Problem Set 1: Personal Loss, Individual Risk Factors, and Concern about COVID-19

Harvard University
Spring 2023
Instructor: Gregory Bruich, Ph.D.
Student: Benjamin Zeisberg

- Posted on: 01/24/2023
- Due at: 11:59pm on 01/31/2023

Suggested Imports

In [56]:
```
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import statsmodels.api as sm

# Uncomment the line below if using Google Colab:
# !pip install stargazer
from stargazer.stargazer import Stargazer
from sklearn.linear_model import LinearRegression
```

Background

The Health and Retirement Study is a survey designed to track households consisting of older Americans. The COVID-19 module of HRS 2020 was administered to 3,210 respondents in June. These data were released on November 13, 2020. Here is a link to the questionnaire.

With over 5.6 million deaths world wide from COVID-19, many of us have had family members or close friends die during the pandemic. In this assignment, we will quantify how this personal loss ("Has anyone you know died from COVID-19?") as well as individual risk factors (e.g., age, diabetes) affect respondents' answer to the following question:

> *Overall, on a scale from 1 to 10, where one is the least concerned and ten is the most concerned, how concerned are you about the coronavirus pandemic?*

In the HRS data, 42.1% of respondents put 10/10 for this question. 19.8% of respondents know someone who has died from COVID-19.

The goal of this assignment is twofold. First, this is an opportunity to get to know other students in the class by working together in study groups. Second, the assignment is designed to help you acquire some familiarity working with data in statistical software. You are welcome to use any software you would like.

You are encouraged to work in groups on your problem sets, and we will facilitate the formation of study groups. However, each student must write up his or her answers individually in his or her own words based on his or her own understanding.

Data Description

File: `covid19.dta`

The data consist of $n = 3210$ individuals surveyed as part of the Health and Retirement Survey.

Variable	Description	Mean	Std. Dev.
`anyone_died`	Has anyone you know died from COVID-19? 1 if yes, 0 if no	0.198	0.339
`concern`	How concerned are you about the coronavirus pandemic? Scale of 1 to 10	7.780	2.659
`college`	1 if respondent has ≥ 16 years of education, 0 otherwise	0.288	0.453
`diabetes`	1 if respondent has diabetes, 0 otherwise	0.265	0.441
`age65`	1 if respondent is 65 or older on March 2020, 0 if younger than 65 in March 2020	0.565	0.496

NOTE — Table reports means and standard deviations for select variables from the HRS 2020.

Data Load

In [57]:
```
# Read dataset into a pandas dataframe
covid = pd.read_stata("covid19.dta")

# Add intercept to dataframe
covid["intercept"] = 1

# Display first 5 rows of data
covid.head()
```

Out[57]:

	hhid_pn	concern	anyone_died	college	diabetes	age65	covid	wanted_test	female
0	010325020	5.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0
1	010372010	10.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0
2	010397010	8.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0
3	010577010	1.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0
4	010773020	10.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0

Instructions

Please submit your Problem Set on Canvas. Your submission should include two files:

1. This notebook as a `.ipynb` file with your code and answers to questions.
2. A `.pdf` version of this notebook. Different editors will have different ways of converting notebooks to PDFs, but feel free to message in the Slack if you have any problems with this.

Please remember to restart and run your entire notebook before submitting!

Questions

Note: Short answers should be very succinct. Show your work and intuition clearly: credit is given for explanations and not just having the correct answer

Sample code is provided at the end of this notebook.

1

Join our Slack workspace (Instructions on Canvas). Respond to the following prompt on the `#welcome` channel:

> *If you were not at Harvard College right now, what would you be doing instead? That is, what is your "unobserved counterfactual"?*

Please feel free to react (kindly and respectfully, please) to the posts of others.

