Applied Econometrics Methods

Investigating net worth dynamics through ordinary least square regression analysis

Peng Li 07/05/2024

Contents

1. INTRODUCTION

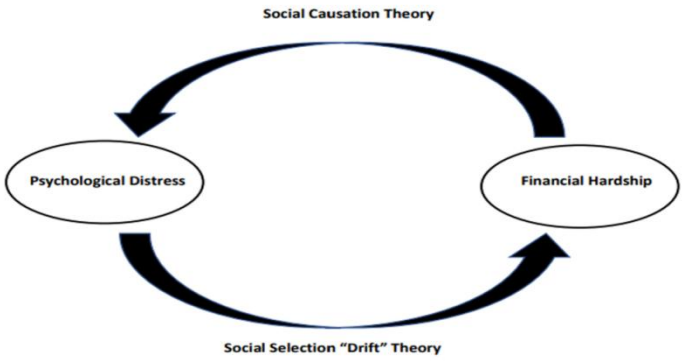
1.1 Main target

In this project we will take "wealth"as net worth (wealth) as dependent variable, psychological distress (pd) as key independent variable and others as control variables.

We will describe the raw data through tables and graphs, providing initial insights of the data. We will select the appropriate regression models such like linear regression and multiple regression. After running the regressions, we will provide technical and logical interpretation of the parameters. We will also conduct CLRM tests and discuss the possible reasons when a test fails.

1.2 Theoretical Framework

Many researchers generally believe that inequalities in wealth are strongly associated with psychological distress [1], and the stress of navigating financial hardship increases the risk of subsequent psychological distress, a key tenet of social causation theory[2]. Other researchers have proposed social drift theory, which posits that a decline in mental health increases the risk of subsequent financial hardship[2], therefore there is a negative effect of poor mental health on individuals wealth[3]. Few researchers examine the relationship from the perspective of social drift theory, we will have more study from this side.



1.3 Variables description

Detailed Variables description please see below:

Variable name (variable name in data set)	Variable description	Variable type
Dependent variable		
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
Socio-economic status (socioeconomic)	Captures the respondents' socio-economic status, where higher score means higher socio-economic status	Continuous
Year (Time ID variable)	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
Key (Cross sectional ID variable)	Captures the unique identifier of the respondents.	Categorical

2. DATA OVERALL STUDY

2.1 Overall data status

- To show the variables' type, length, format etc.

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
5	age	Num	8	15		age
8	black	Num	8	15		black
16	childbirth	Num	8	15		childbirth
14	divorce	Num	8	15		divorce
11	education	Num	8	15		education
13	employed	Num	8	15		employed
17	familydeath	Num	8	15		familydeath
9	hispanic	Num	8	15		hispanic
12	income	Num	8	15		income
1	key	Char	8	$8.00	$8.00	key
18	laidoff	Num	8	15		laidoff
6	male	Num	8	15		male
15	marriage	Num	8	15		marriage
19	missedwork	Num	8	15		missedwork
10	otherrace	Num	8	15		otherrace
4	pd	Num	8	15		pd
20	socioeconomic	Num	8	15		socioeconomic
2	time	Num	8	15		time
3	wealth	Num	8	15		wealth
7	white	Num	8	15		white

Except the variable “key”, all the other variables are number type.

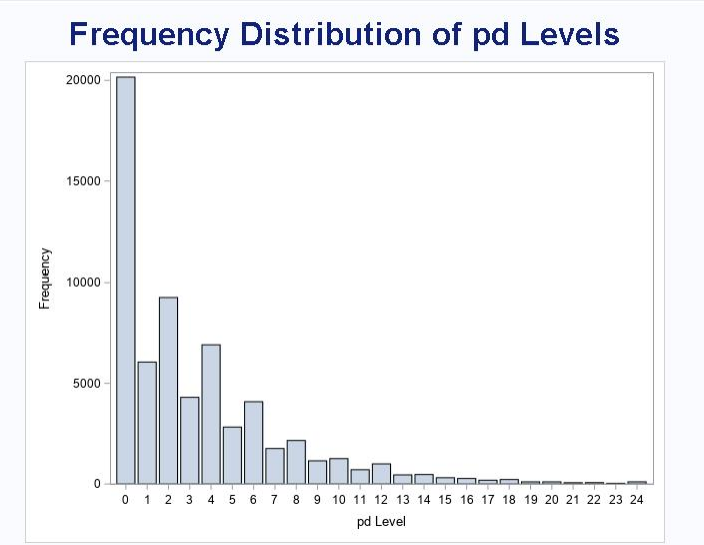
- To show the variables’ amount, mean, standard deviation, minimum, maximum.

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
time	time	65266	4.105997	2.5453903	0	8
wealth	wealth	65266	259227.66	1203131.59	-3197000	100555000
pd	pd	64054	3.3589003	3.9838327	0	24
age	age	65258	45.6783383	16.1038715	16	101
male	male	65266	0.6961052	0.4599413	0	1
white	white	65262	0.5845515	0.492803	0	1
black	black	65262	0.3524716	0.4777435	0	1
hispanic	hispanic	65262	0.0098219	0.0986186	0	1
otherrace	otherrace	65262	0.053155	0.2243441	0	1
education	education	63370	13.1717216	2.6441235	0	17
income	income	64738	10.6636751	1.0788477	0	16
employed	employed	65266	0.7161309	0.4508775	0	1
divorce	divorce	65266	0.0511599	0.2203254	0	1
marriage	marriage	65266	0.0817731	0.2740207	0	1
childbirth	childbirth	65266	0.1206754	0.3257522	0	1
familydeath	familydeath	65266	0.0205007	0.1417066	0	1
laidoff	laidoff	65265	0.05018	0.218318	0	1
missedwork	missedwork	65266	1.072963	3.6301397	0	78
socioeconomic	socioeconomic	65228	40868.58	97293.89	-96352	6278577

2.2 Data overall study findings:

- For the amount of different variables, we will see that for some variables, part of their values are missing. We will deal with this during data wrangling part, to delete the related observations with missing values based on the variables we want to check.

