Applied Econometrics Methods

Investigating financial distress through binary response regression model and panel data model

James/Peng Li 11/06/2024

Contents

1. INTRODUCTION

1.1 Main target

Financial distress is common in today's world, either for a company or for an individual. It is a term commonly used in corporate finance that describes any situation where an individual's or company's financial condition leaves them struggling to pay their bills, especially loan payments due to creditors. Severe, prolonged financial distress may eventually lead to bankruptcy[1].

For this study we will discuss about the individual financial distress.There are many reasons that cause the financial distress for an individual. The reasons could be:

- **Lost or reduced income**

Someone maybe unexpectedly fired or laid off from a job, or the company that you work for may go out of business, leaving you suddenly unemployed. A severe economic crisis or other circumstance may force you to take a substantial pay cut to remain employed[1].

- **Unexpected expenses**

Large unexpected expenses, such as high medical bills or an expensive car repair, are another common cause of financial difficulties[1]. Health problems can also be a common reason in many countries, as the medical cost is so high and when people are sick they can also not work. Psychological distress is one type of them.

- **Divorce**

Divorce is one of the most frequent and severe causes of financial distress. In fact, divorce is such a financial strain often on both parties that, according to studies, the rate of bankruptcy filings for single mothers in the United States is 300% higher than the national average[1]. And when divorce happens, it is also related to wealth division and other property loss etc.

- **Finance management**

Some people are not so good at managing their treasure or spend too much without good balance. High loan for housing or education can also be the reasons. According to the research, families with student loans in 2007 have higher levels of financial distress than families without such loans, and these families also transitioned to financial distress at higher rates during the early stages of the Great Recession[2].

- **Other reasons**

It can be weak background, such like low education, family status etc.

So there are many reasons which can cause financial distress. In this study we will take financial distress as dummy dependent variable. Psychological distress will be one variable to be discussed based on former study. But thinking of above so many reasons for the financial distress, this study will not be limited to psychological distress, and we are open to add other variables' for analysis step by step. In the primary data analysis and exploratory data analysis, we will try to get the primary impression and findings. And then based on these, we will choose logit model for analysis.

After the model analysis we will summarize the results for finding the most important variables and the best related models and provide technical and logical interpretation.

1.2 Variables description

Detailed Variables description please see below:

Most of the variables are continuous and dummy variables.

Variable name (variable name in data set)	Variable description	Variable type
Dependent variables		
Financial distress (financial_distress)	A dummy variable that is equal to 1 if net wealth is negative and is equal to 0 if net wealth is positive or zero.	Dummy
Net worth (wealth)	Captures the values of assets minus debts, where positive net worth means surplus wealth and negative net worth means deficit wealth.	Continuous
Key independent variables		
Psychological distress (pd)	Captures the psychological distress score of the respondents, where higher score means higher psychological distress (maximum score is 24 and minimum score is 0).	Continuous
Control variables		
Education (education)	Captures the respondents' years of schooling.	Continuous
Income (income)	Captures the combined labor income of all household members (in logs).	Continuous
Age (age)	This variable is equal to the respondents' age in years	Continuous
Male (male)	It takes the value of one if the respondent is male, and zero otherwise.	Dummy
Employed (employed)	Equal to one if the respondent is employed, and zero otherwise.	Dummy
Divorce (divorce)	Equal to one if the respondent recently experienced a divorce, and zerootherwise.	Dummy
Marriage (marriage)	Equal to one if the respondent recently got (re)married, and zero otherwise.	Dummy
Birth of child (childbirth)	Equal to one if a household member recently gave birth, and zero other-wise.	Dummy
Death of family member (familydeath)	Equal to one if a household member recently died, and zero otherwise.	Dummy
Lay off (laidoff)	Equal to one if the respondent was recently laid off from work, and zero otherwise.	Dummy
Missed work with illness (missedwork)	Captures the total number of weeks of work missed due to illness.	Continuous
White (white)	Equal to one for "White" ethnicity, and zero otherwise.	Dummy
Black (black)	Equal to one for "Black" ethnicity, and zero otherwise.	Dummy
Hispanic (Hispanic)	Equal to one for "Hispanic" ethnicity, and zero otherwise	Dummy
Other ethnicity (otherethnicity)	Equal to one for reports of ethnicity other than "Black", "Hispanic" or "White", and zero otherwise.	Dummy
College degree (collegedegree)	Equal to one for respondents with college degree. (Other options: No college degree = 5; Did not study in college = 0; No answer = 9)	Dummy
Student loan (studentloan)	Captures the respondents' student loan outstanding. This variable is only captured for those who took student loan	Continuous
Socio-economic status (socioeconomic)	Captures the respondents' socio-economic status, where higher score means higher socio-economic status. Socio-economic status variable is measured as: Total income – Poverty threshold.	Continuous
Year (year)	Captures the year of the biannual (data collected once in two years) surveys. Also, the year variable has been encoded as categorical number, where: **Year variable** / **Actual year** 0 / 2003 1 / 2005 2 / 2007 3 / 2009 4 / 2011 5 / 2013 6 / 2015 7 / 2017 8 / 2019	Categorical
ID(id)	Captures the unique identifier of the respondents.	Categorical
Head	This variable is equal to one if the respondent is the head of the household.	Dummy
Notmoved	This variable is equal to one if the head did not change within the family	Dummy
Nofamchange	This variable is equal to one if the respondent did not move in or out of the family.	Dummy

2. VARIABLES OVERALL PRIMARY STUDY

Except the variable "id", all the other variables are numeric type. To show the variables' amount, mean, standard deviation, minimum, maximum as below:

Variable	N	Mean	Std Dev	Minimum	Maximum
year	210388	3.61	3	0	8
financial_distress	210388	0.15	0	0	1
wealth	210388	220330.30	1058502	-3197000	100555000
pd	201659	3.37	4	0	24
age	204856	43.46	14	16	101
male	204905	0.73	0	0	1
white	210367	0.54	0	0	1
black	210367	0.36	0	0	1
hispanic	210367	0.02	0	0	1
otherrace	210367	0.09	0	0	1
education	197122	12.93	3	0	17
income	203764	10.74	1	0	16
employed	204905	0.75	0	0	1
divorce	210388	0.07	0	0	1
marriage	210388	0.10	0	0	1
childbirth	210388	0.19	0	0	1
familydeath	210388	0.03	0	0	1
laidoff	210381	0.06	0	0	1
missedwork	204905	1.08	4	0	75
studentloan	79823	9664.37	29444	0	700000
collegedegree	204832	1.28	2	0	9
socioeconomic	210302	37318.87	102922	-96352	6278577
head	210388	0.36	0	0	1
nofamichange	210390	0.86	0	0	1
notmoved	210390	0.88	0	0	1

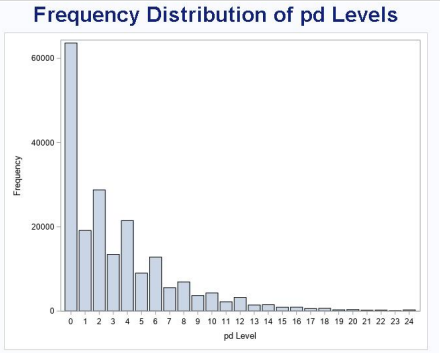
Based the different color marked above, we can generally divided the variables into different groups:

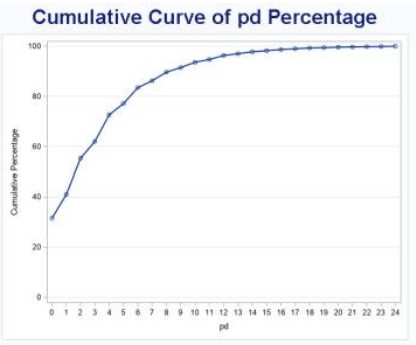
1. Financial distress and pd (key variables we want to study):

- Financial distress: it is dummy variable with 0 or 1. Around 15% observations have financial distress. No financial distress value is missing.

- pd: it has 25 levels from 0 to 24. The average is 3.37, which is relative low compared to maximum level 25. So this means most of the observations are in the low pd level area. If we look further into its distribution in below chart, pd is missing 8,731 values. 63,649 observations are with 0 pd which around 31.5% of the whole data set and have no psychological distress. And pd 0 to 6 together have occupied over 80% of whole data set.

	pd	COUNT	PERCENT
1	.	8731	.
2	0	63649	31.562687507
3	1	19142	9.4922616893
4	2	28773	14.268145731
5	3	13450	6.6696750455
6	4	21476	10.649661061
7	5	9005	4.4654590175
8	6	12817	6.3557788147
9	7	5549	2.7516748571
10	8	6920	3.4315354137
11	9	3687	1.8283339697



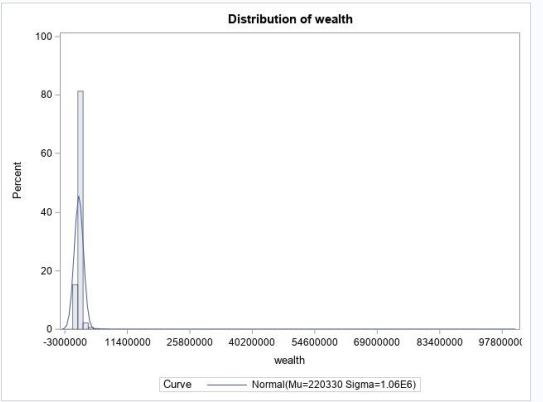


2. Wealth:

- Top 1% and bottom 1% has big gap compared to other groups. Obviously not normal distribution.

- Mean is around 220,330, which means on average positive wealth is higher than debt.

Quantiles (Definition 5)	
Level	Quantile
100% Max	100555000
99%	2782000
95%	905000
90%	487000
75% Q3	159700
50% Median	33010
25% Q1	1145
10%	-7900
5%	-27200
1%	-98770
0% Min	-3197000

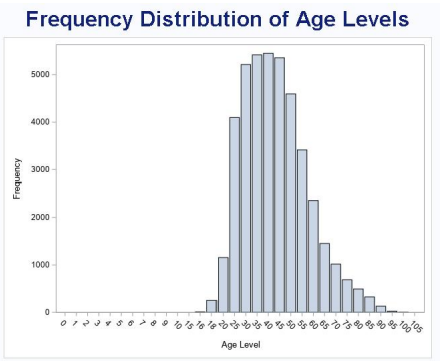


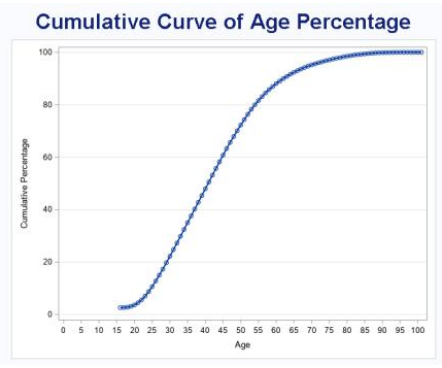
3. Gender and age:

- 73% are male.

- Age average is around 43. If we look into the distribution of the age, we can find that the youngest ones are 16 and oldest ones are 101. The age is missing 5,534 values.

	age	COUNT	PERCENT
1	.	5534	.
2	16	11	0.0053696255
3	17	36	0.0175733198
4	18	251	0.1225250908
5	19	670	0.3270590073
6	20	1155	0.5638106768
7	21	1762	0.8601163744
8	22	2308	1.126645058
9	23	3003	1.4659077596
10	24	3561	1.7382942164
11	25	4007	1.9560414323





4. White, black, hispanic and other race:

- in general very few missing values compared to the whole 210,388 rows.

- around 54% are white, 36% black, 2% hispanic and 9% other race.

5. Education: the average is around 13, overall high education level.

6. Income and employed: income 10.7 on average while the maximum 16, overall on above medium level. And 75% are employed.

7. Divorce, marriage, childbirth an family death, head, nofamichange, notmoved:

- 7% recently experienced a divorce; This could relate to pd and financial distress;

- 10% recently got (re)married;

- 19% a household member recently gave birth;

- 3% a household member recently died; This could relate to pd;

- 36% respondent is the head of the household; as head they can experience more of the pd and financial distress.

- 86% the head did not change within the family;

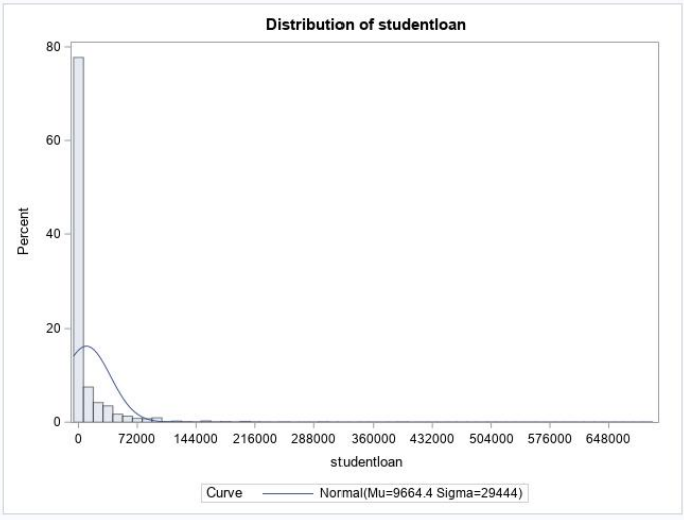
- 88% respondent did not move in or out of the family;

8. Laidoff, missedwork: 6% recently are laid off from work, and on average around 1 week missed due to sickness. Laid off may leads to high pd in short time, but financial distress in relative longer time.

9. Studentloan:

- 79,823 observations have the values. In them around 80% is 0 and most of them are below 100,000; On average is 9,664.

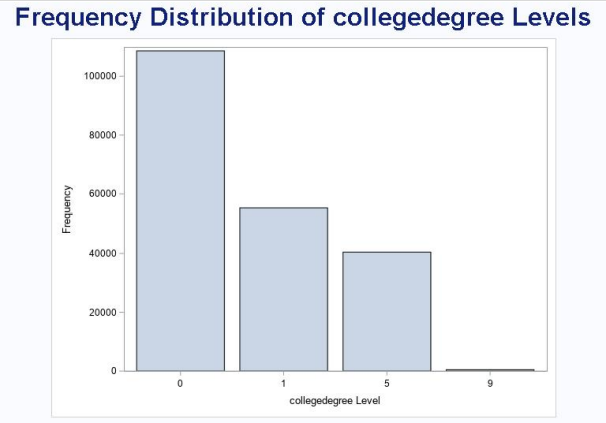
- This is a big load for the students who have loan and can cause pd and financial distress.