(No Answer Necessary)

2

Using four separate *linear regressions*, estimate the difference in mean level of concern about COVID-19 for the following groups:

1. Individuals who know someone who died of COVID-19 vs. those who do not
2. Individuals with diabetes vs. those without diabetes
3. Individuals over 65 vs. those under 65
4. Individuals with a college education vs. those less than a college education

Report appropriate standard errors for these differences in means, and discuss how you chose between three possibilities: Pooled variance formula, HC_1 unequal variance formula, and HC_2 unequal variance formula. Sample code is provided in Table 2a and Table 2b.

In [58]:
```
covid.head()
```

Out[58]:

	hhid_pn	concern	anyone_died	college	diabetes	age65	covid	wanted_test	female
0	010325020	5.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0
1	010372010	10.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0
2	010397010	8.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0
3	010577010	1.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0
4	010773020	10.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0

1. Individuals who know someone who died of COVID-19 vs. those who do not

We use a linear regression of the form $\bar{Y} = \beta_0 + \beta_1 \cdot X_1 + \epsilon$ to estimate the difference in mean level of concern about Covid-19.

We estimate that the average difference in concern about covid-19 of those who know somone that died of covid-19 vs those that do not know someone that died of covid is $\beta_1 = 0.7833$. We find this estimate statistically significant at the $\alpha = 0.05$ level. Standard errors were calculated using the HC2 variance formula, since it can resolve unequal variance(heterskedasticity) unlike the pooled-variance formula. I choose the HC_2 formula throughout this problem set, since it's been shown(in lecture) to provide more consistent standard error results then the HC_1 formula.

In [59]:
```
# 1. Individuals who know someone who died of COVID-19 vs those who do not
# Response variable -> difference in mean level of concern.
Y = covid[['concern']]
X = covid[['anyone_died', 'intercept']]
mod_1 = sm.OLS(Y,X)
res_1 = mod_1.fit(cov_type='HC2')
print(res_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                concern   R-squared:                       0.
014
Model:                            OLS   Adj. R-squared:                  0.
013
Method:                 Least Squares   F-statistic:                     5
0.74
Date:                Tue, 31 Jan 2023   Prob (F-statistic):           1.29e
-12
Time:                        21:12:17   Log-Likelihood:                -766
5.6
No. Observations:                3210   AIC:                          1.534e
+04
Df Residuals:                    3208   BIC:                          1.535e
+04
Df Model:                           1
Covariance Type:                  HC2
==============================================================================
====
                 coef    std err          z      P>|z|      [0.025      0.
975]
------------------------------------------------------------------------------
----
anyone_died     0.7833      0.110      7.124      0.000       0.568
0.999
intercept       7.6290      0.053    144.206      0.000       7.525
7.733
==============================================================================
===
Omnibus:                      485.238   Durbin-Watson:                   1.
782
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              730.
044
Skew:                          -1.150   Prob(JB):                     2.97e-
159
Kurtosis:                       3.406   Cond. No.
2.63
==============================================================================
===

Notes:
[1] Standard Errors are heteroscedasticity robust (HC2)
```

2. Individuals with diabetes vs. those without diabetes

I use a linear regression of the form $\bar{Y} = \beta_0 + \beta_1 \cdot X_1 + \epsilon$ to estimate the difference in mean level of concern about Covid-19.

I estimate that the average difference in concern about covid-19 for those with diabetes vs those without diabetes is $\beta_1 = 0.2514$. I find this estimate to be statistically signifiant at the $\alpha = 0.05$ level. Standard errors were calculated using the HC2 variance formula, as mentioned before.

In [60]:
```
# # 2. Individuals with diabetes s. those without diabetes
# technically we don't need to define Y again, but to make the code more rea
Y = covid[['concern']]
X = covid[['diabetes', 'intercept']]
mod_2 = sm.OLS(Y,X)
res_2 = mod_2.fit(cov_type='HC2')
print(res_2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                concern   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     5.723
Date:                Tue, 31 Jan 2023   Prob (F-statistic):             0.0168
Time:                        21:12:17   Log-Likelihood:                -7685.0
No. Observations:                3210   AIC:                         1.537e+04
Df Residuals:                    3208   BIC:                         1.539e+04
Df Model:                           1
Covariance Type:                  HC2
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
diabetes       0.2514      0.105      2.392      0.017       0.045       0.457
intercept      7.7168      0.055    140.598      0.000       7.609       7.824
==============================================================================
Omnibus:                      482.959   Durbin-Watson:                   1.772
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              726.762
Skew:                          -1.152   Prob(JB):                    1.53e-158
Kurtosis:                       3.358   Cond. No.                         2.46
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC2)
```