- For the dependent variable “wealth” we want to study, we see big difference between minimum and maximum, from -3,197,000 to 100,555,000. We will check them more in details to understand its their distribution in next part. The mean is 259 ,227 which is positive number.
- For the key independent variable “pd” psychological distress, combined with the original excel data sheet, we see pd has 25 levels from 0 to 24. This variable has high data quality, as it has detailed level information.



- For other control variables, nearly all of them are binomial data 0 or 1, except “age” from 16 to 99, “education” level 0 to 17, “income” level 0 to 16, “misswork” from 0 to 78 weeks of “work missed due to illness”, and also “socioeconomic” from -96,352 to 6,278,577 which also has big range.
- If we check the binomial data in detail, we can find that in the statistic around 69% are male, 58% are white and 35% are black, 71% employed, 5% recently experienced divorce, 8% recently got married, 12% a household member recently gave birth, 2% a household member recently died, 2% recently laid off from work etc. All these give us an overall impression about the data distribution and status.
- Combined with the general data status and variable descriptions, we can generally divide them into following groups for better understanding:

  -> Independent variable: wealth, we want to research;

  -> Key depend variable: pd psychological distress, we will focus on its influence on wealth;

  -> Time variable: time, 0 to 8, represents 2003 to 2019;

  -> Physiological variable: age (16 to 99), male/female;

  -> Ethnicity variable: white, black, hispanic and otherrace;

  -> Education: 0 to 17;

  -> Income: 0 to 16;

-> Family variable: divorce, marriage, childbirth, familydeath;

-> Employment variable: employed, laidoff, missedwork(0 to 78);

-> Socioeconomic: -96,352 to 6,278,577;

- If we check the average "pd" level, we can find that "female" has overall higher "pd" level than "male". The reason is that female are more emotional. And from "wealth" side "male" has overall higher "wealth" than "female" with around 3 times.

male=0

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
pd	pd	19239	4.1972036	4.5003368	0	24.0000000
wealth	wealth	19834	102214.62	562409.41	-990023.00	50475000.00

male=1

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
pd	pd	44815	2.9990182	3.6822394	0	24.0000000
wealth	wealth	45432	327773.99	1387779.86	-3197000.00	100555000

3. Primary DATA Wrangling

3.1 Delete the observations with missing values

As we find in data overall study, some "pd" values are missing, so we will exclude these missing observations first. After that we get below the "means" for all variables.

The MEANS Procedure

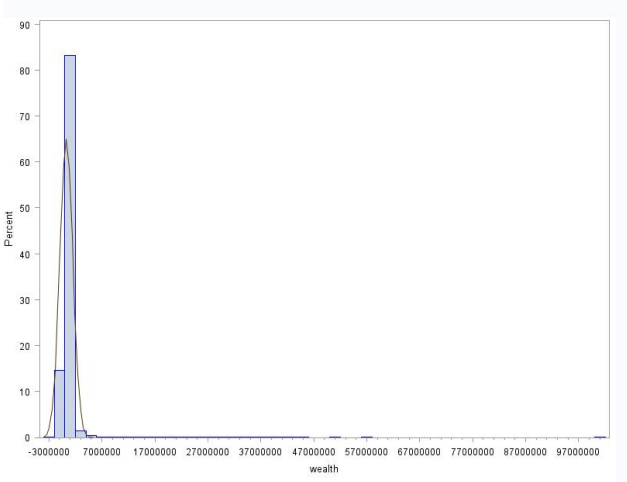
Variable	Label	N	Mean	Std Dev	Minimum	Maximum
time	time	64054	4.1111874	2.5484490	0	8.0000000
wealth	wealth	64054	261408.80	1213088.54	-3197000.00	100555000
pd	pd	64054	3.3589003	3.9838327	0	24.0000000
age	age	64047	45.4069824	15.8500017	16.0000000	99.0000000
male	male	64054	0.6996441	0.4584163	0	1.0000000
white	white	64050	0.5863232	0.4924958	0	1.0000000
black	black	64050	0.3519906	0.4775948	0	1.0000000
hispanic	hispanic	64050	0.0098205	0.0986112	0	1.0000000
otherrace	otherrace	64050	0.0518657	0.2217576	0	1.0000000
education	education	62254	13.2054968	2.6082897	0	17.0000000
income	income	63550	10.6738788	1.0757413	0	16.0000000
employed	employed	64054	0.7223905	0.4478232	0	1.0000000
divorce	divorce	64054	0.0513629	0.2207386	0	1.0000000
marriage	marriage	64054	0.0824148	0.2749978	0	1.0000000
childbirth	childbirth	64054	0.1213976	0.3265913	0	1.0000000
familydeath	familydeath	64054	0.0197490	0.1391375	0	1.0000000
laidoff	laidoff	64053	0.0505519	0.2190825	0	1.0000000
missedwork	missedwork	64054	1.0798389	3.6365866	0	78.0000000
socioeconomic	socioeconomic	64016	41279.55	97888.80	-96352.00	6278577.00

After the first step wrangling, we find that some variables are still missing some values. "education" is missing around 1800 values, while "income" is missing 500 values. Compared to overall around 64,000 observations, 1,800 observations takes 2.8% and 500 observations takes 0.8%; "education" and "income" are quite related the "wealth" we will study, so we would like to prepare these data also for analysis.