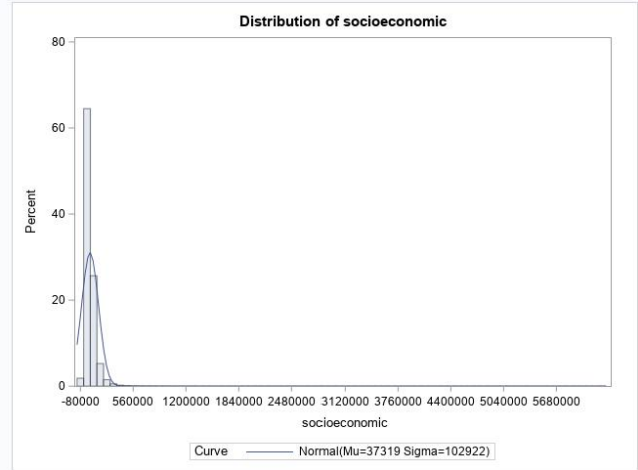
10. Colleagedegree:

- Equal to one for respondents with college degree. (Other options: No college degree = 5; Did not study in college = 0; No answer = 9)

- 204,832 observations, in which around half did not study in college. Around one quarter have college degree. Less then one quarter have no college degree.



11. Socioeconomic: this indicates socio-economic status; It has similar distribution as wealth. Mean is around 37,318.



From above data overall study we can find that except student loan, all the other variables are have high volume values. So it is a high quality data set. But there are so many variables, how to the choose the right variables for analysis will be a challenge and how to find which variable are really correlated with the financial distress will also be hard.

The variables overall study part gives us the general impression about the variables' distribution and speciality. In next part we will further explore the data and have data wrangling to have more preparation for the data analysis.

3. EXPLORATORY DATA ANALYSIS AND DATA WRANGLING

1. Correlation matrix

NAME	financial_distress	pd	age	male	white	black	hispanic	otherrace	education	income	employed	divorce	marriage	childbirth	familydeath	laidoff	missedwork	studentloan	collegedegree	socioeconomic	head	nofamichange	notmoved
financial_d	1	0.1	-0.17	-0.09	-0.07	0.08	-0.03	0.01	0.01	-0.11	0	0.07	0.05	0.04	-0.01	0.03	0.03	0.41	0.06	-0.08	0.01	-0.1	-0.05
pd	0.1	1	-0.1	-0.15	-0.07	0.08	-0.01	-0.01	-0.13	-0.22	-0.15	0.06	0.06	0	0.04	0.04	0.03	0	-0.03	-0.11	0.01	-0.1	-0.04
age	-0.17	-0.1	1	0.04	0.08	-0.08	-0.02	0	-0.03	0.14	-0.3	-0.13	-0.23	-0.26	0.15	-0.07	-0.01	-0.1	-0.05	0.06	0.12	0.02	0.02
male	-0.09	-0.15	0.04	1	0.28	-0.33	0.04	0.04	0.09	0.41	0.18	-0.04	-0.01	0.11	-0.04	0.03	-0.04	0.01	0	0.18	-0.1	0.13	0.04
white	-0.07	-0.07	0.08	0.28	1	-0.81	-0.14	-0.33	0.22	0.3	0.12	-0.04	-0.01	-0.02	-0.02	-0.01	-0.01	-0	0.03	0.21	0.03	0.13	0.07
black	0.08	0.08	-0.08	-0.33	-0.81	1	-0.1	-0.23	-0.14	-0.3	-0.14	0.05	0.01	0	0.02	0	0.01	0.02	0	-0.2	0	-0	-0.04
hispanic	-0.03	-0.01	-0.02	0.04	-0.14	-0.1	1	-0.04	-0.18	-0.05	0.03	-0.02	-0.01	0.05	-0.01	0.05	-0.01	-0	-0.05	-0.02	-0	0.01	0
otherrace	0.01	-0.01	0	0.04	-0.33	-0.23	-0.04	1	-0.09	0	0.01	-0.01	0	0.01	0	0	-0.01	-0	-0.04	-0.02	-0	-0.2	-0.06
education	0.01	-0.13	-0.03	0.09	0.22	-0.14	-0.18	-0.09	1	0.38	0.21	-0.04	-0.02	-0.03	-0.05	-0.11	0	0.2	0.27	0.27	0.04	0.07	0.05
income	-0.11	-0.22	0.14	0.41	0.3	-0.3	-0.05	0	0.38	1	0.35	-0.06	-0.08	0	-0.03	-0.06	0.02	0.08	0.09	0.39	-0.1	0.07	0.04
employed	0	-0.15	-0.3	0.18	0.12	-0.14	0.03	0.01	0.21	0.35	1	-0.01	0.02	0.06	-0.1	0.01	0.08	0.09	0.07	0.15	-0.1	0.05	0.03
divorce	0.07	0.06	-0.13	-0.04	-0.04	0.05	-0.02	-0.01	-0.04	-0.06	-0.01	1	0.17	0	-0.03	0.03	0.01	-0	0.01	-0.05	0.02	-0.2	-0.14
marriage	0.05	0.06	-0.23	-0.01	-0.01	0.01	-0.01	0	-0.02	-0.08	0.02	0.17	1	0.08	-0.02	0.04	-0.01	0.04	-0.01	-0.02	-0	-0.2	-0.1
childbirth	0.04	0	-0.26	0.11	-0.02	0	0.05	0.01	-0.03	0	0.06	0	0.08	1	-0.03	0.03	-0.01	0.07	0	-0.04	-0.1	-0	-0.08
familydeath	-0.01	0.04	0.15	-0.04	-0.02	0.02	-0.01	0	-0.05	-0.03	-0.1	-0.03	-0.02	-0.03	1	-0.01	0.01	-0	-0.01	-0.03	0.04	-0.1	-0.08
laidoff	0.03	0.04	-0.07	0.03	-0.01	0	0.05	0	-0.11	-0.06	0.01	0.03	0.04	0.03	-0.01	1	0	0	-0.03	-0.05	-0	-0	-0.02
missedwork	0.03	0.03	-0.01	-0.04	-0.01	0.01	-0.01	-0.01	0	0.02	0.08	0.01	-0.01	-0.01	0.01	0	1	0.01	0.02	-0.02	0	-0	-0.01
studentloan	0.41	0	-0.12	0.01	-0.01	0.02	-0.01	-0.01	0.2	0.08	0.09	-0.01	0.04	0.07	-0.02	0	0.01	1	0.05	0.06	-0	-0	0
collegedegr	0.06	-0.03	-0.05	0	0.03	0	-0.05	-0.04	0.27	0.09	0.07	0.01	-0.01	0	-0.01	-0.03	0.02	0.05	1	0.05	0.02	0	0
socioeconon	-0.08	-0.11	0.06	0.18	0.21	-0.2	-0.02	-0.02	0.27	0.39	0.15	-0.05	-0.02	-0.04	-0.03	-0.05	-0.02	0.06	0.05	1	0	0.07	0.05
head	0.01	0.01	0.12	-0.06	0.03	0	-0.04	-0.03	0.04	-0.06	-0.07	0.02	-0.04	-0.12	0.04	-0.02	0	-0	0.02	0	1	-0	0.07
nofamichang	-0.09	-0.06	0.02	0.13	0.13	-0.02	0.01	-0.2	0.07	0.07	0.05	-0.24	-0.17	-0.03	-0.12	-0.01	-0.01	-0	0	0.07	-0	1	0.41
notmoved	-0.05	-0.04	0.02	0.04	0.07	-0.04	0	-0.06	0.05	0.04	0.03	-0.14	-0.1	-0.08	-0.08	-0.02	-0.01	0	0	0.05	0.07	0.41	1

From the correlation matrix we can find that financial distress is relatively correlated with age and student loan. Even so the correlations are relatively low.

2. Financial distress=0 vs Financial distress=1

Divide the data sets into two parts to study the data status with financial distress=0 and financial distress=1, to get the the primary impression about the distribution of the variables in each part. Below are the extracted parts from SAS.

For financial_distress=0:

Variable	Label	Mean	Std Dev	Minimum	Maximum	Median	N
year	year	3.5676525	2.5665134	0	8.0000000	3.0000000	178227
wealth	wealth	266214.54	1143415.27	0	100555000	57500.00	178227
pd	pd	3.1942670	3.8676934	0	24.0000000	2.0000000	170837
age	age	44.4740515	14.4078152	16.0000000	101.0000000	43.0000000	173671
male	male	0.7462797	0.4351407	0	1.0000000	1.0000000	173707
white	white	0.5530006	0.4971844	0	1.0000000	1.0000000	178215
black	black	0.3423000	0.4744808	0	1.0000000	0	178215
hispanic	hispanic	0.0189041	0.1361869	0	1.0000000	0	178215
otherrace	otherrace	0.0857952	0.2800622	0	1.0000000	0	178215
education	education	12.9117165	2.7742973	0	17.0000000	12.0000000	167166
income	income	10.7904324	1.0343253	1.0000000	16.0000000	11.0000000	172708
employed	employed	0.7506548	0.4326352	0	1.0000000	1.0000000	173707
divorce	divorce	0.0638343	0.2444583	0	1.0000000	0	178227
marriage	marriage	0.0956701	0.2941392	0	1.0000000	0	178227
childbirth	childbirth	0.1838330	0.3873490	0	1.0000000	0	178227
familydeath	familydeath	0.0268927	0.1617702	0	1.0000000	0	178227
laidoff	laidoff	0.0562058	0.2303194	0	1.0000000	0	178220
missedwork	missedwork	1.0367631	3.5323754	0	75.0000000	0	173707
studentloan	studentloan	4317.41	16180.92	0	500000.00	0	66640
collegedegree	collegedegree	1.2283931	1.9025431	0	9.0000000	0	173718
socioeconomic	socioeconomic	40775.96	109598.22	-96352.00	6278577.00	20682.00	178149
head	head	0.3551594	0.4785629	0	1.0000000	0	178227
nofamichange	nofamichange	0.8778524	0.3274571	0	1.0000000	1.0000000	178227
notmoved	notmoved	0.8830985	0.3213038	0	1.0000000	1.0000000	178227

For financial_distress=1:

Variable	Label	Mean	Std Dev	Minimum	Maximum	Median	N
year	year	3.8750661	2.4119959	0	8.0000000	4.0000000	32161
wealth	wealth	-33946.95	89262.60	-3197000.00	-1.0000000	-14000.00	32161
pd	pd	4.3152294	4.4102477	0	24.0000000	3.0000000	30822
age	age	37.8003527	11.9488324	17.0000000	95.0000000	36.0000000	31185
male	male	0.6307776	0.4826020	0	1.0000000	1.0000000	31198
white	white	0.4565501	0.4981163	0	1.0000000	0	32152
black	black	0.4431451	0.4967647	0	1.0000000	0	32152
hispanic	hispanic	0.0089886	0.0943824	0	1.0000000	0	32152
otherrace	otherrace	0.0913162	0.2880628	0	1.0000000	0	32152
education	education	13.0200294	2.5562736	0	17.0000000	13.0000000	29956
income	income	10.4704727	0.9822597	0	14.0000000	11.0000000	31056
employed	employed	0.7510097	0.4324351	0	1.0000000	1.0000000	31198
divorce	divorce	0.1123099	0.3157523	0	1.0000000	0	32161
marriage	marriage	0.1417555	0.3488047	0	1.0000000	0	32161
childbirth	childbirth	0.2224122	0.4158731	0	1.0000000	0	32161
familydeath	familydeath	0.0204596	0.1415683	0	1.0000000	0	32161
laidoff	laidoff	0.0752153	0.2637426	0	1.0000000	0	32161
missedwork	missedwork	1.3117828	3.8470130	0	75.0000000	0	31198
studentloan	studentloan	36693.23	55236.39	0	700000.00	18000.00	13183
collegedegree	collegedegree	1.5549913	2.1104739	0	9.0000000	1.0000000	31114
socioeconomic	socioeconomic	18164.24	47951.10	-74581.00	2092739.00	8199.00	32153
head	head	0.3702621	0.4828823	0	1.0000000	0	32161
nofamichange	nofamichange	0.7931345	0.4050645	0	1.0000000	1.0000000	32161
notmoved	notmoved	0.8355151	0.3707208	0	1.0000000	1.0000000	32161

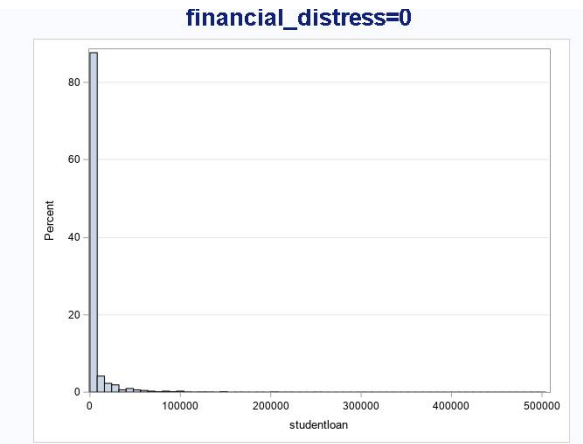
For the above two parts, we will focus on the mean values of each variable and their difference in both parts. Below the summary based on above two charts, for better comparing the difference.