3. Individuals with aged 65 vs. under 65

We use a linear regression of the form $\bar{Y} = \beta_0 + \beta_1 \cdot X_1 + \epsilon$ to estimate the difference in mean level of concern about Covid-19.

I estimate that the average difference in concern about covid-19 for those aged above 65 vs those under age 65 to be $\beta_1 = 0.2514$. I find this estimate to not be statistically signifiant at the $\alpha = 0.05$ level. Standard errors were calculated using the HC2 variance formula, as mentioned before.

In [61]:
```
# 3. Individuals over 65 vs those under 65
Y = covid[['concern']]
X = covid[['age65', 'intercept']]
mod_3 = sm.OLS(Y,X)
res_3 = mod_3.fit(cov_type='HC2')
print(res_3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                concern   R-squared:                       0.
001
Model:                            OLS   Adj. R-squared:                  0.
001
Method:                 Least Squares   F-statistic:                     3.
609
Date:                Tue, 31 Jan 2023   Prob (F-statistic):             0.0
575
Time:                        21:12:17   Log-Likelihood:                -768
6.0
No. Observations:                3210   AIC:                         1.538e
+04
Df Residuals:                    3208   BIC:                         1.539e
+04
Df Model:                           1

Covariance Type:                  HC2

==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.9
75]
------------------------------------------------------------------------------
age65          0.1800      0.095      1.900      0.057      -0.006       0.
366
intercept      7.6817      0.072    106.830      0.000       7.541       7.
823
==============================================================================
Omnibus:                      482.054   Durbin-Watson:                   1.
773
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              724.
972
Skew:                          -1.151   Prob(JB):                     3.75e-
158
Kurtosis:                       3.354   Cond. No.
2.80
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC2)
```

4. Individuals with college degree vs. those without college degree

I use a linear regression of the form $\bar{Y} = \beta_0 + \beta_1 \cdot X_1 + \epsilon$ to estimate the difference in mean level of concern about Covid-19.

I estimate the average difference in concern about covid-19 for those with a college degree vs those without to be $\beta_1 = -0.1375$, with a $P > |z| == 0.166$. Suggesting, that the coefficient is not statistically significant at the $\alpha = 0.05$ level. Standard errors were calculated using the HC_2 variance formula, as mentioned before.

In [62]:
```
# 4. Individuals with a college education vs. those less than a college educ
X = covid[['college', 'intercept']]
mod_4 = sm.OLS(Y,X)
res_4 = mod_4.fit(cov_type='HC2')
print(res_4.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                concern   R-squared:                       0.
001
Model:                            OLS   Adj. R-squared:                  0.
000
Method:                 Least Squares   F-statistic:                     1.
923
Date:                Tue, 31 Jan 2023   Prob (F-statistic):              0.
166
Time:                        21:12:17   Log-Likelihood:                -768
7.0
No. Observations:                3210   AIC:                         1.538e
+04
Df Residuals:                    3208   BIC:                         1.539e
+04
Df Model:                           1
Covariance Type:                  HC2
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.9
75]
------------------------------------------------------------------------------
college       -0.1375      0.099     -1.387      0.166      -0.332       0.
057
intercept      7.8235      0.057    137.086      0.000       7.712       7.
935
==============================================================================
Omnibus:                      486.904   Durbin-Watson:                   1.
768
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              734.
835
Skew:                          -1.158   Prob(JB):                     2.71e-
160
Kurtosis:                       3.366   Cond. No.
2.43
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC2)
```

In [63]:

```
# We can also use the Sklearn Library to do an OLS regression, but I don't t
x = covid[['anyone_died']]
reg = LinearRegression(fit_intercept=True).fit(x,Y)
parameters = reg.get_params()
# This is what we would use to print out the coef and parameters
# print(reg.coef_)
# print(parameters)
```

3

Later this semester, it will save you a lot of time and effort to generate tables of regression output automatically. Generating tables is easy to do using stargazer in Python. Use this library to create a table of the four regressions from the first question, with each regression listed in a separate column.