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
time	time	61714	4.1535308	2.5425273	0	8.0000000
wealth	wealth	61714	264789.99	1227526.01	-3197000.00	100555000
pd	pd	61714	3.3425479	3.9615627	0	24.0000000
age	age	61714	45.4985417	15.8664134	16.0000000	99.0000000
male	male	61714	0.6994199	0.4585140	0	1.0000000
white	white	61714	0.5921833	0.4914328	0	1.0000000
black	black	61714	0.3512169	0.4773545	0	1.0000000
hispanic	hispanic	61714	0.0097547	0.0982837	0	1.0000000
otherrace	otherrace	61714	0.0468451	0.2113087	0	1.0000000
education	education	61714	13.2187024	2.6065646	0	17.0000000
income	income	61714	10.6737693	1.0765508	0	16.0000000
employed	employed	61714	0.7255404	0.4462452	0	1.0000000
divorce	divorce	61714	0.0519007	0.2218284	0	1.0000000
marriage	marriage	61714	0.0815212	0.2736361	0	1.0000000
childbirth	childbirth	61714	0.1197297	0.3246478	0	1.0000000
familydeath	familydeath	61714	0.0199955	0.1399856	0	1.0000000
laidoff	laidoff	61714	0.0501021	0.2181574	0	1.0000000
missedwork	missedwork	61714	1.0900768	3.6582632	0	78.0000000
socioeconomic	socioeconomic	61714	41716.36	98683.67	-96352.00	6278577.00

Show the distribution of “wealth ” to have the overall impression of the values:

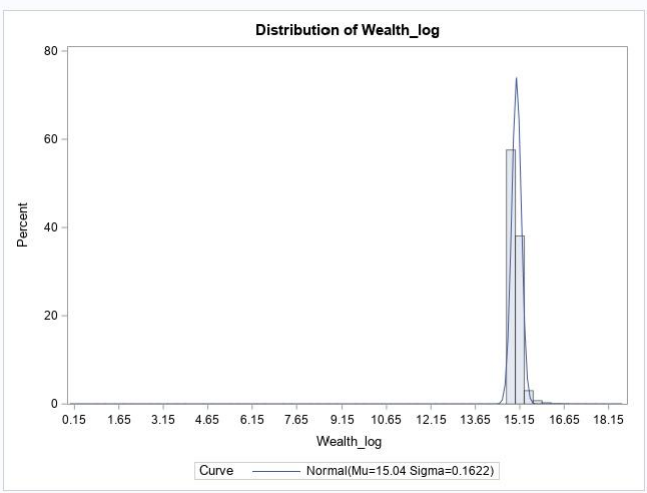


Quantiles (Definition 5)	
Level	Quantile
100% Max	100555000
99%	3220000
95%	1091600
90%	600065
75% Q3	202000
50% Median	43000
25% Q1	2000
10%	-7000
5%	-27900
1%	-102700
0% Min	-3197000

Through above diagram we will find that the top 1% richest peoples is very discrete. So we decide to use log for wealth to have better distribution.

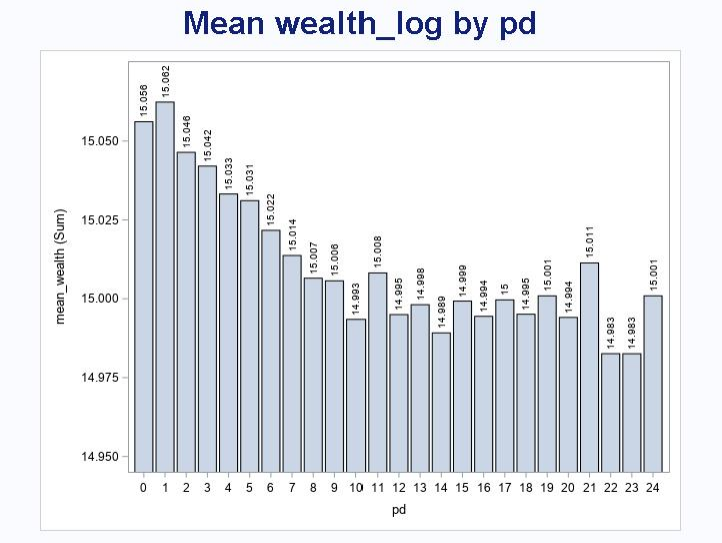
3.2 Apply log for “wealth” data

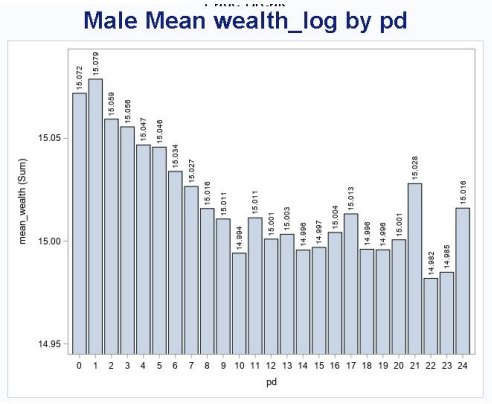
Because there is negative “wealth” values, so we will not directly apply log. The lowest value is -319700, so we transfer all the values in to positive with plus 3197000+1, then to apply log to have the “wealth_log” distribution.