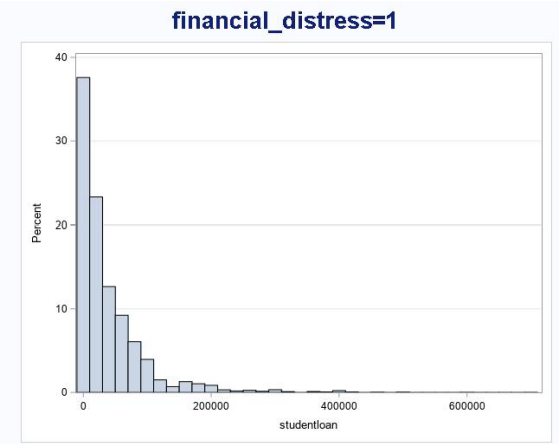
Variable	financial_distress =0/mean	financial_distress =1/mean	Change percentage	Absolute change
studentloan	4317.41	36693.23	750%	**750%**
wealth	266214.54	-33946.95	-113%	113%
divorce	0.0638343	0.1123099	76%	76%
socioeconomic	40775.96	18164.24	-55%	55%
hispanic	0.0189041	0.0089886	-52%	52%
marriage	0.0956701	0.1417555	48%	48%
pd	3.194267	4.3152294	35%	35%
laidoff	0.0562058	0.0752153	34%	34%
black	0.3423	0.4431451	29%	29%
collegedegree	1.2283931	1.5549913	27%	27%
missedwork	1.0367631	1.3117828	27%	27%
familydeath	0.0268927	0.0204596	-24%	24%
childbirth	0.183833	0.2224122	21%	21%
white	0.5530006	0.4565501	-17%	17%
male	0.7462797	0.6307776	-15%	15%
age	44.4740515	37.8003527	-15%	15%
nofamichange	0.8778524	0.7931345	-10%	10%
year	3.5676525	3.8750661	9%	9%
otherrace	0.0857952	0.0913162	6%	6%
notmoved	0.8830985	0.8355151	-5%	5%
head	0.3551594	0.3702621	4%	4%
income	10.7904324	10.4704727	-3%	3%
education	12.9117165	13.0200294	1%	1%
employed	0.7506548	0.7510097	0%	0%

*** Change Percentage** = [(financial_distress=1/mean)-(financial_distress=0/mean)]/(financial_distress=0/mean).

Through this factor of "change percentage", we can primarily compare the different influence of each variable on financial distress (0 and 1). According to this factor we primarily rank the influence of each variable on financial distress, this is not a perfect way , just for first impression. More exploratory analysis please see below:

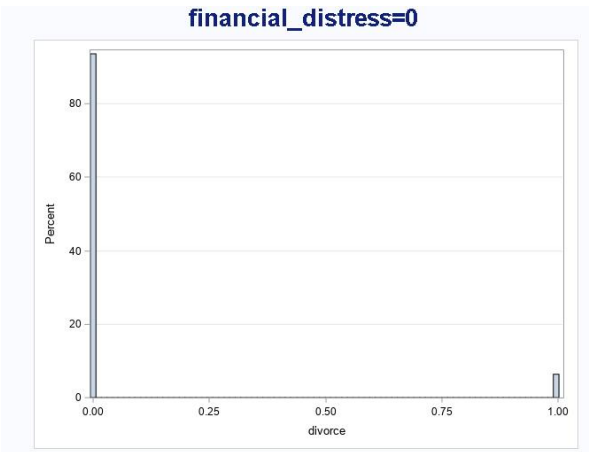
1. **Student loan:** student loan is ranking number 1 with biggest difference in two situations (financial distress=0 or 1). People with higher student loan are more likely to have financial distress compared to people who do not have lower student loan. From below two charts we can also find that in the chart financial distress=1, more people with higher student loan. And the means of student loan for each part have big gap and difference. So the general primary impression is that the student loan has very high influence on financial distress. In the model analysis, we will focus on this variable for more analysis.

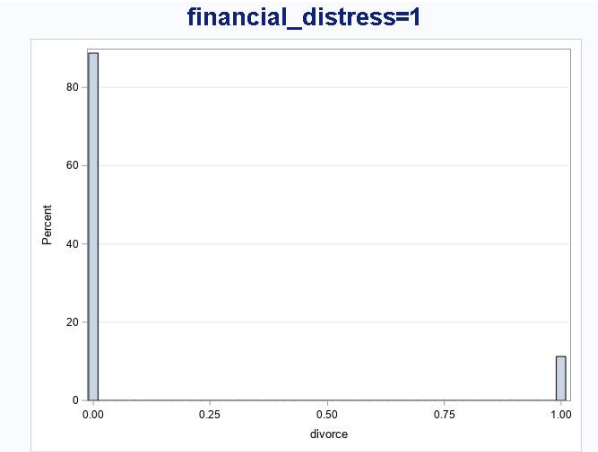




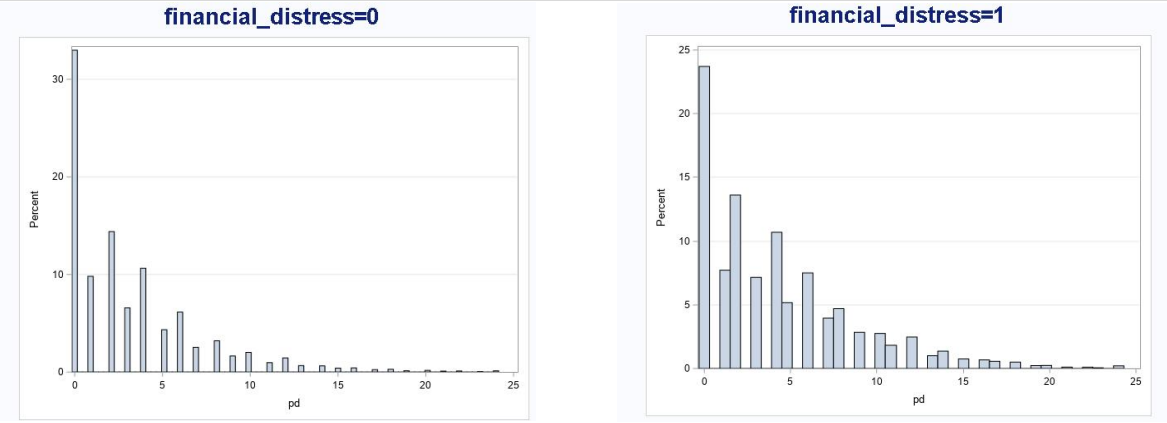
2. **wealth:** it is directly related to financial distress, so we will not take it into analysis of financial distress.

3. **divorce:** people who are recently divorced are more likely to have financial distress; This is also reasonable. When people is divorced, normally there will be big influence on the financial status, about the financial division of property etc. From below charts we can also find that the percentage of recently divorced are higher in the chart of financial distress=1.





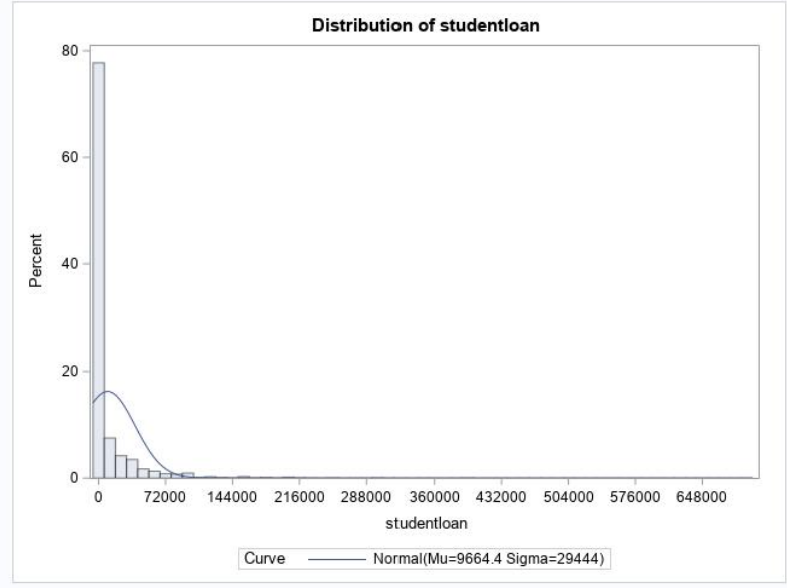
4. **pd:** from below charts, we can find that in the chart of financial distress=1, there are higher percent of observations with higher pd compared to the chart of financial distress=0.



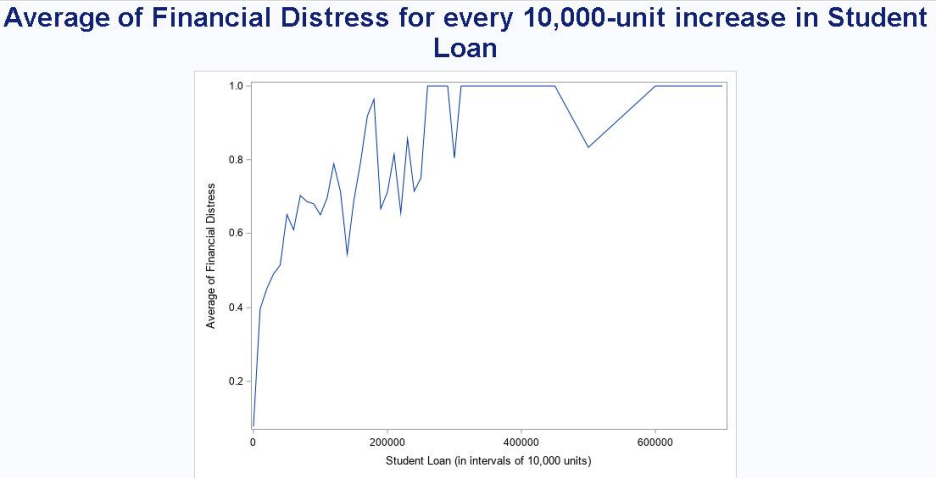
Primary summary and deeper study about student loan:

Because student loan seems highly related with financial distress, we would like to summarize above findings about student loan and have more primary investigation about it.

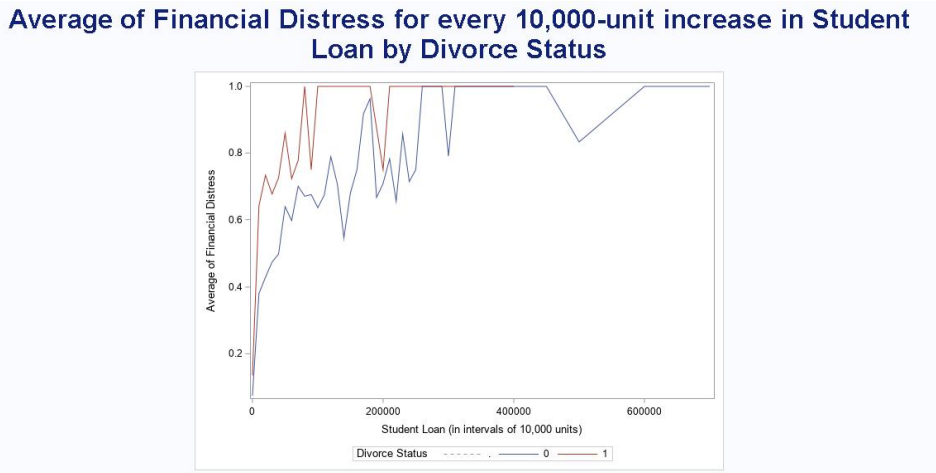
- From the "Variables overall primary study" part, we found that 79,823 observations have the student loan values. In them around 80% is 0 which means no student loan. For the observation who have student loan, most of them are below 100,000;



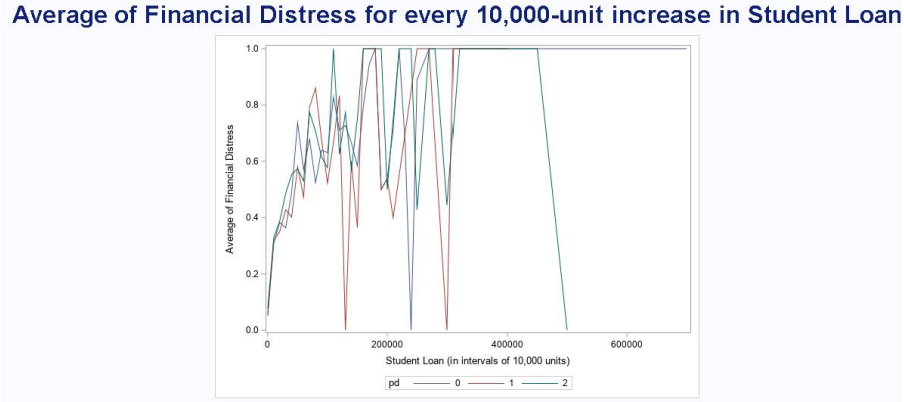
- From the two parts distribution (financial distress=0 and financial distress=1) , we can find that there is big gap of average student loan for the two part: 4,317 vs 36,693.
- **Student loan on financial distress:** for all observations with student loan, if we take 10,000 of student loan as one step or unit, and for each step we calculate the average of financial distress in this step, we can have blow curve. We can find the trend that higher student loan will leads to higher financial distress.

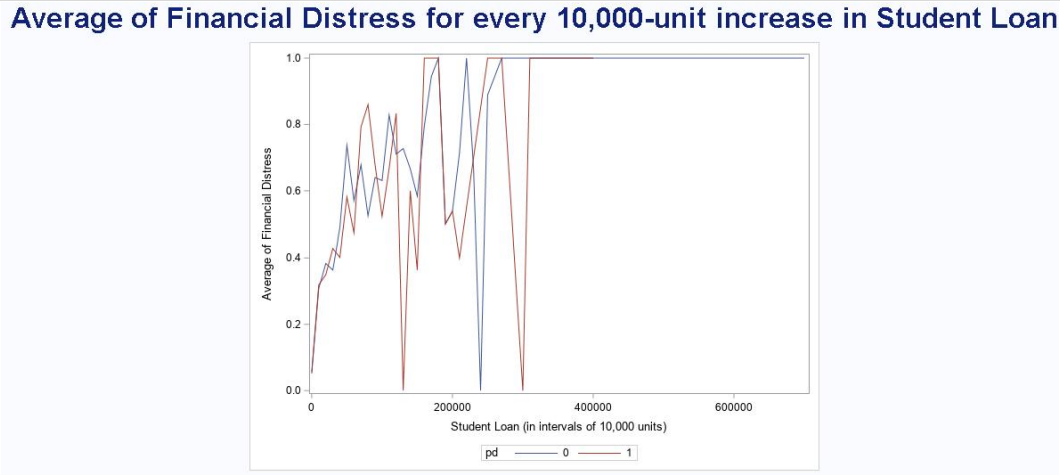


- **Student loan&divorce on financial distress:** if we further take recently divorced and not recently divorced for comparing. From below chart we can find that, for recently divorced people, when they have student loan, they are more likely to have financial distress. And for not recently divorced people, the curve is similar as the above curve, because in the data set most of the observations are the not recently divorced, so this curve looks similar as above one for the whole section but has slight difference.

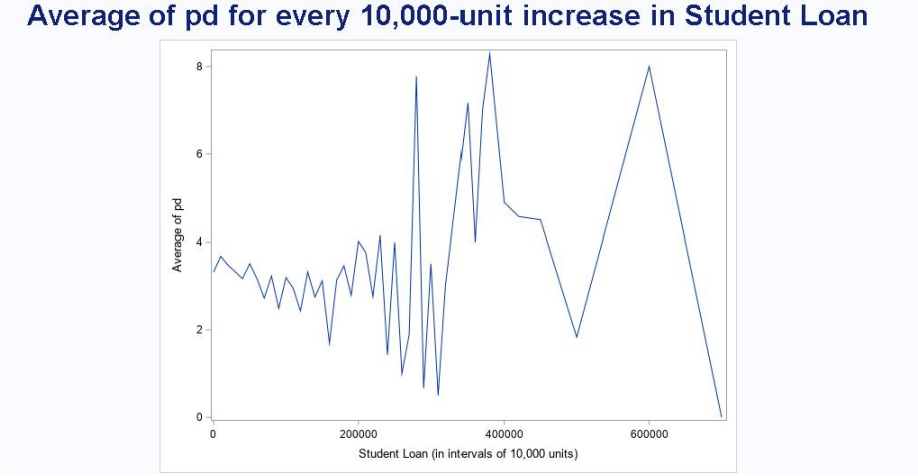


- **Student loan&pd on financial distress:** If we take different pd level to check the student loan influence on financial distress, we can find that there is not very clear conclusion there. We take the representative pd level 0, 1, 2 for analysis, as they have high portion.