In [64]:
```
# Your Code Here
models = [res_1, res_2, res_3, res_4]
table = Stargazer(models)
table.covariate_order(['intercept','anyone_died', 'diabetes', 'age65', 'coll
table.custom_columns(['OLS Model', 'OLS Model', 'OLS Model', 'OLS Model'], [
table.add_custom_notes(["The Standard errors reported in parentheses are het
table
```

Out[64]:

	Dependent variable:concern			
	OLS Model	OLS Model	OLS Model	OLS Model
	(1)	(2)	(3)	(4)
intercept	7.629***	7.717***	7.682***	7.823***
	(0.053)	(0.055)	(0.072)	(0.057)
anyone_died	0.783***			
	(0.110)			
diabetes		0.251**		
		(0.105)		
age65			0.180*	
			(0.095)	
college				-0.137
				(0.099)
Observations	3,210	3,210	3,210	3,210
R^2	0.014	0.002	0.001	0.001
Adjusted R^2	0.013	0.001	0.001	0.000
Residual Std. Error	2.636 (df=3208)	2.652 (df=3208)	2.653 (df=3208)	2.654 (df=3208)
F Statistic	50.745*** (df=1; 3208)	5.723** (df=1; 3208)	3.609* (df=1; 3208)	1.923 (df=1; 3208)
Note:				*p<0.1; **p<0.05; ***p<0.01
				The Standard errors reported in parentheses are heteroskedasticity robust (HC2).

4

Calculate 95% confidence intervals for the differences in means you estimated above. What critical values did you use? Discuss briefly.

I shall use the critical value 1.96. Since we know from lecture for a 95% ($\alpha = 0.05$) confidence interval we use: $\bar{Y} \pm 1.96 \cdot \hat{SE}(\bar{Y})$
If I was to calculate a 99% confidence interval, which is $\alpha = 0.01$ I would look for $\alpha/2 = 0.005$ in the normal distribution, which corresponds to a Z critical value of 2.576.

In [65]:
```
# Your Code Here
# We shall use 1.96 as our critical value.
# We can make use of conf_int function from statsmodels for documentation se
# In conf_int we specify the confidence interval using the alpha parameter.
interval_1 = res_1.conf_int(alpha=0.05)
interval_1.rename(columns={0:'Lower_bound', 1:'Upper bound'}, inplace=True)
interval_1
```

Out[65]:

	Lower_bound	Upper bound
anyone_died	0.567781	0.998812
intercept	7.525337	7.732715

In [66]:
```
interval_2 = res_2.conf_int(alpha=0.05)
interval_2.rename(columns={0:'Lower_bound', 1:'Upper bound'}, inplace=True)
interval_2
```

Out[66]:

	Lower_bound	Upper bound
diabetes	0.045437	0.457449
intercept	7.609255	7.824403

In [67]:
```
interval_3 = res_3.conf_int(alpha=0.05)
interval_3.rename(columns={0:'Lower_bound', 1:'Upper bound'}, inplace=True)
interval_3
```

Out[67]:

	Lower_bound	Upper bound
age65	-0.005700	0.365675
intercept	7.540787	7.822654

In [68]:
```
interval_4 = res_4.conf_int(alpha=0.05)
interval_4.rename(columns={0:'Lower_bound', 1:'Upper bound'}, inplace=True)
interval_4
```

Out[68]:

	Lower_bound	Upper bound
college	-0.331831	0.056846
intercept	7.711597	7.935307

In [69]:
```
# Let's concat all frames to one to make it more readible. Let's actively nc
frames = [interval_1.iloc[0], interval_2.iloc[0], interval_3.iloc[0], interv
df = pd.DataFrame(frames)
df
```