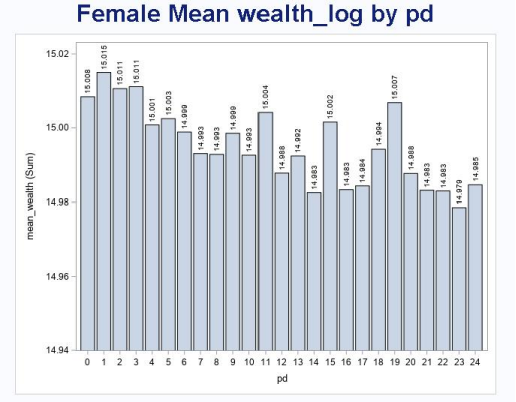


We see that after using log to transform the "wealth" data, the data has much better distribution.

Then we draft the below chart to show the distribution of "wealth_log" according to different "pd" level.







This is more clear to show the association of "wealth_log" with "pd". In general lower "pd" has higher "wealth_log". And rich people almost locate in the low "pd" level area. The right psychological level or adjustment will help people to accumulate more wealth, while on the other side, high psychological

distress cases more problem to people and also reduces the "wealth". Of course higher "wealth" also ensure lower "pd", because there maybe less worries for life cost etc. They influence each other, so we keep this good interaction. If we do not have "wealth", we at least need to avoid too much psychological distress and stay in good mode for future wealth.

4. Simple linear regression through ordinary least square

Firstly we use simple linear regression to study the "pd" influence on "wealth" directly.

Simple regression

We just take the variable "pd" for analysis:

Number of Observations Read	61714
Number of Observations Used	61714

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	21.77210	21.77210	839.31	<.0001
Error	61712	1600.83814	0.02594		
Corrected Total	61713	1622.61024			

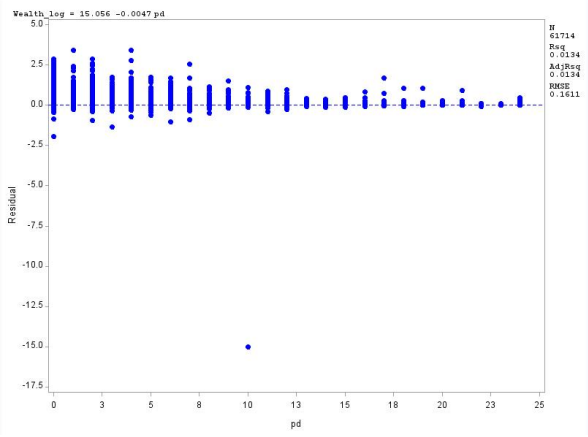
Root MSE	0.16106	R-Square	0.0134
Dependent Mean	15.03993	Adj R-Sq	0.0134
Coeff Var	1.07089		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	15.05578	0.00084828	17748.6	<.0001
pd	pd	1	-0.00474	0.00016366	-28.97	<.0001

- The R-Square is 0.0134 which means this model only explain 1.34% of the variation. The probability of the model is less than 0.0001 and this means there is less than 0.01% chance that the parameter coefficients are zero, hence, the parameter estimates do have an impact on the price. For "pd", it has negative contribution to wealth. The "pd" level increase 1 leads to 0.00474 unit decrease of "wealth_log".

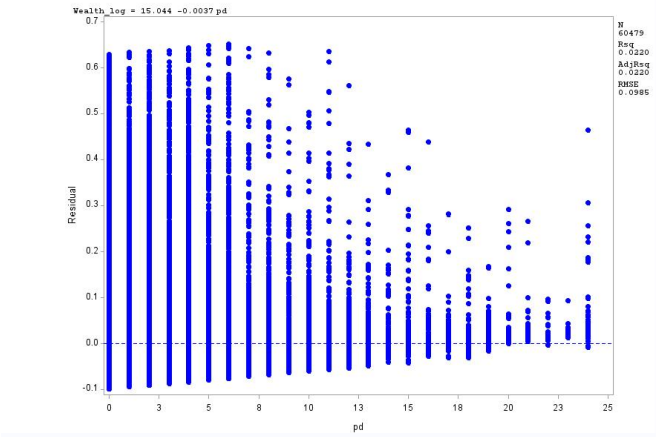
The least squares residuals

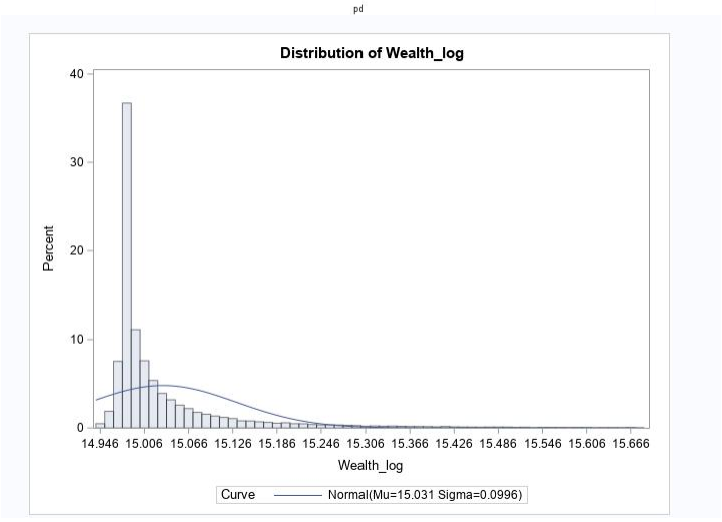


We can find that there are still many discrete dots up and down. So we will check more details of the “log wealth” for further wrangling.

