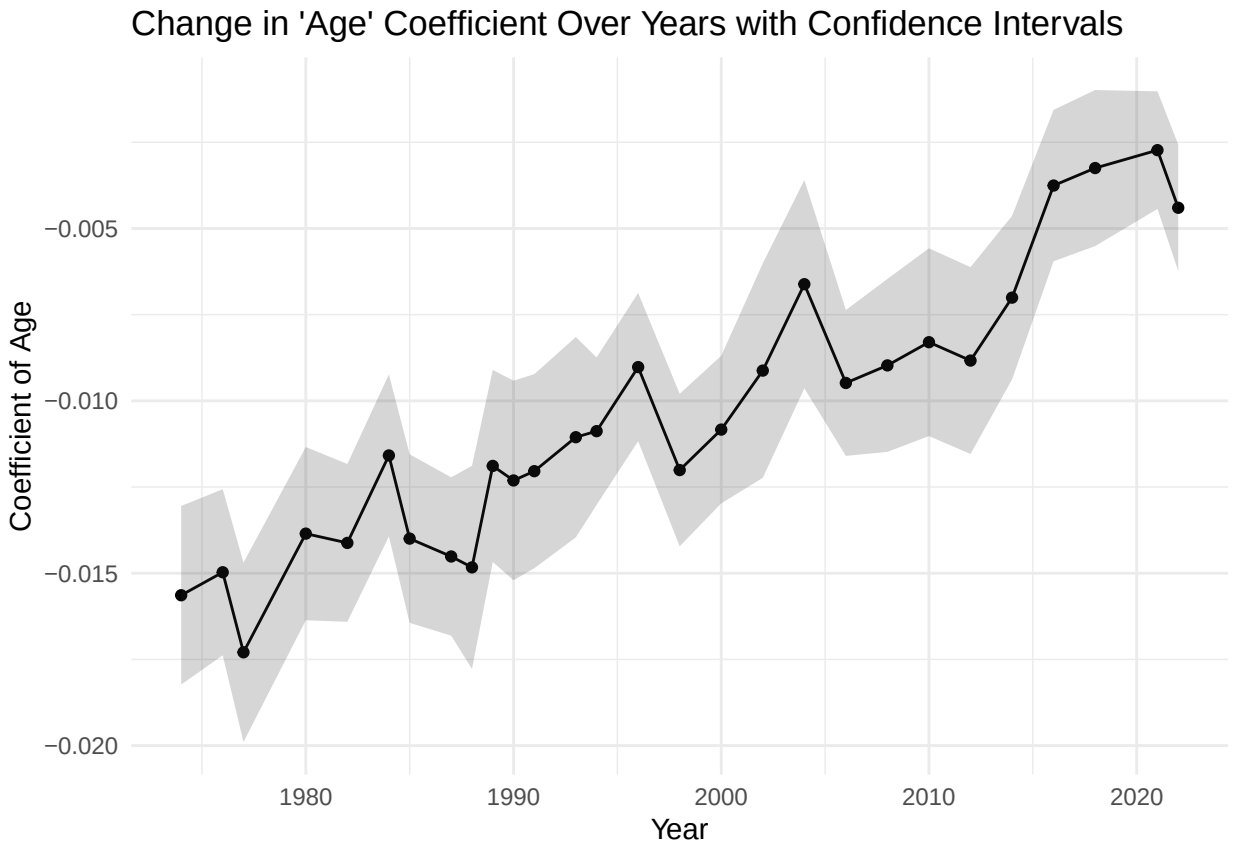
year	coef	conf.low	conf.high	se	t_statistic	p_value
1974	-0.0156387	-0.0182284	-0.0130490	0.0013201	-11.846458	0.0000000
1976	-0.0149720	-0.0173804	-0.0125636	0.0012277	-12.195023	0.0000000
1977	-0.0172938	-0.0198951	-0.0146924	0.0013261	-13.040929	0.0000000
1980	-0.0138517	-0.0163643	-0.0113391	0.0012808	-10.814492	0.0000000
1982	-0.0141204	-0.0164095	-0.0118313	0.0011671	-12.098657	0.0000000
1984	-0.0115835	-0.0139332	-0.0092339	0.0011978	-9.671043	0.0000000
1985	-0.0139956	-0.0164398	-0.0115513	0.0012460	-11.232164	0.0000000
1987	-0.0145142	-0.0168100	-0.0122183	0.0011705	-12.400363	0.0000000
1988	-0.0148299	-0.0177768	-0.0118831	0.0015015	-9.876757	0.0000000
1989	-0.0118876	-0.0146720	-0.0091032	0.0014188	-8.378607	0.0000000
1990	-0.0123095	-0.0152031	-0.0094159	0.0014743	-8.349448	0.0000000
1991	-0.0120397	-0.0148544	-0.0092249	0.0014343	-8.394284	0.0000000
1993	-0.0110543	-0.0139591	-0.0081495	0.0014803	-7.467704	0.0000000
1994	-0.0108790	-0.0130172	-0.0087408	0.0010902	-9.978617	0.0000000
1996	-0.0090209	-0.0111704	-0.0068713	0.0010960	-8.230802	0.0000000
1998	-0.0120072	-0.0142185	-0.0097960	0.0011274	-10.650182	0.0000000
2000	-0.0108334	-0.0129743	-0.0086924	0.0010916	-9.924665	0.0000000
2002	-0.0091219	-0.0122330	-0.0060108	0.0015851	-5.754717	0.0000000
2004	-0.0066162	-0.0096374	-0.0035949	0.0015393	-4.298090	0.0000192
2006	-0.0094810	-0.0115968	-0.0073651	0.0010788	-8.788270	0.0000000