Out[69]:

	Lower_bound	Upper bound
anyone_died	0.567781	0.998812
diabetes	0.045437	0.457449
age65	-0.005700	0.365675
college	-0.331831	0.056846

Let's also do it manually to see if we get the same results. We know the formula for our confidence interval is: $\bar{Y} \pm 1.96 \cdot \hat{SE}(\bar{Y})$

In [70]:
```
# We get the coef using the params function
beta_0 = res_1.params[0]
beta_1 = res_1.params[1]
# We can get the std using the bse function
std_0 = res_1.bse[0]
std_1 = res_1.bse[1]

conf_beta_0 = {'0.025': [beta_0 - (1.96*std_0)], '0.975': [beta_0 + (1.96* s
conf_beta_1 = {'0.025': [beta_1 - (1.96*std_1)], '0.975': [beta_1 + (1.96* s
df = pd.DataFrame([conf_beta_0, conf_beta_1], index=['Intercept', '$Beta_1$'
df.head()
```

Out[70]:

	0.025	0.975
Intercept	[0.567776934412547]	[0.9988156169027074]
$Beta_1$	[7.525334971872793]	[7.732717026575056]

We get pretty much the same results. Hence, we shall stick to using the inbuilt conf_int function.

5

As a general matter, give a definition of a 95% confidence interval in words.

The 95% confidence interval contains a set valued function of the data that contains the true value of the parameter $E[Y_i]$ in 95% of the contained samples.

6

Discuss your results. In particular, what do these data say about which groups are the most concerned about COVID-19? Do the results make sense to you? Is there anything that is surprising?

Interestingly, I find that two of the results are not statistically significant at the

$alpha = 0.05$ level. Furthermore, I find that the those that those who do know someone who died of Covid-19 are more so concerned than those that did not know someone, which seems to make sense. Those who are already suffering with a health condition, are also more concerned than those that do not, which also seems to make sense, since they might be at an increased health risk from covid-19. Stunningly, the estimated average difference between those that attend college vs those that do not is statistically insiginificant suggesting that it has no impact on the level of concern about Covid-19.

Sample Code

This sequence allows you to import all necessary libraries for this problem set:

```
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import statsmodels.api as sm
# Uncomment the line below if using Google Colab:
# !pip install stargazer
from stargazer.stargazer import Stargazer
```

Reads in data and displays first 5 rows in dataset. We add an intercept column because Python does not automatically add a constant to regressions like Stata and R:

```
# Read Dataset into a pandas dataframe
covid = pd.read_stata(“covid19.dta”)
# Add intercept to dataframe
covid[“intercept”] = 1
# Display first 5 rows of data
covid.head()
```

Estimates regression of yvar on an intercept and xvar1, with HC2 heteroskedasticity-robust standard errors. `cov_type="HC2"` corresponds to HC2 `cov_type="HC1"` corresponds to HC1

```
mod = sm.OLS(covid[[“yvar”]], covid[[“xvar1”,
“intercept”]])
res = mod.fit(cov_type=”HC2”)
```

These lines show how to make a regression table with two columns, corresponding to regressions of some variable yvar on xvar1 and a second regression of yvar on xvar2:

```
# Estimate Regressions:
mod1 = sm.OLS(covid[[“yvar”]], covid[[“xvar1”,
“intercept”]])
res1 = mod.fit(cov_type=”HC2”)
mod2 = sm.OLS(covid[[“yvar”]], covid[[“xvar2”,
“intercept”]])
res2 = mod.fit(cov_type=”HC2”)
# Create Table
table = Stargazer(models)
# Make sure the order of the independent variables is
correct
```

```
table.covariate_order([“xvar1”, “xvar2”])
# Add note about heteroskedasticity
table.add_custom_notes(["Standard errors reported in parentheses
are heteroskedasticity robust (HC2)."])
# Display table
table
```