Quantiles (Definition 5)	
Level	Quantile
100% Max	18.4575
99%	15.6745
95%	15.2715
90%	15.1497
75% Q3	15.0390
50% Median	14.9911
25% Q1	14.9783
10%	14.9755
5%	14.9690
1%	14.9451
0% Min	0.0000

Noticed that there is big variance of the top 1% richest and lowest 1%, so we will delete the two parts for the log transformation, to eliminate the discrete dots in the “the least squares residuals” chart. Below chart is is the new distribution of residual.





Model 1: Simple regression with updated “wealth_log” data

With the adjusted “”wealth_log” date we still just take the variable “pd” for analysis:

Number of Observations Read	60479
Number of Observations Used	60479

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13.18686	13.18686	1358.75	<.0001
Error	60477	586.93974	0.00971		
Corrected Total	60478	600.12661			

Root MSE	0.09851	R-Square	0.0220
Dependent Mean	15.03113	Adj R-Sq	0.0220
Coeff Var	0.65541		

Parameter Estimates

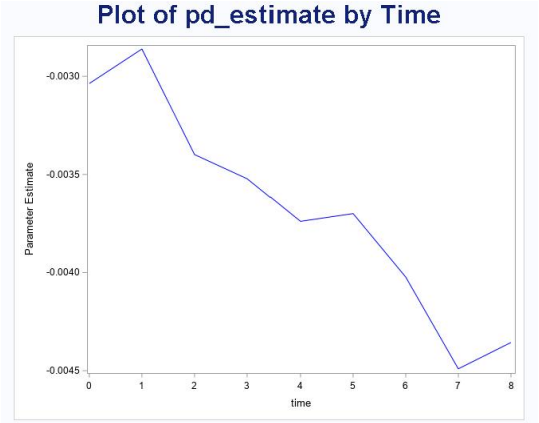
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	15.04359	0.00052415	28701.0	<.0001
pd	pd	1	-0.00372	0.00010096	-36.86	<.0001

We can find that the R-Square has improved from 0.0134 in first linear regression to 0.022 in this second linear regression after further data wrangling. The "pd" increase 1 level will lead to 0.00372 unit decrease of "wealth_log". The estimated regression equation to Model 4 is as follows:

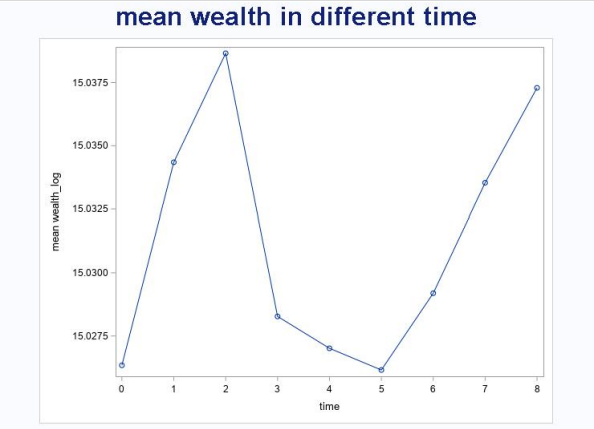
Log(wealth) = 15.04359– 0.00372pd + μi

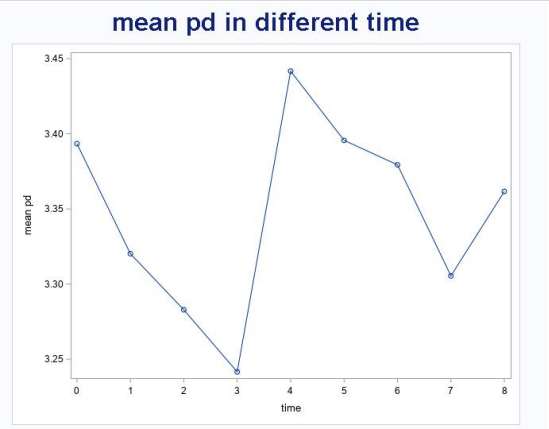
Model 2: Study the "pd" coefficient in different time

According to the time from 0 to 8, we divide the original data set into 9 time groups, to get the different coefficient of "pd" in different time based on model 1. Then we make below chart to see the "pd" coefficient value trend according to the time.



Through this we can find that the influence of "pd" on "wealth_log" is increasing year after year. At the beginning we see the slight decrease of the absolute value of the "pd" coefficient in 2005 (time 1) compared to 2003 (time 0). But then a big curve change happened in year 2007 (from time 1 to time 2) . Global financial crisis could be one reason behind it. After that the trend kept going down and seems that there is some change in the year 2019 (time 8), but that is already 12 years after 2007 (time 2). From the time side, we see that the influence of "pd" on "wealth_log" in 2017 (time 7 with around -0.0045) is around 50% higher than in 2003 (time 0 with around -0.0030). So that means recently the "pd" plays a more important role on "wealth_log" or "wealth".