2008	-0.0089714	-0.0114807	-0.0064621	0.0012790	-7.014283	0.0000000
2010	-0.0082983	-0.0110213	-0.0055752	0.0013879	-5.978846	0.0000000
2012	-0.0088312	-0.0115414	-0.0061210	0.0013814	-6.392935	0.0000000
2014	-0.0070062	-0.0093794	-0.0046331	0.0012099	-5.790762	0.0000000
2016	-0.0037546	-0.0059502	-0.0015589	0.0011195	-3.353787	0.0008140
2018	-0.0032474	-0.0055114	-0.0009834	0.0011542	-2.813601	0.0049642
2021	-0.0027279	-0.0044336	-0.0010222	0.0008698	-3.136150	0.0017326
2022	-0.0043991	-0.0062322	-0.0025660	0.0009347	-4.706242	0.0000027

Note that every single beta is statistically significant. Now let's visualize it.

```
# Plot coefficients
ggplot(lm_health_v_age_0, aes(x = year, y = coef)) +
  geom_point() +
  labs(
    title = "Change in 'Age' Coefficient Over Years",
    x = "Year",
    y = "Coefficient of Age"
  ) +
  theme_minimal()
```




```
# Plot coefficients with CI
ggplot(lm_health_v_age_0, aes(x = year, y = coef)) +
  geom_line() +
  geom_point() +
  geom_ribbon(aes(ymin = conf.low, ymax = conf.high), alpha = 0.2) +  # Add shaded area for confidence
  labs(
    title = "Change in 'Age' Coefficient Over Years with Confidence Intervals",
    x = "Year",
    y = "Coefficient of Age"
  ) +
  theme_minimal()
```



Regress the srh vs age coefficients from each year on the year of the survey

The relationship looks surprisingly strong and linear, so let's do another regression of the coefficients on year. It is super statistically significant (which I'm not sure totally how to interpret since it's on coefficients):