- **Student loan on pd:** for all observations with student loan, same as above, if we take 10,000 of student loan as one step or unit, and for each step we calculate the average of pd in this step, we can find blow curve. According to the distribution of student loan, we can find that 99% are below 150,000, so we just focus on the curve before 200,000. And we can find that there is even slight decrease of pd in this area. But for some cases with higher student loan than 200,000, we can also find high pd. This is not related to financial distress, just would like to know the correlation between student loan and pd, as additional small study.



Distribution of student loan

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	700000
99%	150000
95%	60000
90%	30000
75% Q3	2100
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

So after more deeper analysis from student loan side. We can find that in general when student loan increase, the financial distress also increase. And for recently divorced people, their situation is even more sever (in below model 2, we will add new variable student loan*divorce for more analysis). And there is not very clear association between student loan and pd.

So in general, in the green marked variables in the chart (financial distress=0 vs financial distress=1) will be highly considered in the model analysis in next section.

3. Data wrangling:

Considering that the values of student loan and socioeconomic are much bigger compared to other variables, so we will divide them through 1000 and use the 1000 as the unit.

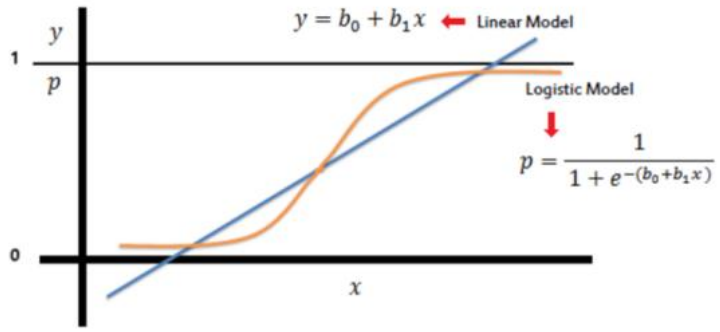
4. Binary response regression model

4.1 Methodology

Binary choice models are a class of econometric models used to analyze situations where the dependent variable can take one of two possible outcomes. These models are particularly useful in understanding decision-making processes where choices are dichotomous, such as “yes or no” decisions. Common examples include whether or not an individual decides to purchase a product, vote for a particular candidate, or adopt a technology[4].

In statistics, the logistic model (or logit model) is a statistical model that models the log-odds of an event as a linear combination of one or more independent variables. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination)[5].

The purpose of this report is to create a Binary Choice Model that estimates the probability of a person has financial distress (1) as opposed to not have financial distress (0). The Logit Model will be used which utilizes the logistic function F, transforming a regression model, so that the fitted model is bound within 0 and 1 and does not produce illogical results where the likelihood of a financial distress is negative or more than one[6]. We use the log to transform below linear model into logistic model, to locate the y between 0 and 1.



A Logistic Model with two independent variables is written as:

$$\text{Logit } (P_i = 1) = \log\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 Q_i + \beta_2 R_i + \mu_i$$

The Logit (Pi = 1) result must be transformed using the following formula to get the probability of financial distress

$$P_i = \frac{e^{\beta_0 + \beta_1 Q_1 + \beta_2 R_i}}{1 + e^{\beta_0 + \beta_1 Q_1 + \beta_2 R_i}}$$

4.2 Model analysis

Model 1: Use “pd, student loan and divorce” for analysis

Based on above analysis, we found that student loan and divorce has big influence on financial distress, so firstly we take the two independent variables together with pd for analysis.

Model Information		
Data Set	WORK.ASSESSMENT2_DATA_WITHHEAD2	
Dependent Variable	financial_distress	financial_distress
Number of Observations	76772	
Name of Distribution	Logistic	
Log Likelihood	-28191.82245	

Number of Observations Read	210390
Number of Observations Used	76772
Missing Values	133618

Class Level Information		
Name	Levels	Values
financial_distress	2	0 1

Response Profile		
Ordered Value	financial_distress	Total Frequency
1	1	12704
2	0	64068

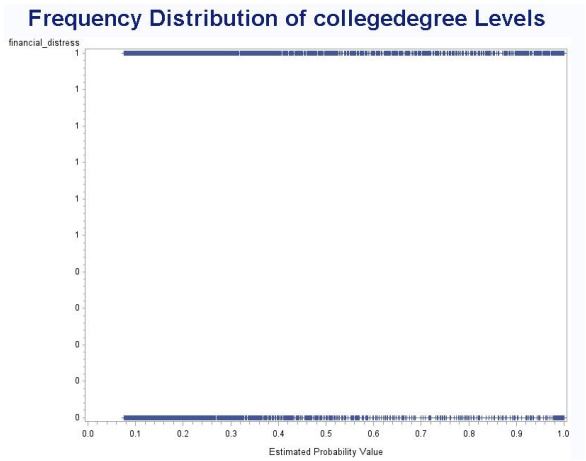
Because there are not so many student loan observations, we there are 76,772 observations are used for this model. Accordingly there are 12,704 financial distress=1 and 64,098 financial distress=0.

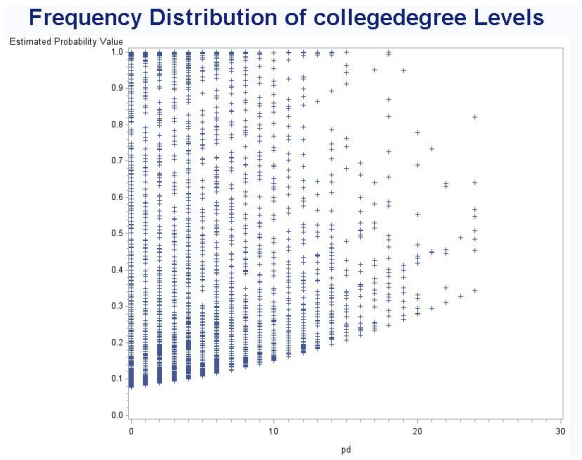
Type III Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
pd	1	1001.1030	<.0001
student_loan	1	6992.8126	<.0001
divorce	1	335.5891	<.0001

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-2.4908	0.0168	-2.5238	-2.4578	21856.7	<.0001
pd	1	0.0769	0.0024	0.0721	0.0817	1001.10	<.0001
student_loan	1	0.0417	0.0005	0.0407	0.0427	6992.81	<.0001
divorce	1	0.6783	0.0370	0.6057	0.7509	335.59	<.0001

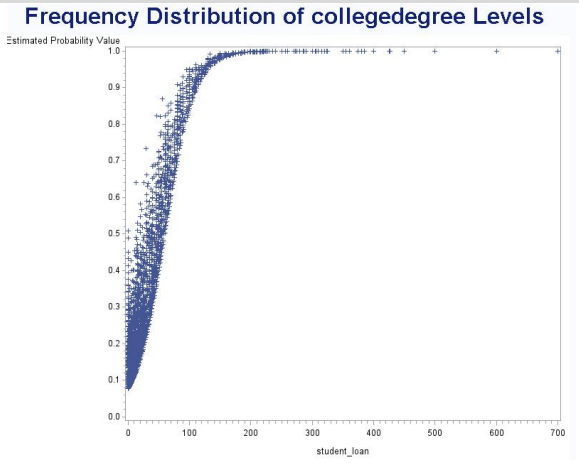
The estimates from logistic regression suggest that all the three variables are significant explanatory variables, where all increase the likelihood of having financial distress.

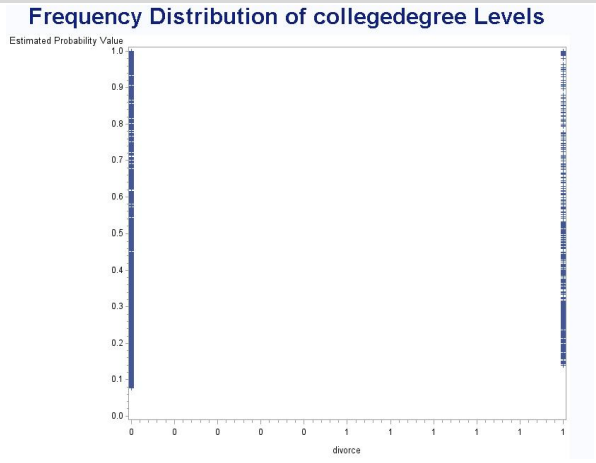
$Zi = x_iB_j = -2.4908 + 0.0769pd_i + 0.0417student\_loan_i + 0.6783divorce_i + error_i$





From the pd and estimated probability value chart, we can find that the lower boundary is increasing according to the pd increase. This is also some indication that overall the higher pd may more leads to financial distress. If we take 0 pd level for analysis, we can find that there are more dots are located near the 0.1 financial distress compared to higher pd such like pd=10, that means lower pd in general may lead to higher likelihood of lower financial stress.





From above student_loan and estimated probability value, we can find the clear trend that higher student loan may leads to higher financial distress. And the student loan>150T, it is very likely leading to financial distress. And from the student loan distribution chart when financial distress=0 in section 3, we can also find that nearly all the student loan observations are below 150T. So this conclusion is reasonable and consistent with former findings.

From above divorce and estimated probability value, fore divorce=1, we can find that there are more dots near 0 rather near 1, it is reasonable, as only small part of them have financial distress. According to the distribution chart of divorce when financial distress=1 in section 3, we can find that only around 11% of divorced have financial distress.

We use the SAS code to calculate probabilities:

	student_loan	socio_economic	Zi	probabilities
80019	60	-10.386	0.4726	0.6159989566
80020	85	4.557	1.3613	0.7959709005
80021	85	-16.845	1.8227	0.8608897903
80022	100	-18.54	2.4482	0.9204297206
80023	.	-26.699	.	.
80024	.	-15.046	.	.
80025	.	-15.662	.	.
80026	.	-17.352	.	.
80027	.	-18.39	.	.
80028	60	-10.386	0.4726	0.6159989566
80029	85	4.557	1.3613	0.7959709005
80030	85	-16.845	1.8227	0.8608897903
80031	100	26.273	2.4482	0.9204297206

From above chart, we can also find higher student_loan normally have higher probabilities of financial distress.

We also use code in SAS to calculate the average marginal effects:

Variable	Mean	Minimum	Maximum
MEffpd	0.0083527	1.935778E-13	0.0192221
MEffStudentLoan	0.0045313	1.050163E-13	0.0104280
MEffDivorce	0.0736806	1.707588E-12	0.1695615

The mean marginal effect of pd is 0.0083527 which is the expected instantaneous change in the probability of financial distress to a term deposit for every 1 increase in the pd level, holding all other variables constant. For

student loan, one thousand increase of student loan will increase the probability of financial distress with 0.0045313. And for divorce, the recently divorced people will increase the probability of financial distress with 0.0736806 compared to the recently not divorced people.

We can use log odds ratios instead of parameter estimates. The log odds ratios can be obtained from SAS:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
pd	1.080	1.075	1.085
student_loan	1.043	1.042	1.044
divorce	1.970	1.832	2.119

The log odds ratios can be directly related with the probability, unlike the original parameter estimates. An odds ratio is can be interpreted as the multiplicative change in the odds for a unit change in the independent variable. The results in the above table suggests that 1 increase in pd increases the odds of having financial distress by a factor of 1.080. And 1000 dollars increase in student loan increases the odds of having financial distress by a factor of 1.043. On the other hand, the odds of having financial distress is 1.970 times as large for the divorced individuals as compared to their counterparts.

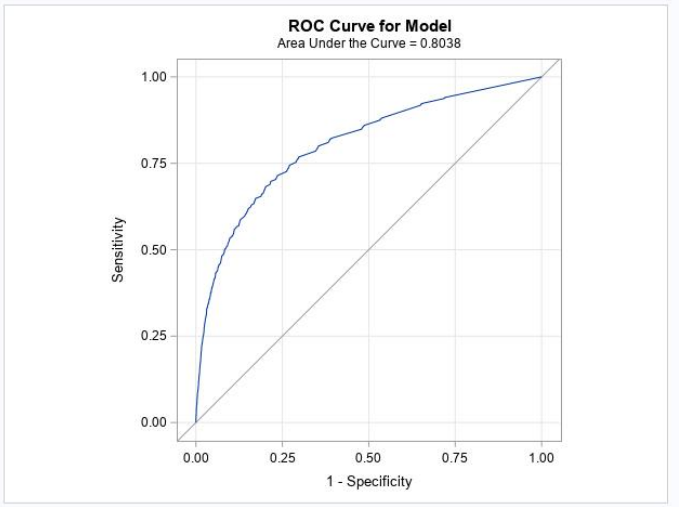
Model fit: in order to get the model fit stats, we now use the Logistic procedure with a link function to probit (Like calling probit procedure and linking to logitistic).

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	68888.196	56802.829
SC	68897.444	56839.823
-2 Log L	68886.196	56794.829

The above table shows that the AIC and SC for the main-effects model are smaller than for the saturated model, indicating that the main-effects model might be the preferred model.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12091.3671	3	<.0001
Score	13887.3221	3	<.0001
Wald	7375.8059	3	<.0001

Similar to F Test in OLS, results of the likelihood ratio test and the efficient score test for testing the joint significance of the explanatory variables are included in the "Testing Global Null Hypothesis: BETA=0" table. In the above table, the small P-values (<0.0001) suggests that the hypothesis that all slope parameters are equal to zero should be rejected.



From the above ROC curve, we find that our model has 0.8038 area under the curve, which is not bad given the inadequacy of the data and absence of many important variables related to financial distress.

Model 2: Use "pd, student loan, divorce and student loan*divorce" for analysis

For this model we would like to add student loan*divorce as additional variable for further study.

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-2.4739	0.0168	-2.5070	-2.4409	21557.6	<.0001
pd	1	0.0771	0.0024	0.0723	0.0819	1001.95	<.0001
student_loan	1	0.0403	0.0005	0.0393	0.0413	6418.25	<.0001
divorce	1	0.3974	0.0450	0.3091	0.4856	77.89	<.0001
student_loan*divorce	1	0.0374	0.0032	0.0313	0.0436	140.57	<.0001

The estimates from logistic regression suggest that all the four variables are significant explanatory variables, where all increase the likelihood of having financial distress.

$Zi = x_iB_j = -2.4739 + 0.0771pd_i + 0.0403student\_loan_i + 0.3974divorce_i + 0.0374student\_loan_i*divorce_i + error_i$

Compared with model 1 as below, other coefficients are nearly the same, only the divorce's coefficient has been changed, so this means this variable is partly explained by student_loan*divorce.