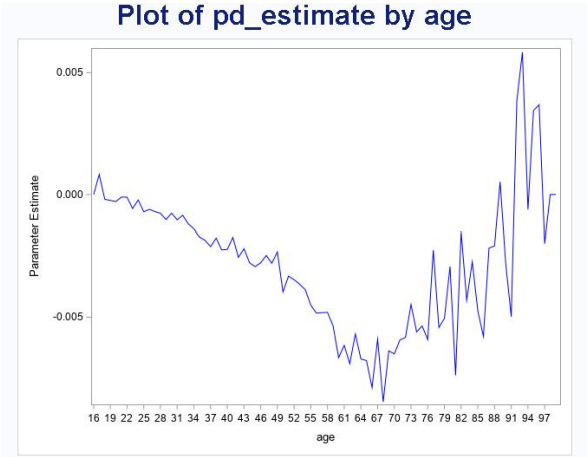




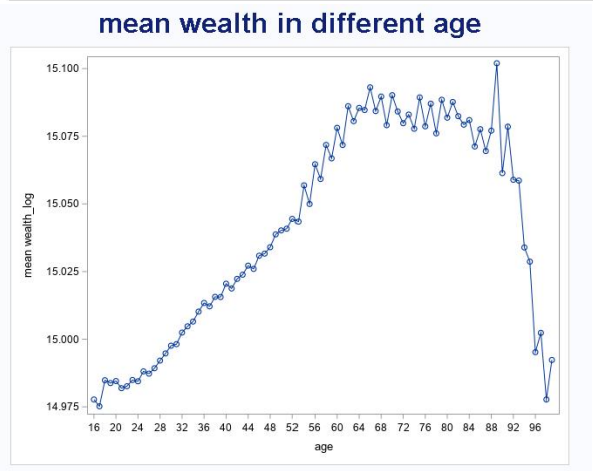
If we further check the average of "wealth_log" and "pd" in different times, we also see some changes around financial crisis time. And seems that they have opposite trends with some time gap. The "pd" variation is not so big between 3.25 to 3.45, but on the contrary, we see that the "ph" influence on wealth is overall becoming bigger, and in 2017 the coefficient is around 50% bigger than 2003. So the average of "wealth_log and "pd" supports the conclusion that, no matter the overall situation of wealth and "pd" in different time, people should regulate their "pd" level to have less influence on the wealth, as its influence on wealth is becoming bigger year after year in general.

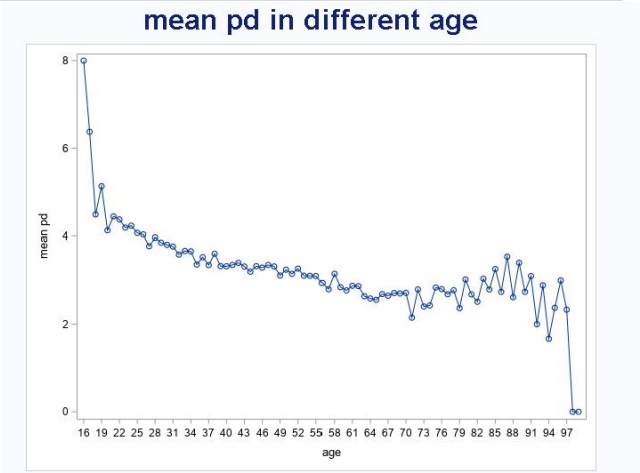
Model 3: Study the pd coefficient in different age

According to the age from 16 to 97, we divide the original data set into 82 age groups, to get the different coefficients of "pd" based on model 1 in different age group. Then we make below chart to see the "pd" coefficient value trend according to the age.



It is interesting to find that the curve changes at around 65 year of the "age", which is the general retirement year. After retirement it seems that the "pd" has smaller influence on "wealth". I think another hiding factor is "income", but after retirement if people can get the pension, the "income" will not change much. Retirement is also a big change of each person's status, from daily working to non working. At this time, even their "pd" increases, this will not influence so much on the "wealth", because the income which means pension is very stable and they also have enough time to handle the "pd" increase, which will not leads to bad results such like health problem or relationship problem/family problem which will also has negative influence on wealth. So in general their "wealth" will be relative stable during this secure situation even the "pd" increase.





Through above charts (mean of wealth_log and mean of pd in different age) we can find that after around 65 years old, the wealth_log is nearly stable and after around 88 it decreased suddenly, which may because of heavy disease that much money should be spent in short time. And the "pd" level seems to keep on decreasing along the age, especially before the retirement.

If we combine above three charts together, we should say, before the retirement people should pay high attention to the "pd" level, as its influence on "wealth" is becoming bigger along the age increasing.

5. Multiple regression analysis through ordinary least square

Different from simple linear regression to have more analysis of "pd" from "time" and "age" sides, in this part we will study more from the gender side and also other control variables' sides.