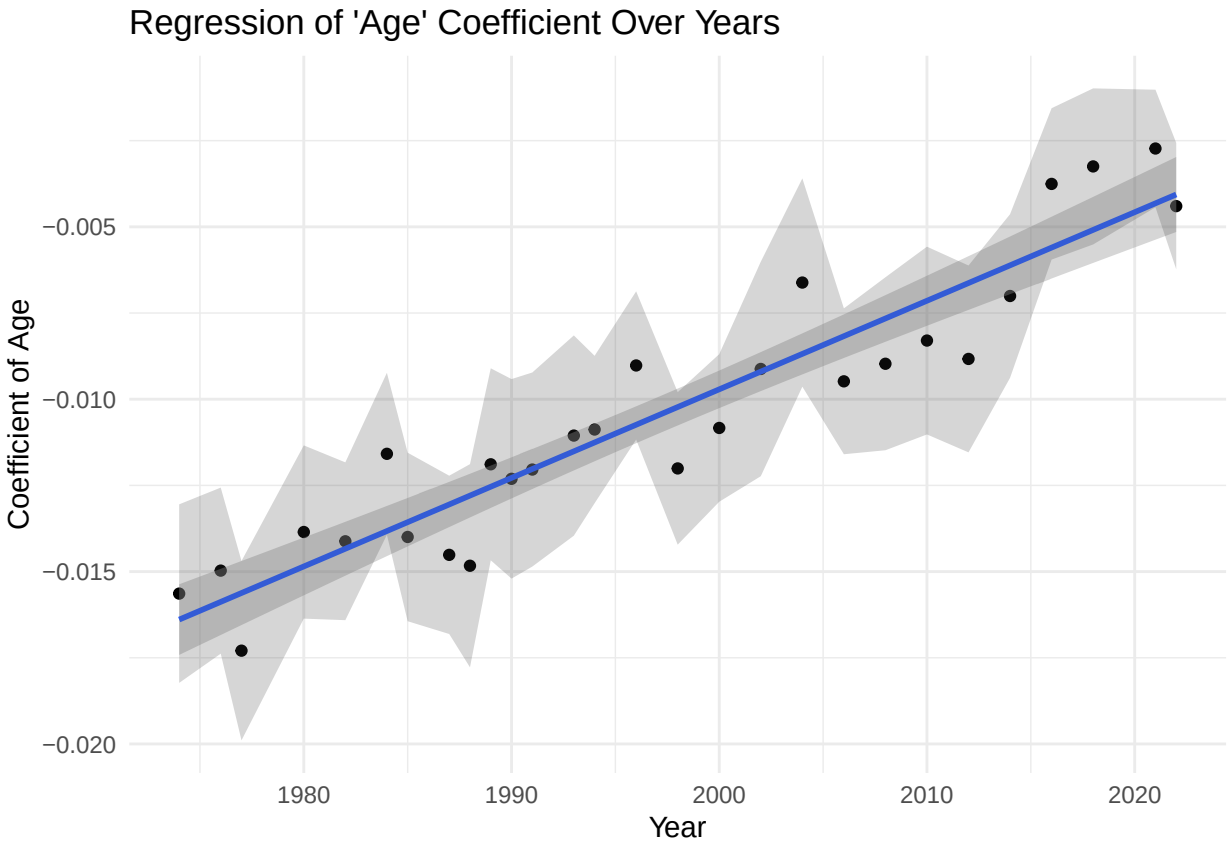
```
# Perform linear regression of 'coef' (age coefficient) vs 'year'
lm_coef_vs_year <- lm(coef ~ year, data = lm_health_v_age_0)

# View the summary of the regression
summary(lm_coef_vs_year)
```

```
##
## Call:
## lm(formula = coef ~ year, data = lm_health_v_age_0)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.201e-03 -1.193e-03  3.136e-05  9.298e-04  2.240e-03
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.235e-01  3.689e-02  -14.19 9.34e-14 ***
## year         2.569e-04  1.847e-05   13.91 1.49e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001373 on 26 degrees of freedom
## Multiple R-squared:  0.8815, Adjusted R-squared:  0.877
```

```
## F-statistic: 193.5 on 1 and 26 DF,  p-value: 1.489e-13
```

```
ggplot(lm_health_v_age_0, aes(x = year, y = coef)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE) +  # Adds the regression line with standard error shading
  geom_ribbon(aes(ymin = conf.low, ymax = conf.high), alpha = 0.2) +  # Confidence intervals for the co
  labs(
    title = "Regression of 'Age' Coefficient Over Years",
    x = "Year",
    y = "Coefficient of Age"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



So basically this shows that as years pass, the predictive power of someone's age on their self-rated health decreases. Interesting!