Model 1 equation: $Zi = x_iB_j = -2.4908 + 0.0769pd_i + 0.0417student\_loan_i + 0.6783divorce_i + error_i$

Variable	Mean	Minimum	Maximum
MEffpd	0.0083310	1.034154E-14	0.0192774
MEffStudentLoan	0.0043540	5.404773E-15	0.0100749
MEffDivorce	0.0429311	5.329146E-14	0.0993392
MEffStudentLoanDivorce	0.0576353	0	7.9004077

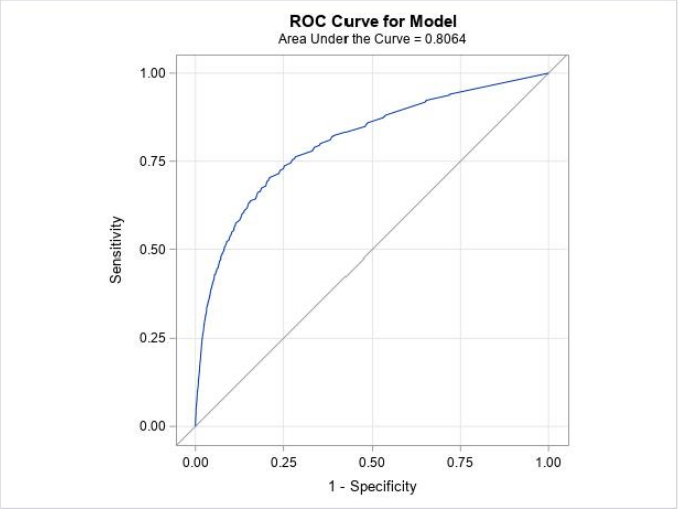
In the above chart, the mean for pd and student loan is nearly the same as the model 1. The mean for divorce is smaller than model 1, which is partly contributed by studentloan*divorce for explaining financial distress.

Model fit: in order to get the model fit stats, we now use the Logistic procedure with a link function to probit (Like calling probit procedure and linking to logitistic).

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	68888.196	56611.108
SC	68897.444	56657.351
-2 Log L	68886.196	56601.108

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12285.0875	4	<.0001
Score	14035.3476	4	<.0001
Wald	7389.7959	4	<.0001

The above table shows that the AIC and SC for the main-effects model are smaller than for the saturated model, indicating that the main-effects model might be the preferred model.



From the above ROC curve, we find that this model has 0.8064, which is not bad given the inadequacy of the data and absence of many important variables related to financial distress. And this result is similar as model 1.

Model 3: Use all the green marked variables on slide 10 for analysis (student_loan divorce socio_economic hispanic marriage pd laidoff black collegedegree missedwork familydeath childbirth white male age)

In Model 3 we input most of the variables in a Stepwise Logistic Regression which determined the best variables to use to predict whether the people will have financial distress.

Summary of Stepwise Selection								
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	student_loan		1	1	12425.5869		<.0001	
2	age		1	2	1568.9296		<.0001	age
3	male		1	3	1036.8536		<.0001	male
4	socio_economic		1	4	675.1771		<.0001	
5	pd		1	5	363.6901		<.0001	pd
6	collegedegree		1	6	271.6835		<.0001	collegedegree
7	divorce		1	7	96.0352		<.0001	divorce
8	missedwork		1	8	52.9403		<.0001	missedwork
9	laidoff		1	9	14.5775		0.0001	laidoff

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.9337	0.0457	417.0118	<.0001
student_loan	1	0.0441	0.000549	6436.6370	<.0001
divorce	1	0.3788	0.0393	92.9355	<.0001
socio_economic	1	-0.00826	0.000272	923.8685	<.0001
pd	1	0.0492	0.00264	348.2286	<.0001
laidoff	1	0.2079	0.0545	14.5473	0.0001
collegedegree	1	0.0897	0.00548	267.2458	<.0001
missedwork	1	0.0200	0.00278	51.8172	<.0001
male	1	-0.4351	0.0249	305.7757	<.0001
age	1	-0.0265	0.000918	833.0488	<.0001

After the stepwise analysis, from above chart we can find that only 9 variables are kept in the end, which are significant. And the most important one is student loan.

$Zi = x_iB_j = -0.9337 + 0.0441\ student\_loan_i + 0.3788divorce_i + 0.00826socio\_economic_i + 0.0492pd_i + 0.2079laidoff_i + 0.0897collegedegree_i + 0.0200missedwork_i + 0.4351male_i + 0.0265age_i + error_i$

Variable	Mean	Minimum	Maximum
MEffStudentLoan	0.0044923	6.510804E-22	0.0110157
MEffDivorce	0.0386209	5.597387E-21	0.0947032
MEffSocioEconomic	-0.000841836	-0.0020643	-1.22009E-22
MEffPd	0.0050186	7.273531E-22	0.0123062
MEffLaidoff	0.0211911	3.071258E-21	0.0519631
MEffCollegedegree	0.0091404	1.324733E-21	0.0224134
MEffMissedwork	0.0020416	2.958892E-22	0.0050062
MEffMale	-0.0443638	-0.1087854	-6.42971E-21
MEffAge	-0.0027013	-0.0066239	-3.915E-22

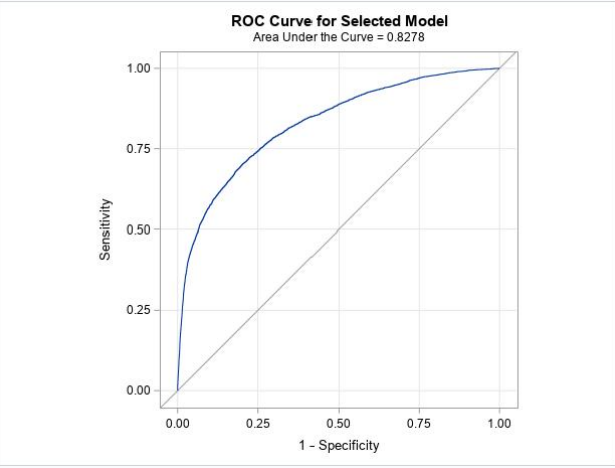
From above chart and also above formula we can find that higher socio_economic will decrease the probability of having financial distress. Increase of collegedegree which means no college degree with value 5 or without answer with value 9 (this part is very small portion) will more likely to have financial distress. And male has lower probability of having financial distress compared to female. And along with the age increase, it will also be less likely to have financial distress. This means, we or the government should care more about the financial distress of people who have student loan, divorced, lower socioEconomic, higher pd, Laid off, no college degree, missed work, female and young people.

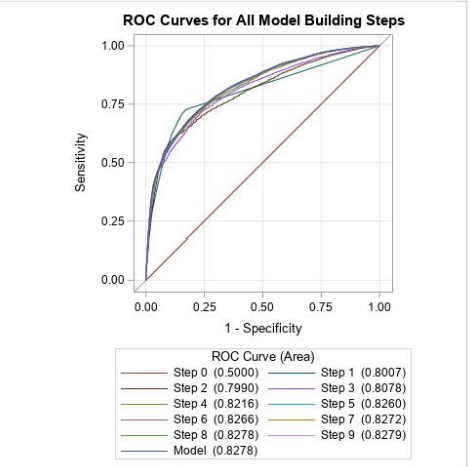
Model fit: we now use the Logistic procedure with a link function to probit (Like calling probit procedure and linking to logitistic).

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	66014.044	50892.226
SC	66023.253	50993.523
-2 Log L	66012.044	50870.226

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	15141.8182	10	<.0001
Score	15875.3925	10	<.0001
Wald	9415.9891	10	<.0001

The above table shows that the AIC and SC for the main-effects model are smaller than for the saturated model, indicating that the main-effects model might be the preferred model.





From the above ROC curve, we find that this model has 0.8064, which is higher than model 1 and 2. This also indicates better performance. And from above right chart, we can find that the ROC results is improving after each step.

Model 4: use all variables except student_loan in model 3 (divorce socio_economic hispanic marriage pd laidoff black collegedegree missedwork familydeath childbirth white male age)

Because there are only 76,772 observations have the student loan values, while there are in total 210,390 observations, so we want to exclude student loan variable to have bigger size of data set for analysis. And also we would like to compare the results excluded student loan with the results with student loan in model 3.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	age		1	1	5611.1024		<.0001	age
2	male		1	2	1382.0617		<.0001	male
3	pd		1	3	1003.6516		<.0001	pd
4	collegedegree		1	4	560.4034		<.0001	collegedegree
5	socio_economic		1	5	604.0971		<.0001	
6	divorce		1	6	203.1944		<.0001	divorce
7	hispanic		1	7	141.9022		<.0001	hispanic
8	white		1	8	84.3252		<.0001	white
9	missedwork		1	9	85.9268		<.0001	missedwork
10	laidoff		1	10	49.3195		<.0001	laidoff
11	marriage		1	11	18.3104		<.0001	marriage

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.3286	0.0266	152.1221	<.0001
divorce	1	0.2822	0.0217	168.4159	<.0001
socio_economic	1	-0.00419	0.000145	831.1749	<.0001
hispanic	1	-0.8089	0.0633	163.4804	<.0001
marriage	1	0.0839	0.0196	18.3039	<.0001
pd	1	0.0398	0.00151	697.7952	<.0001
laidoff	1	0.1732	0.0251	47.7515	<.0001
collegedegree	1	0.0814	0.00310	688.2841	<.0001
missedwork	1	0.0150	0.00161	86.2845	<.0001
white	1	-0.1329	0.0141	88.9493	<.0001
male	1	-0.2591	0.0148	307.9719	<.0001
age	1	-0.0328	0.000547	3593.8688	<.0001

Compared to model 3, more variables are kept after the stepwise analysis. That means without student loan, more variables are needed to interpret the financial distress. This also shows the importance of student loan for explaining the financial distress.

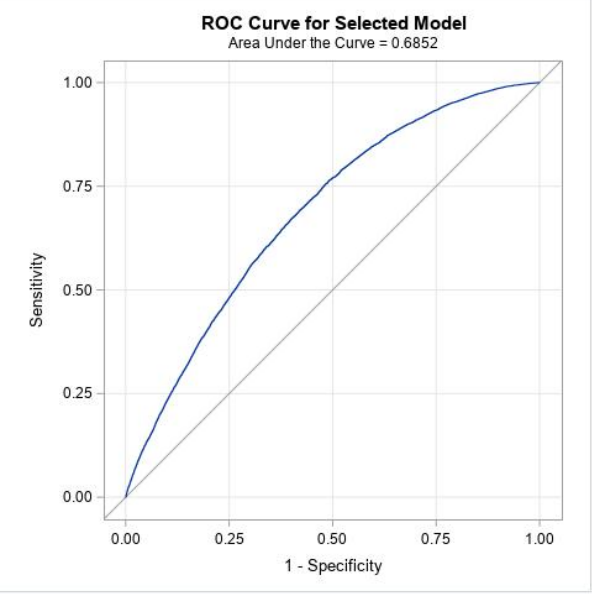
$$Zi = x_iB_j = -0.3286 + 0.2822divorce_i - 0.00419socio\_economic_i - 0.8089mariage_i + 0.0398pd_i + 0.1732laidoff_i + 0.0814\ collegedegree_i + 0.0150missedwork_i - 0.1329white - 0.2591male - 0.0328age + error_i$$

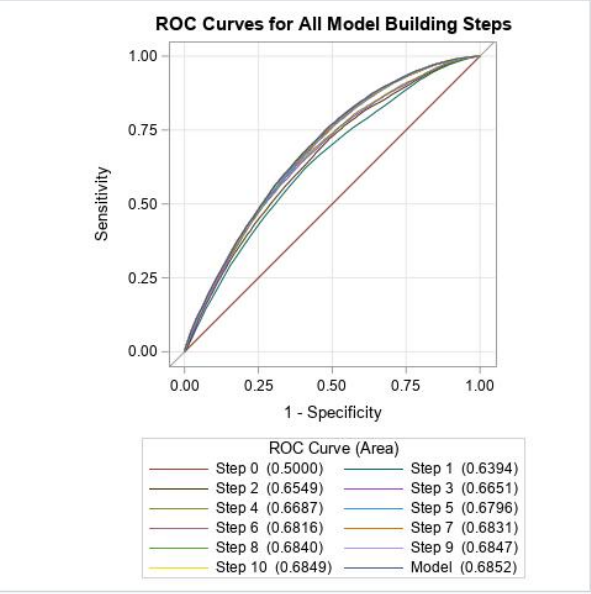
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
divorce	1.326	1.271	1.384
socio_economic	0.996	0.996	0.996
hispanic	0.445	0.393	0.504
marriage	1.087	1.046	1.130
pd	1.041	1.038	1.044
laidoff	1.189	1.132	1.249
collegedegree	1.085	1.078	1.091
missedwork	1.015	1.012	1.018
white	0.876	0.852	0.900
male	0.772	0.750	0.794
age	0.968	0.967	0.969

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	166604.39	155734.78
SC	166614.57	155856.99
-2 Log L	166602.39	155710.78

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	10891.6076	11	<.0001
Score	9975.1003	11	<.0001
Wald	9558.4834	11	<.0001

The above table shows that the AIC and SC for the main-effects model are smaller than for the saturated model, indicating that the main-effects model might be the preferred model.





From the above ROC curve, we find that our model has 0.6852, which is much lower compared to other three models. This also indicate the importance of student loan in the model and necessary of having the student loan in the model.

4.3 Model analysis results in summary

Four models are discussed below that considers various variables that may affect whether a client subscribes to a term deposit. The following table summarizes the key estimates for each model and model performance:

Test	Model 1	Model 2	Model 3	Model 4
AIC	56802.829	56611.108	50892.226	155734.78
SC	56839.823	56657.351	50993.523	155856.99
-2Log L	56794.829	56601.108	50870.226	155710.78
Wald Chi-Square	7375.8059	7389.7959	9415.9891	9558.4834
AUC	0.8038	0.8064	0.8278	0.6852

Compared to model 1, model 2 has improved results, but not so big. Model 3 is the preferred model with the smallest Akaike information criterion (AIC), Schwarz criterion (SC) and negative of twice the log likelihood (-2 Log L) and highest Wald Chi-Square results. AUC measures the model's ability to predict whether a person will have financial distress. From the AUC results for all four models the third one is the best with highest value.

5. Panel data models

5.1 Methodology

Panel data is a data type increasingly used in research in economics, social sciences, and medicine. Its primary characteristic is that the data variation goes jointly over space (e.g. individuals, firms, countries) and time (e.g. years, months). Panel data allow examination of problems which cannot be handled by cross-section data or time-series data[7].

Because the data are measured over time for different ids, so would to use panel data for more analysis.

Fixed effect model is a type of model used for panel data that employs dummies to account for variables that affect the dependent variable y cross-sectionally but do not vary over time. There is time invariant variables such like gender, white, black, hispanic, so we would capture the heterogeneity that is encapsulated in u_i through fixed effect model that allows for different intercepts for each cross sectional unit. And then we will also try time fixed effects model and random effects model for more analysis.