Correlation matrix

	pd	age	male	white	black	hispanic	otherrace	education	income	employed	divorce	marriage	childbirth	family death	laidoff	missed work	socioeconomic	Wealth_log
pd	1	-0.11	-0.14	-0.06	0.07	-0.01	-0.01	-0.13	-0.22	-0.13	0.05	0.03	-0.02	0.04	0.03	0.03	-0.13	-0.15
age	-0.11	1	-0.03	0.13	-0.12	-0.02	0	-0.06	0.08	-0.38	-0.11	-0.21	-0.25	0.11	-0.07	-0.03	0.03	0.32
male	-0.14	-0.03	1	0.25	-0.28	0.04	0.03	0.08	0.38	0.16	-0.04	0.11	0.15	-0.06	0.03	-0.04	0.22	0.19
white	-0.06	0.13	0.25	1	-0.89	-0.12	-0.26	0.21	0.29	0.08	-0.03	0.01	0	0	-0.01	-0.01	0.24	0.27
black	0.07	-0.12	-0.28	-0.89	1	-0.07	-0.16	-0.16	-0.3	-0.09	0.03	-0.01	-0.02	0	0	0.01	-0.24	-0.28
hispanic	-0.01	-0.02	0.04	-0.12	-0.07	1	-0.02	-0.12	-0.03	0.02	-0.01	-0.01	0.05	0	0.04	0	-0.03	-0.03
otherrace	-0.01	0	0.03	-0.26	-0.16	-0.02	1	-0.05	0.01	0.01	0	0	0.02	0.01	0.01	-0.01	-0.01	0
education	-0.13	-0.06	0.08	0.21	-0.16	-0.12	-0.05	1	0.37	0.21	-0.02	0	0	-0.04	-0.09	0	0.32	0.28
income	-0.22	0.08	0.38	0.29	-0.3	-0.03	0.01	0.37	1	0.35	-0.03	-0.01	0.06	-0.02	-0.04	0.03	0.46	0.38
employed	-0.13	-0.38	0.16	0.08	-0.09	0.02	0.01	0.21	0.35	1	0.01	0.06	0.1	-0.07	0.03	0.1	0.19	-0.02
divorce	0.05	-0.11	-0.04	-0.03	0.03	-0.01	0	-0.02	-0.03	0.01	1	0.13	-0.02	-0.02	0.02	0.02	-0.04	-0.08
marriage	0.03	-0.21	0.11	0.01	-0.01	-0.01	0	0	-0.01	0.06	0.13	1	0.1	-0.01	0.03	0	0.02	-0.07
childbirth	-0.02	-0.25	0.15	0	-0.02	0.05	0.02	0	0.06	0.1	-0.02	0.1	1	-0.03	0.03	-0.01	-0.01	-0.08
familydeath	0.04	0.11	-0.06	0	0	0	0.01	-0.04	-0.02	-0.07	-0.02	-0.01	-0.03	1	-0.01	0.01	-0.03	0.02
laidoff	0.03	-0.07	0.03	-0.01	0	0.04	0.01	-0.09	-0.04	0.03	0.02	0.03	0.03	-0.01	1	0	-0.05	-0.06
missedwork	0.03	-0.03	-0.04	-0.01	0.01	0	-0.01	0	0.03	0.1	0.02	0	-0.01	0.01	0	1	-0.01	-0.03
socioeconomic	-0.13	0.03	0.22	0.24	-0.24	-0.03	-0.01	0.32	0.46	0.19	-0.04	0.02	-0.01	-0.03	-0.05	-0.01	1	0.38
Wealth_log	-0.15	0.32	0.19	0.27	-0.28	-0.03	0	0.28	0.38	-0.02	-0.08	-0.07	-0.08	0.02	-0.06	-0.03	0.38	1

Through the correlation analysis, we see that "wealth_log" has some correlation with "age" "income" and "socioeconomic". I think this is because "wealth" is accumulated indicator, so "income" by "age"(means

the years to accumulate the bigger wealth) and also "Socioeconomic" are most related with "wealth_log". "age" is correlated with "employed" with 0.38 and "male" is correlated with "income" with 0.38, which is reasonable. "Education" is somehow correlated to "income" and similar situation for "employed" and "income". Besides these, we see "white" is quite correlated with "black", as they belong to the same category, but not so correlated with "hispanic" which is about nationality.

For "time" and "age", we already studied it in model 2 and model 3, so we will study influence of gender side "male" for more understanding.

Model 4: From the gender side to study psychological distress on change in wealth

In this model we only take "pd" and "male " as independent variables for analysis.

Number of Observations Read	60479
Number of Observations Used	60479

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	31.23610	15.61805	1660.28	<.0001
Error	60476	568.89050	0.00941		
Corrected Total	60478	600.12661			

Root MSE	0.09699	R-Square	0.0520
Dependent Mean	15.03113	Adj R-Sq	0.0520
Coeff Var	0.64525		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	15.01506	0.00083096	18069.5	<.0001
pd	pd	1	-0.00311	0.00010037	-30.96	<.0001
male	male	1	0.03797	0.00086686	43.80	<.0001

Overall, the probability of the model is less than 0.0001 and this means there is less than 0.01% chance that the parameter coefficients are zero, hence, the parameter estimates do have an impact on the price. The adjusted R-square of 0.052 which is higher than model 1 with 0.022, but the predictive variables still only explain 5.2% of the variation. The adjusted R-square value is extremely low and further action needs to be taken to improve the Model.

The estimated regression equation to Model 4 is as follows:

Log(wealth) = 15.01506– 0.00311pd + 0.03797male + μi

The coefficient of the "male" is 0.03797 and its probability was less than 0.01%. This means for same level "pd" on average the "male" has 0.03797 units higher "wealth_log" than the average "female" and their values are significantly different. And the coefficient of the "pd" is -0.00311 and its probability was less than 0.01%. This means on average for male and female, one unit increase of "pd" will lead to 0.00311 unit decrease of "wealth_log".

Model 5: Add the new parameter pd*male for more study from gender side

Based on model 4, we added pd*male as a new variable for helping the study of "pd" influence on "wealth_log" from gender side.