5.2 Model analysis

Model 5: Panel model with fixed one-way estimate-cross sectional (Fixone) ,with same variables for model 3

Model Description	
Estimation Method	FixOne
Number of Cross Sections	22566
Time Series Length	4

Fit Statistics			
SSE	3777.6075	DFE	51203
MSE	0.0738	Root MSE	0.2716
R-Square	0.6278		

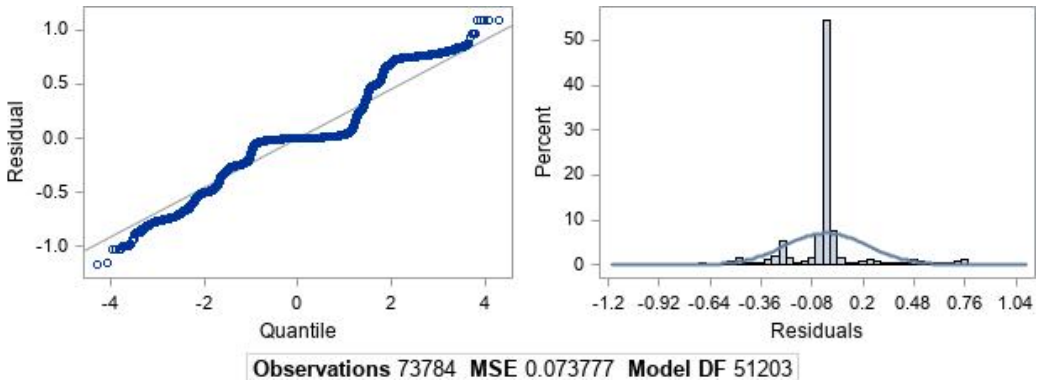
F Test for No Fixed Effects			
Num DF	Den DF	F Value	Pr > F
22565	51203	2.52	<.0001

The R-square value shows that these fixed effects models explain 62.78% of the variations in financial distress. F-statistic means that for the fixed effects model we reject the null hypothesis of the dummy coefficients being zero. This means that there is evidence of there being a cross-section fixed effect (based on sector).

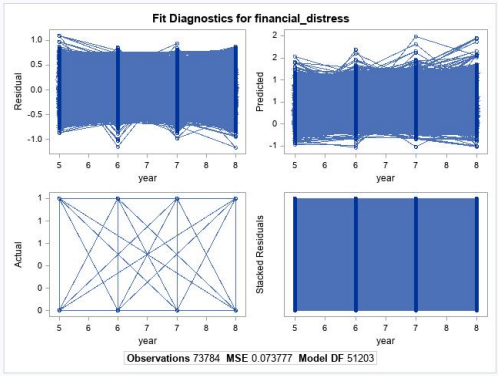
Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	1.061009	0.1937	5.48	<.0001	Intercept
student_loan	1	0.004395	0.000072	60.81	<.0001	
divorce	1	0.029527	0.00572	5.16	<.0001	divorce
socio_economic	1	0.000013	0.000017	0.77	0.4438	
hispanic	1	-0.14506	0.0662	-2.19	0.0285	hispanic
marriage	1	-0.01731	0.00533	-3.25	0.0012	marriage
pd	1	0.001165	0.000432	2.70	0.0070	pd
laidoff	1	0.022092	0.00792	2.79	0.0053	laidoff
black	1	-0.03889	0.0200	-1.95	0.0518	black
collegedegree	1	0.002403	0.000869	2.76	0.0057	collegedegree
missedwork	1	0.001822	0.000412	4.42	<.0001	missedwork
familydeath	1	-0.01695	0.00921	-1.84	0.0657	familydeath
childbirth	1	0.025092	0.00442	5.68	<.0001	childbirth
white	1	0.003015	0.0183	0.16	0.8692	white
male	1	-0.02069	0.00595	-3.48	0.0005	male
age	1	-0.00254	0.000245	-10.38	<.0001	age

P value of some variables such like “socio_economic”, “black”, “familydeath”, “white” are bigger than 5%, that means the estimated coefficients are not significant. One pd level increase can leads to the 0.001165 probability increase of financial distress. And 1000 dolor increase of student loan can leads to 0.004395 probability increase of financial distress. Compared to not recently divorced, the recently divorced can leads to the 0.029527 probability increase of financial distress.



From above chat, we can find that residuals are nearly normal distributed, but not fully, still some relative higher residuals in the left part compared to the right part. The residuals are varies a bit along the straight line in the left chart.



Residuals distributions in different years have no big difference, so there is no clear heteroskedasticity.

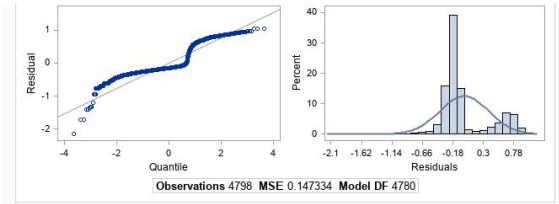
Model 6: Panel model with one way fixed effects - time (fixonetime) ,with same variables for model 3

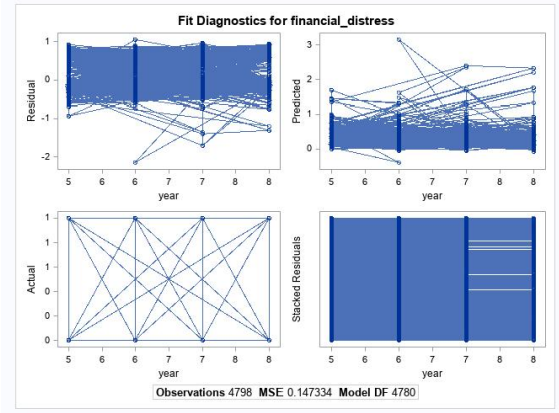
Model Description	
Estimation Method	FixOneTm
Number of Cross Sections	3329
Time Series Length	4

Fit Statistics			
SSE	704.2582	DFE	4780
MSE	0.1473	Root MSE	0.3838
R-Square	0.2217		

F Test for No Fixed Effects			
Num DF	Den DF	F Value	Pr > F
3	4780	7.92	<.0001

The R-square value is much smaller. And the number of cross section is only 3329 cross sections. This reduces hugely of the data size.





The residual distribution is also not good, with some outliers.

Model 7: Panel model with two ways fixed effects (fixtwo) ,with same variables for model 3

Model Description	
Estimation Method	FixTwo
Number of Cross Sections	22566
Time Series Length	4

Fit Statistics			
SSE	3760.3294	DFE	51200
MSE	0.0734	Root MSE	0.2710
R-Square	0.6295		

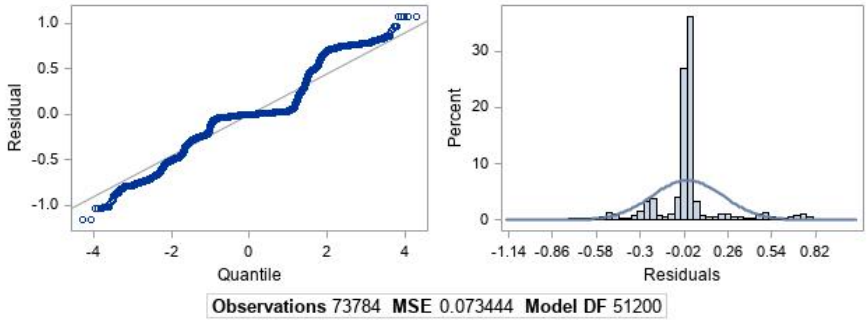
F Test for No Fixed Effects			
Num DF	Den DF	F Value	Pr > F
22568	51200	2.54	<.0001

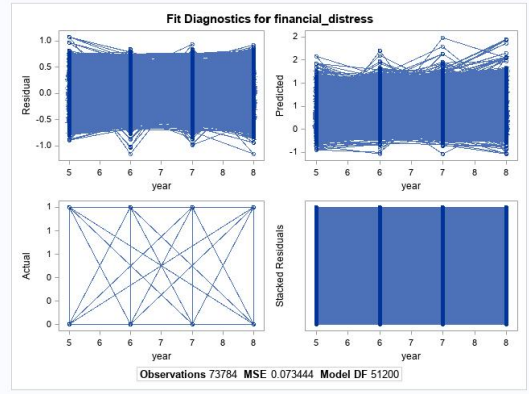
The R-square value shows that these fixed effects models explain 62.95% of the variations in financial distress, which is a little higher than model 5. F-statistic is also nearly same as model 5

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	0.968777	0.1934	5.01	<.0001	Intercept
student_loan	1	0.004462	0.000072	61.76	<.0001	
divorce	1	0.027125	0.00571	4.75	<.0001	divorce
socio_economic	1	7.798E-6	0.000017	0.45	0.6545	
hispanic	1	-0.14671	0.0661	-2.22	0.0264	hispanic
marriage	1	-0.01721	0.00532	-3.24	0.0012	marriage
pd	1	0.001187	0.000431	2.75	0.0059	pd
laidoff	1	0.014375	0.00791	1.82	0.0693	laidoff
black	1	-0.03571	0.0199	-1.79	0.0735	black
collegedegree	1	0.001911	0.000870	2.20	0.0280	collegedegree
missedwork	1	0.00176	0.000411	4.28	<.0001	missedwork
familydeath	1	-0.01268	0.00919	-1.38	0.1679	familydeath
childbirth	1	0.016918	0.00444	3.81	0.0001	childbirth
white	1	0.005854	0.0183	0.32	0.7486	white
male	1	-0.02208	0.00594	-3.72	0.0002	male
age	1	-0.00136	0.000257	-5.30	<.0001	age

P value of variables "socio_economic", "laidoff", "black", "familydeath", "white" are bigger than 5%, that means the estimated coefficients are not significant. Compared to model 5, coefficient of "laidoff" is also not significant.





Above charts are similar as model 5 with similar conclusions.

Model 8: Panel model with one way random effects (Ranone), with same variables for model 3

Model Description	
Estimation Method	RanOne
Number of Cross Sections	3329
Time Series Length	4

Fit Statistics			
SSE	461.7894	DFE	4783
MSE	0.0965	Root MSE	0.3107
R-Square	0.1960		

Variance Component Estimates	
Variance Component for Cross Sections	0.05402
Variance Component for Error	0.097857

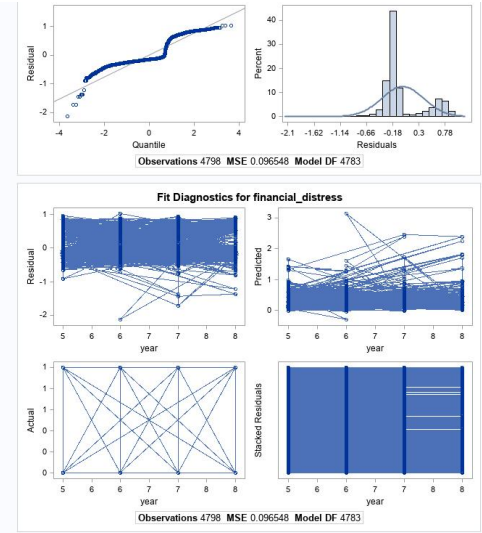
Hausman Test for Random Effects			
Coefficients	DF	m Value	Pr > m
13	13	47.22	<.0001

Similar as model 6 the time fixed effects, there are only only 3329 cross sections. And the R-square is much smaller with 0.1960.

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	0.252908	0.0332	7.61	<.0001	Intercept
student_loan	1	0.007395	0.000238	31.09	<.0001	
divorce	0	0	.	.	.	divorce
socio_economic	1	-0.00022	0.000089	-2.53	0.0115	
hispanic	1	-0.18753	0.2248	-0.83	0.4042	hispanic
marriage	1	-0.02906	0.0129	-2.26	0.0242	marriage
pd	1	0.009804	0.00128	7.67	<.0001	pd
laidoff	1	0.075655	0.0249	3.04	0.0023	laidoff
black	1	-0.04004	0.0245	-1.64	0.1016	black
collegedegree	1	0.020719	0.00288	7.18	<.0001	collegedegree
missedwork	1	-0.00025	0.00152	-0.16	0.8701	missedwork
familydeath	1	-0.08527	0.0520	-1.64	0.1013	familydeath
childbirth	1	0.006882	0.0157	0.44	0.6615	childbirth
white	1	-0.02648	0.0244	-1.08	0.2782	white
male	1	-0.01613	0.0111	-1.46	0.1447	male
age	1	-0.00222	0.000560	-3.96	<.0001	age

Only few variables' P values are less than 5%.



Residuals' distribution is also not good. So this model is not a good one. Main reason is that cross sectional only few id has full time series data. This limit the data quality from cross sectional side.

Model 9: Panel model with two way random effects (Rantwo), with same variables for model 3

Model Description

Estimation Method	RanTwo
Number of Cross Sections	3329
Time Series Length	4

Fit Statistics

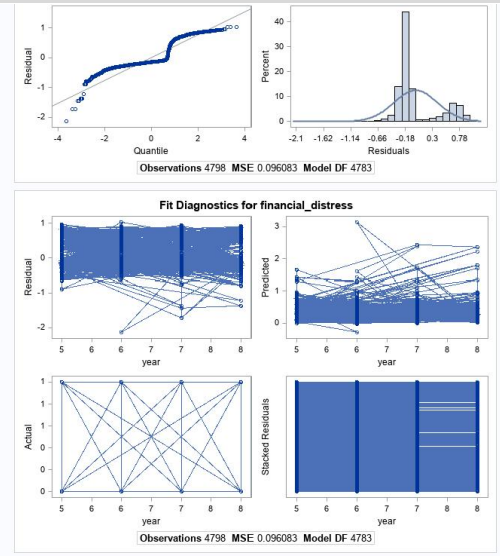
SSE	459.5635	DFE	4783
MSE	0.0961	Root MSE	0.3100
R-Square	0.1959		

Variance Component Estimates

Variance Component for Cross Sections	0.053875
Variance Component for Time Series	0.000829
Variance Component for Error	0.097346

Hausman Test for Random Effects

Coefficients	DF	m Value	Pr > m
13	13	48.78	<.0001



All the results are nearly the same as random one.

5.3 Model analysis results in summary

From above panel models 5-9, we can find that, in general model 5 Fixone and model 7 Fixtwo are the best with high R2 and also relative better residuals distribution. Of course, these models can be further improved through adjusting the variables in the model.

And the relative better results with model 5 and 7 also indicate that, for time invariant variables the two models can well capture the heterogeneity that is encapsulated in ui through the fixed effect model that allows for different intercepts for each cross sectional unit.

And not so many id has full time series data and not so many variables' data vary over time but would be assumed to be the same across entities at each given point in time. The data to be used for analysis is also small amount. So the time fixed effects model is not good, similar reason for random effects model.

Test	**Model 5 Fixone**	**Model 6 Fixonetime**	**Model 7 Fixtwo**	**Model 8 Randomone**	**Model 9 Randomtwo**
R2	0.6278	0.2217	0.6295	0.1960	0.1959
Pr>F or Pr>m	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Number of cross sections	22566	3329	22566	3329	3329

6. Comparing binary response regression model and panel data model

From the independent variable side, binary response regression model is suitable, because the financial distress is 0 or 1 data.

From the variable side, we can find that there are same id which are observed over multiple time periods and also have some time invariant variables. From this side the panel data model is more suitable.

We take best binary response regression model 4 and the best panel model 7 to look at the estimates:

	Model 4			Model 7			Difference	
Variables	**Estimate 1**	**Standard error1**	**Pr > ChiSq**	**Estimate 2**	**Standard error 2**	**Pr > |t|**	**Estimate1/ Estimate2**	**Standard error1/Standard error2**
Intercept	-0.9337	0.0457	<.0001	0.968777	0.1934	<.0001	-0.96	0.24
student_loan	0.0441	0.000549	<.0001	0.004462	0.000072	<.0001	9.88	7.63
divorce	0.3788	0.0393	<.0001	0.027125	0.00571	<.0001	13.96	6.88
socio_economic	-0.00826	0.000272	<.0001	0.000007798	0.000017	0.6545	-1059.25	16.00
pd	0.0492	0.00264	<.0001	0.001187	0.000431	0.0059	**41.45**	6.13
laidoff	0.2079	0.0545	0.0001	0.014375	0.00791	0.0693	14.46	6.89
collegedegree	0.0897	0.00548	<.0001	0.001911	0.00087	0.028	46.94	6.30
missedwork	0.02	0.00278	<.0001	0.00176	0.000411	<.0001	11.36	6.76
male	-0.4351	0.0249	<.0001	-0.02208	0.00594	0.0002	19.71	4.19
age	-0.0265	0.000918	<.0001	-0.00136	0.000257	<.0001	19.49	3.57

From the above chart, we can find that the two models have different intercepts, so the variable coefficients have relative big gaps. But if we calculate the Estimate1/Estimate2 and Standard error1/Standard error2 ratios, we can find that the ratios do not vary big between different variables. And for socioeconomic, for model 7 it is not significant, so we will not consider it, same for laidoff in model 7. From this side we can find some consistence between the two models. For pd, through the different coefficient, we can find that pd plays a more important role in model 4 to explain the financial distress then in model 7.

For model 4 it has high AUC 0.8278, and for model 7 it explain 62.95% of the variance of financial distress.

So in general both models are relative good to interpret the dependent variable.

7. Summary

As introduced in the beginning, for financial distress there are so many reasons behind it. Based on last study about psychological distress influence on wealth, we all know that psychological is one important reason for financial status. For this study besides the psychological distress we are more open to study other variables for getting wider insight.

Through the variables primary study, we got the overall impression about all the data status and each variable's distribution.

In the exploratory data analysis we primarily analyze the data. Through dividing the data into two parts by financial distress=0 and financial distress=1, we got more findings of different variables' distribution in the two areas. And based on comparing the mean of each variable in the two areas, we tried to ranking the influence of different variables on financial distress. This is not strict way, but gives us some general ideas to choose some relative important variables for the model analysis in next part.

Firstly we used binary response regression model logit and tried four models:

Model 1: based on the exploratory data analysis we found that higher student loans and recently divorced ones have clearly more located in the area of financial distress=1, so we besides pd, we also take student loan, divorce for analysis. Though ROC and other results we find that this model is general a good one.

Model 2: we want to further study the influence of student loan and divorce, so we add one new variable loan*divorce for analysis. It slightly improved the key indicators in the final results. But not so big.

Model 3: we take more variables into analysis, and we use logit stepwise for analysis. And we find that some variables which are not significant are not listed in the final results. Through this way we find the most related variables to build up the model. The got good ROC result.

Model 4: For knowing the importance of the variable divorce, we took out this variable out of model 3. After that we found the overall results are obviously worse compared to model 3. So this means student loan has high influence on financial distress.

Secondly we used panel data model for analysis and tried 5 models, and all the 5 models use the variables which are used in model 3, for better comparing later.

Model 5: Panel model with fixed one-way estimate-cross sectional (Fixone), the result is good with 0.6278 R2. And the residuals have also relative good distribution, although not optimal, generally normally distributed and there is no clear heteroskedasticity.

Model 6: Panel model with one way fixed effects - time (fixonetime) , the result is no so good with low R2 values and low number of cross sections, that means low data size for analysis.

Model 7: Panel model with two ways fixed effects (fixtwo), the best model o the 5 panel time models. The R2 has slight increase compared to model 5.

Model 8: Panel model with one way random effects (Ranone), similar like model 6, no so good result.

Model 9: Panel model with two way random effects (Rantwo), similar like model 6 and 8, no so good result.

The data set does have time invariant variables such like gender, white, black, hispanic, so the fixed effects model results are good, and no so many ids have full time series data, so this limits the to be used data number of cross sections and can lead to not good result of time fixed effects model or random effects model.

In conclusion, we find that the best model is model 3 and model 7 with adding more variables for analysis. So this means financial distress is a result which is caused by many reasons together, although the contribution weights are different.

Through this study form the financial distress side, we should care the group with "student loan". Some study clearly shows that student loan debt is positively related to all indicators of financial stress and hardship[8]. If the person is recently divorced, the situation is even worse, and after divorce one partner needs to pay another partner's student loan if it happens after the marriage[9]. And also thinking of other variables such like psychological distress, if some people have experienced certain situations at the same time, it will even tougher. Obviously student loan, divorce, psychological distress etc. also influence each other. So these groups need special support during this period of time. Regarding back to student loan, right policy and more support for university education can also be considered by the public and government, as it does not influence the education itself, but also the life after graduation, about the financial distress, family status etc. If we have more background information or policy details about the student loan, such like how to apply it, how can it be approved and what is the interest etc. , or if we can have more data about how long can people pay it back, we can have more clear understanding and have more deeper study of this part. Same for recently divorced or high psychological distress people. So more information for deeper study can be executed for understand the background reasons of financial distress and its key influencing factors. And more work need to be done to further improve the models.

Reference:

[1] Team, C. (2023, November 1). *Financial distress.* Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/commercial-lending/financial-distress/

[2] Bricker, J., & Thompson, J. (2016). DOES EDUCATION LOAN DEBT INFLUENCE HOUSEHOLD FINANCIAL DISTRESS? AN ASSESSMENT USING THE 2007-2009 SURVEY OF CONSUMER FINANCES PANEL. *Contemporary Economic Policy*, 34(4), 660–677. https://doi.org/10.1111/coep.12164

[3] Carter, K. N., Blakely, T., Collings, S., Gunasekara, F. I., & Richardson, K. (2009). What is the association between wealth and mental health? Journal of Epidemiology and Community Health, 63(3), 221–226. https://doi.org/10.1136/jech.2008.079483

[4] Quickonomics. (2024, April 6). *Binary Choice Models Definition & Examples - Quickono mics.* https://quickonomics.com/terms/binary-choice-models/#:~:text=Binary%20choice%20models%20are%20a%20class%20of%20econometric,are%20dichotomous%2C%20such%20as%20%E2%80%9Cyes%20or%20no%E2%80%9D%20decisions.

[5] Wikipedia contributors. (2024l, June 8). *Logistic regression*. Wikipedia. https://en.wikipedia.org/wiki/Logistic_regression

[6] Bank Data – Proposed Solutions for applied economic lecture on the stream of Massey University

[7] Biørn, E. (2016). Econometrics of panel data. *In Oxford University Press eBooks.* https://doi.org/10.1093/acprof:oso/9780198753445.001.0001

[8]Zhan, M. (n.d.). *Financial Stress and Hardship among Young Adults*: The Role of Student Loan Debt. ScholarWorks at WMU. https://scholarworks.wmich.edu/jssw/vol49/iss3/5/

[9] Helhoski, A. (2024, January 11). Student loans in divorce: *What happens to the debt?* NerdWallet. https://www.nerdwallet.com/article/loans/student-loans/what-happens-to-student-loans-in-divorce

Complete code:

```
*1. Assign a library to a path where you have stored the data.;
%let path=D:\1. BA\Applied Econometric Methods\Assignment\AEMDATA;
libname AEM"&path";
ods graphics on;

* Import the CSV file *;
PROC IMPORT OUT=WORK.Assessment2_data_withhead
    DATAFILE="&path.\Assessment2_data_withhead.xlsx"
    DBMS=xlsx
    REPLACE;
    GETNAMES=YES;
RUN;

* 2. Overall data status;
proc contents data=WORK.Assessment2_data_withhead;
run;
proc means data=WORK.Assessment2_data_withhead;
 var year -- notmoved;

*pd primary study*;
proc freq data=WORK.Assessment2_data_withhead noprint;
   tables pd / out=pd_freq;
run;
proc sgplot data=pd_freq;
   vbar pd / response=Count;
   xaxis label="pd Level";
   yaxis label="Frequency";
   title "Frequency Distribution of pd Levels";
run;
proc freq data=WORK.Assessment2_data_withhead noprint;
   tables pd / out=pd_freq;
run;
data pd_freq_cumulative;
   set pd_freq;
   retain cum_percent 0;
   cum_percent + percent;
run;
proc sgplot data=pd_freq_cumulative;
   series x=pd y=cum_percent / markers lineattrs=(thickness=2);
   xaxis label='pd' values=(0 to 24 by 1);
   yaxis label='Cumulative Percentage' grid;
   title 'Cumulative Curve of pd Percentage';
run;

* Wealth *;
proc freq data=WORK.Assessment2_data_withhead noprint;
   tables wealth / out=wealth_freq;
run;

        proc univariate data=WORK.Assessment2_data_withhead;
 var wealth;
```

```
 histogram / normal;
run;
ods output Histogram=hist_table;

*age primary study*;
/* Step 1: Create the dataset with the frequency distribution for age */
proc freq data=WORK.Assessment2_data_withhead noprint;
   tables age / out=age_freq;
run;
/* Step 2: Plot the frequency distribution of age */
proc sgplot data=age_freq;
   vbar age / response=Count;
   xaxis label="Age Level" values=(0 to 10 by 1 15 16 18 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105);
   yaxis label="Frequency";
   title "Frequency Distribution of Age Levels";
run;
/* Step 3: Calculate cumulative percentage */
data age_freq_cumulative;
   set age_freq;
   retain cum_count 0 total_count 0;
   /* Calculate the total count for cumulative percentage calculation */
   if _N_ = 1 then do;
      total_count = 0;
      do until (last);
         set age_freq end=last;
         total_count + Count;
      end;
      cum_count = 0;
      set age_freq;
   end;
   cum_count + Count;
   cum_percent = (cum_count / total_count) * 100;
run;
/* Step 4: Plot the cumulative percentage curve */
proc sgplot data=age_freq_cumulative;
   series x=age y=cum_percent / markers lineattrs=(thickness=2);
   xaxis label='Age' values=(0 to 101 by 1);
   yaxis label='Cumulative Percentage' grid;
   title 'Cumulative Curve of Age Percentage';
run;

* studentloan*;
proc freq data=WORK.Assessment2_data_withhead noprint;
   tables studentloan / out=studentloan_freq;
        proc univariate data=WORK.Assessment2_data_withhead;
 var studentloan;
 histogram / normal;
run;
ods output Histogram=hist_table;

/* collegedegree */
proc freq data=WORK.Assessment2_data_withhead noprint;
   tables collegedegree / out=collegedegree_freq;
run;
```

```
proc sgplot data=collegedegree_freq;
   vbar collegedegree / response=Count;
   xaxis label="collegedegree Level";
   yaxis label="Frequency";
   title "Frequency Distribution of collegedegree Levels";
run;

/* Calculate the frequency of each level in the socioeconomic variable */
proc freq data=WORK.Assessment2_data_withhead noprint;
   tables socioeconomic / out=socioeconomic_freq;
run;

/* Generate histogram and univariate analysis for socioeconomic */
proc univariate data=WORK.Assessment2_data_withhead;
   var socioeconomic;
   histogram / normal;
run;
ods output Histogram=hist_table;


* 3.Exploratory data analysis*;
* show the correlation of wealth, pd...socioecomomic;
proc corr data=WORK.Assessment2_data_withhead pearson nosimple noprob plots=none outp=corr_out;
 var financial_distress pd --notmoved;
proc print data=corr_out noobs;
run;

*overall study about financial distrees=0 and financial distress=1*;
proc means data=WORK.Assessment2_data_withhead  (where=(financial_distress=0))  mean std min max median n
vardef=df
qmethod=os;
var year wealth pd        age    male white  black  hispanic      otherrace    education    income       employed
       divorce       marriage    childbirth    familydeath laidoff        missedwork studentloan collegedegree
       socioeconomic     head  nofamichange      notmoved;
run;
proc means data=WORK.Assessment2_data_withhead  (where=(financial_distress=1))  mean std min max median n
vardef=df
qmethod=os;
var year wealth pd        age    male white  black  hispanic      otherrace    education    income       employed
       divorce       marriage    childbirth    familydeath laidoff        missedwork studentloan collegedegree
       socioeconomic     head  nofamichange      notmoved;
run;

*student loan study*
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sort data=WORK.Assessment2_data_withhead out=_HistogramTaskData;
by financial_distress;
run;
proc sgplot data=_HistogramTaskData;
by financial_distress;
histogram studentloan /;
yaxis grid;
run;
ods graphics / reset;
```

```
proc datasets library=WORK noprint;
delete _HistogramTaskData;
run;

*divorce study*;
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sort data=WORK.Assessment2_data_withhead out=_HistogramTaskData;
by financial_distress;
run;
proc sgplot data=_HistogramTaskData;
by financial_distress;
histogram divorce /;
yaxis grid;
run;

*pd study*;
proc sgplot data=_HistogramTaskData;
by financial_distress;
histogram pd /;
yaxis grid;
run;
ods graphics / reset;
proc datasets library=WORK noprint;
delete _HistogramTaskData;
run;

*Deeper study of student loan*;
*for every 10,000 increase of student loan, to calculate the average of financial distress*;
/* Step 1: Sort the dataset by variable studentloan */
proc sort data=Assessment2_data_withhead;
   by studentloan;
run;
/* Step 2: Create intervals of 10,000 units in studentloan */
data Assessment2_intervals;
   set Assessment2_data_withhead;
   interval = floor(studentloan / 10000);
run;
/* Step 3: Calculate the average of financial_distress within each interval */
proc means data=Assessment2_intervals noprint;
   by interval;
   var financial_distress;
   output out=avg_distress mean=avg_financial_distress;
run;
/* Step 4: Prepare data for plotting by creating X-axis values */
data plot_data;
   set avg_distress;
   X_axis_value = interval * 10000;
run;
/* Step 5: Plot the data */
proc sgplot data=plot_data;
   series x=X_axis_value y=avg_financial_distress;
   xaxis label='Student Loan (in intervals of 10,000 units)';
   yaxis label='Average of Financial Distress';
   title 'Average of Financial Distress for every 10,000-unit increase in Student Loan';
```

```
run;

*student loan and divorce*;
/* Step 1: Sort the dataset by variable studentloan */
proc sort data=Assessment2_data_withhead;
   by studentloan;
run;
/* Step 2: Create intervals of 10,000 units in studentloan */
data Assessment2_intervals_0 Assessment2_intervals_1;
   set Assessment2_data_withhead;
   interval = floor(studentloan / 10000);
   /* Include the 'divorce' variable */
   if divorce = 0 then output Assessment2_intervals_0;
   else if divorce = 1 then output Assessment2_intervals_1;
run;
/* Step 3: Calculate the average of financial_distress within each interval for each value of divorce */
proc means data=Assessment2_intervals_0 noprint;
   by interval;
   var financial_distress;
   output out=avg_distress_0 mean=avg_financial_distress_0;
run;

proc means data=Assessment2_intervals_1 noprint;
   by interval;
   var financial_distress;
   output out=avg_distress_1 mean=avg_financial_distress_1;
run;
/* Step 4: Merge the datasets */
data plot_data;
   merge avg_distress_0(in=a) avg_distress_1(in=b);
   by interval;
   if a then y_curve = 0;
   else if b then y_curve = 1;
   X_axis_value = interval * 10000;
run;
/* Step 5: Plot the data */
proc sgplot data=plot_data;
   series x=X_axis_value y=avg_financial_distress_0 / group=y_curve;
   series x=X_axis_value y=avg_financial_distress_1 / group=y_curve;
   xaxis label='Student Loan (in intervals of 10,000 units)';
   yaxis label='Average of Financial Distress';
   title 'Average of Financial Distress for every 10,000-unit increase in Student Loan';
run;

*pd and student loan*;
/* Step 1: Sort the dataset by variable studentloan */
proc sort data=Assessment2_data_withhead;
   by studentloan;
run;
/* Step 2: Create intervals of 10,000 units in studentloan */
data Assessment2_intervals;
   set Assessment2_data_withhead;
   interval = floor(studentloan / 10000);
run;
```

```
/* Step 3: Sort the dataset by interval and pd */
proc sort data=Assessment2_intervals;
   by interval pd;
run;
/* Step 4: Calculate the average of financial_distress within each interval for each value of pd */
proc means data=Assessment2_intervals noprint;
   by interval pd; /* Include pd in the BY statement */
   var financial_distress;
   output out=avg_distress mean=avg_financial_distress;
run;
/* Step 5: Prepare data for plotting by creating X-axis values */
data plot_data;
   set avg_distress;
   X_axis_value = interval * 10000;
run;
/* Step 6: Filter the data for specific values of pd */
data plot_data_filtered;
   set plot_data;
   where pd in (0, 1); /* Filter for pd values */
run;
/* Step 7: Plot the data */
proc sgplot data=plot_data_filtered;
   series x=X_axis_value y=avg_financial_distress / group=pd;
   xaxis label='Student Loan (in intervals of 10,000 units)';
   yaxis label='Average of Financial Distress';
   title 'Average of Financial Distress for every 10,000-unit increase in Student Loan';
run;

*pd distribution of student loan*;
/* Step 1: Sort the dataset by variable studentloan */
proc sort data=Assessment2_data_withhead;
   by studentloan;
run;
/* Step 2: Create intervals of 10,000 units in studentloan */
data Assessment2_intervals;
   set Assessment2_data_withhead;
   interval = floor(studentloan / 10000);
run;
/* Step 3: Calculate the average of pd within each interval */
proc means data=Assessment2_intervals noprint;
   by interval;
   var pd; /* Change variable to pd */
   output out=avg_pd mean=avg_pd; /* Change output variable to avg_pd */
run;
/* Step 4: Prepare data for plotting by creating X-axis values */
data plot_data;
   set avg_pd; /* Change dataset to avg_pd */
   X_axis_value = interval * 10000;
run;
/* Step 5: Plot the data */
proc sgplot data=plot_data;
   series x=X_axis_value y=avg_pd; /* Change y variable to avg_pd */
   xaxis label='Student Loan (in intervals of 10,000 units)';
   yaxis label='Average of pd'; /* Change y-axis label */
```

```
  title 'Average of pd for every 10,000-unit increase in Student Loan'; /* Change title */
run;

*3.2 data wrangling*;
data Assessment2_data_withhead1;
   set WORK.Assessment2_data_withhead;
   student_loan = studentloan/1000;
   socio_economic = socioeconomic/1000;
run;
data Assessment2_data_withhead2;
   set WORK.Assessment2_data_withhead1;
   drop studentloan socioeconomic;
run;

proc contents data=WORK.Assessment2_data_withhead2;
run;
proc means data=WORK.Assessment2_data_withhead2;
 var year -- socio_economic;


*4. Binary response regression model*;

*model 1*;
ods graphics / imagemap=on ;
proc probit data=WORK.Assessment2_data_withhead2;
model financial_distress (Event = '1')= pd student_loan divorce /
d=logistic ;
output  out=ccdyhat  prob=PredictedValues ;
run;
*plotting*;
ods graphics / imagemap=on;
proc gplot data=ccdyhat;
plot financial_distress*PredictedValues PredictedValues*pd  PredictedValues*student_loan
PredictedValues*divorce;
run;
*possibility*;
Data WORK.Assessment2_data_withhead3 ;
set WORK.Assessment2_data_withhead2;
Zi= -2.4908 + 0.0769*pd+ 0.0417*student_loan + 0.6783*divorce;
probabilities = exp(Zi)/(1+exp(Zi));
run;
*Meff*;
proc logistic data=WORK.Assessment2_data_withhead2 outest=logparms(rename=(pd=tpd
student_loan=tstudent_loan divorce=tdivorce));
   model financial_distress(event="1") = pd student_loan divorce;
   output out=outlog p=p;
run;
data outlog;
   if _n_=1 then set logparms;
   set outlog;
   MEffpd = p * (1 - p) * tpd;
   MEffStudentLoan = p * (1 - p) * tstudent_loan;
   MEffDivorce = p * (1 - p) * tdivorce;
run;
```

```
proc print data=outlog (obs=5) noobs;
   var pd student_loan divorce head MEff:;
run;
proc means data=outlog mean min max;
   var MEff:;
run;
**odds;
ods graphics / imagemap=on ;
proc logistic data=WORK.Assessment2_data_withhead2 ;
model financial_distress (Event = '1')= pd student_loan divorce/ expb;
output  out=ccdyhat  prob=PredictedValues;
run;
*model fit*;
proc logistic data=WORK.Assessment2_data_withhead2 plots(only)=(roc);
model financial_distress (Event = '1')= pd student_loan divorce/
link=probit ;
run;


*model 2*;
ods graphics / imagemap=on ;
proc probit data=WORK.Assessment2_data_withhead2;
model financial_distress (Event = '1')= pd student_loan divorce student_loan*divorce/
d=logistic ;
output  out=ccdyhat  prob=PredictedValues ;
run;
ods graphics / imagemap=on;
proc gplot data=ccdyhat;
plot financial_distress*PredictedValues PredictedValues*pd  PredictedValues*student_loan  ;
run;

*Meff*;
proc logistic data=WORK.Assessment2_data_withhead2 outest=logparms(rename=(pd=tpd
student_loan=tstudent_loan divorce=tdivorce));
   model financial_distress(event="1") = pd student_loan divorce student_loan*divorce;
   output out=outlog p=p;
run;
data outlog;
   if _n_ = 1 then set logparms;
   set outlog;
   MEffpd = p * (1 - p) * tpd;
   MEffStudentLoan = p * (1 - p) * tstudent_loan;
   MEffDivorce = p * (1 - p) * tdivorce;
   MEffStudentLoanDivorce = p * (1 - p) * (student_loan * divorce); /* Calculate the interaction term effect */
run;
proc print data=outlog(obs=5) noobs;
   var pd student_loan divorce MEffpd MEffStudentLoan MEffDivorce MEffStudentLoanDivorce;
run;
proc means data=outlog mean min max;
   var MEffpd MEffStudentLoan MEffDivorce MEffStudentLoanDivorce;
run;

**odds;
ods graphics / imagemap=on ;
```

```
proc logistic data=WORK.Assessment2_data_withhead2 ;
model financial_distress (Event = '1')= pd student_loan divorce student_loan*divorce/ expb;
output  out=ccdyhat  prob=PredictedValues;
run;
*model fit*;
proc logistic data=WORK.Assessment2_data_withhead2 plots(only)=(roc);
model financial_distress (Event = '1')= pd student_loan divorce student_loan * divorce/
link=probit ;
run;


*model 3*;
/* Stepwise Logistic Regression */
proc logistic data=WORK.Assessment2_data_withhead2;
   model financial_distress(Event='1') = student_loan divorce socio_economic hispanic marriage pd laidoff black
collegedegree missedwork familydeath childbirth white male age
   / selection=stepwise slentry=0.05 slstay=0.05;
   output out=ccdyhat prob=PredictedValues;
run;
/* Calculation of Effective Sample Size (Meff) */
proc logistic data=WORK.Assessment2_data_withhead2 outest=logparms(rename=(
   student_loan=tstudent_loan
   divorce=tdivorce
   socio_economic=tsocio_economic
   pd=tpd
   laidoff=tlaidoff
   collegedegree=tcollegedegree
   missedwork=tmissedwork
   male=tmale
   age=tage
));
   model financial_distress(event="1") = student_loan divorce socio_economic  pd laidoff collegedegree missedwork
male age;
   output out=outlog p=p;
run;
data outlog;
   if _n_ = 1 then set logparms;
   set outlog;
   MEffStudentLoan = p * (1 - p) * tstudent_loan;
   MEffDivorce = p * (1 - p) * tdivorce;
   MEffSocioEconomic = p * (1 - p) * tsocio_economic;
   MEffPd = p * (1 - p) * tpd;
   MEffLaidoff = p * (1 - p) * tlaidoff;
   MEffCollegedegree = p * (1 - p) * tcollegedegree;
   MEffMissedwork = p * (1 - p) * tmissedwork;
   MEffMale = p * (1 - p) * tmale;
   MEffAge = p * (1 - p) * tage;
run;
/* Print a subset of the data for verification */
proc print data=outlog(obs=5) noobs;
   var student_loan divorce socio_economic  pd laidoff collegedegree missedwork male age
      MEffStudentLoan MEffDivorce MEffSocioEconomic  MEffPd MEffLaidoff  MEffCollegedegree MEffMissedwork
MEffMale MEffAge;
run;
```

```
/* Summary statistics for Meff variables */
proc means data=outlog mean min max;
   var MEffStudentLoan MEffDivorce MEffSocioEconomic  MEffPd MEffLaidoff  MEffCollegedegree MEffMissedwork
MEffMale MEffAge;
run;

proc logistic data=WORK.Assessment2_data_withhead2 plots(only)=(roc);
   model financial_distress(Event='1') = student_loan divorce socio_economic hispanic marriage pd laidoff black
collegedegree missedwork familydeath childbirth white male age
   / selection=stepwise slentry=0.05 slstay=0.05
   link=probit;
run;


*model 4*;
/* Stepwise Logistic Regression */
ods graphics / imagemap=on;
proc logistic data=WORK.Assessment2_data_withhead2 plots=roc;
   model financial_distress(Event='1') = divorce socio_economic hispanic marriage pd laidoff black collegedegree
missedwork familydeath childbirth white male age
   / selection=stepwise slentry=0.05 slstay=0.05;
   output out=ccdyhat p=PredictedValues;
run;


*5. panel data analysis*;
*have new data set with student_loan,as still want to take student_loan into consideration. This is also good for the
limit memory of my computer*;
data Assessment2_data_withhead3;
   set WORK.Assessment2_data_withhead2;
   where student_loan is not missing;
run;

*fixed effects data analysis*;
ods graphics / imagemap=on;
proc sort data=WORK.Assessment2_data_withhead3;
   by id year;
run;

proc panel data=WORK.Assessment2_data_withhead3 (where=(student_loan~=. and divorce~=. and
socio_economic~=. and hispanic~=. and marriage~=. and pd~=. and laidoff~=. and black~=. and collegedegree~=. and
missedwork~=. and familydeath~=. and childbirth~=. and familydeath~=. and white~=. and male~=. and age~=. and
financial_distress~=.));
   id id year;
   model financial_distress = student_loan divorce socio_economic hispanic marriage pd laidoff black collegedegree
missedwork familydeath childbirth white male age / fixone ;
run;

proc panel data=WORK.Assessment2_data_withhead3 (where=(student_loan~=. and divorce~=. and
socio_economic~=. and hispanic~=. and marriage~=. and pd~=. and laidoff~=. and black~=. and collegedegree~=. and
missedwork~=. and familydeath~=. and childbirth~=. and familydeath~=. and white~=. and male~=. and age~=. and
financial_distress~=.));
   id id year;
```

```
  model financial_distress = student_loan divorce socio_economic hispanic marriage pd laidoff black collegedegree
missedwork familydeath childbirth white male age / fixtwo ;
run;

*fixed time data analysis*;
proc panel data=WORK.Assessment2_data_withhead3 (where=(student_loan~=. and divorce=~. and
socio_economic~=. and hispanic~=. and marriage~=. and pd~=. and laidoff~=. and black~=. and collegedegree~=. and
missedwork~=. and familydeath~=. and childbirth~=. and familydeath~=. and white~=. and male~=. and age~=. and
financial_distress~=.));
  id id year;
  model financial_distress = student_loan divorce socio_economic hispanic marriage pd laidoff black collegedegree
missedwork familydeath childbirth white male age / fixonetime ;
run;

*random effects data analysis*;
proc panel data=WORK.Assessment2_data_withhead3 (where=(student_loan~=. and divorce=~. and
socio_economic~=. and hispanic~=. and marriage~=. and pd~=. and laidoff~=. and black~=. and collegedegree~=. and
missedwork~=. and familydeath~=. and childbirth~=. and familydeath~=. and white~=. and male~=. and age~=. and
financial_distress~=.));
  id id year;
  model financial_distress = student_loan divorce socio_economic hispanic marriage pd laidoff black collegedegree
missedwork familydeath childbirth white male age / ranone ;
run;

proc panel data=WORK.Assessment2_data_withhead3 (where=(student_loan~=. and divorce=~. and
socio_economic~=. and hispanic~=. and marriage~=. and pd~=. and laidoff~=. and black~=. and collegedegree~=. and
missedwork~=. and familydeath~=. and childbirth~=. and familydeath~=. and white~=. and male~=. and age~=. and
financial_distress~=.));
  id id year;
  model financial_distress = student_loan divorce socio_economic hispanic marriage pd laidoff black collegedegree
missedwork familydeath childbirth white male age / rantwo ;
run;
```