Number of Observations Read	60479
Number of Observations Used	60479

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	32.99949	10.99983	1172.96	<.0001
Error	60475	567.12711	0.00938		
Corrected Total	60478	600.12661			

Root MSE	0.09684	R-Square	0.0550
Dependent Mean	15.03113	Adj R-Sq	0.0549
Coeff Var	0.64426		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	15.00795	0.00097814	15343.3	<.0001
pd	pd	1	-0.00141	0.00015919	-8.87	<.0001
male	male	1	0.04840	0.00115	42.01	<.0001
pd_male		1	-0.00281	0.00020489	-13.71	<.0001

The adjusted R-square of 0.0549 which is a little higher than model 4. The predictive variables only explain 5.49% of the variation. The adjusted R-square value is still low and further action needs to be taken to improve the Model.

The estimated regression equation to Model 5 is as follows:

Log(wealth) = 15.00795– 0.00141pd + 0.04840male -0.00281(pd*male) + μi

From the equation we can find that for different "pd" level, the difference of "male" and "female" influence on "wealth" is different. Because we add the new variable "pd*male", so the difference of "male" and "female" influence on "wealth" is not just related to the coefficient of "male" in the equation but also related to the coefficient of "pd*male" (in another word, also related to "pd" level). For "pd" level 0, the difference of "male" and "female" influence on "wealth" is 0.04840 units. When the "pd" level reach 17 (0.04840/0.00281=17.22), then the difference of "male" and "female" influence on "wealth_log" is nearly zero with 0.04840-0.00281*1*17=0.00063 units. Then when "pd" level is higher than 17, then under the same "pd" level, the influence of "male" on "wealth_log" is lower than the influence of "female" on "wealth_log".

In model 5, on average, for "male"one unit increase of "pd" level will leads to -0.00141-0.00281= -0.00422 units decrease of "wealth_log", while for "female" one unit increase of "pd" level will leads to -0.00141 units decrease of "wealth_log". So on average compared to "female", for "male" the "pd" level has higher negative influence on "wealth_log".

Model 6: Add all other control variables for more study

(for "white" and "black", we only only chose "white", as they are very correlated with 00.89)

Based on model 4, we added all other control variables excluding "black" for improving the model.

Source	DF	Squares	Square	F Value	Pr > F
Model	17	192.77368	11.33963	1683.07	<.0001
Error	60461	407.35293	0.00674		
Corrected Total	60478	600.12661			

Root MSE	0.08208	R-Square	0.3212
Dependent Mean	15.03113	Adj R-Sq	0.3210
Coeff Var	0.54608		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	14.69088	0.00413	3553.92	<.0001
pd	pd	1	-0.00059802	0.00008799	-6.80	<.0001
time	time	1	-0.00048571	0.00013391	-3.63	0.0003
male	male	1	0.01240	0.00082276	15.07	<.0001
age	age	1	0.00164	0.00002547	64.44	<.0001
white	white	1	0.02213	0.00077549	28.54	<.0001
hispanic	hispanic	1	0.02276	0.00346	6.58	<.0001
otherrace	otherrace	1	0.01631	0.00166	9.83	<.0001
education	education	1	0.00563	0.00014542	38.74	<.0001
income	income	1	0.01644	0.00041780	39.36	<.0001
employed	employed	1	-0.01399	0.00090304	-15.49	<.0001
divorce	divorce	1	-0.00997	0.00153	-6.54	<.0001
marriage	marriage	1	-0.00443	0.00127	-3.50	0.0005
childbirth	childbirth	1	-0.00799	0.00108	-7.42	<.0001
familydeath	familydeath	1	0.00575	0.00241	2.38	0.0173
laidoff	laidoff	1	-0.00424	0.00154	-2.76	0.0059
missedwork	missedwork	1	-0.00054837	0.00009222	-5.95	<.0001
socioeconomic	socioeconomic	1	2.611591E-7	4.857326E-9	53.77	<.0001

The adjusted R-square of 0.3210 which is much higher than models 1 to 5. The predictive variables explain 32.10% of the variation. From this side the model is much improved. Overall, most of the probabilities of the variables are less than 0.0001 and this means there is less than 0.01% chance that the variable coefficients are zero, hence, the variable estimates do have impacts on the "wealth_log". For "time","marriage", "familydeath" and "laidoff" the probabilities are also smaller than 0.05, so they also have impact on the price. In model 6, on average, one unit increase of "pd" level will leads to -0.00059802 units decrease of "wealth_log".

The estimated regression equation to Model 5 is as follows:

Log(wealth) = 14.69088 - 0.00059802 pd -0.00048571time+ 0.01240male + 0.00164age + 0.02213white+0.02276hispanic +0.01631otherrace +0.00563education+0.01644 income - 0.01399 employed - 0.00997divorce - 0.00443 marriage - 0.00799childbirth + 0.00575familydeath- 0.00424laidoff -0.00054837missedwork+ 2.611591E-7socioeconomic

From the above equation and chart, we can find that:

- Overall on average the "male" has 0.01240 units higher "wealth_log" than the average "female".
- On average "time" increase 1 will lead to 0.00048571 unite decrease of "wealth_log", This is similar results as we studied in model 2. So in recent years it is a bit harder to accumulate the wealth.

- On average "age " increase 1 will lead to 0.00164 unit increase of "wealth_log". In model 3, we see some curve change of this influence around "age" 65, but here we study the overall influence. And this is reasonable, as the wealth needs time for accumulation.
- "White" compared to non "white" has 0.02213 higher "wealth_log". Similar interpretation for other variables such like "hispanic", "education" etc.
- It is interesting to find that "employed " compare to not "employed" has -0.01399 units difference on "wealth_log". So that means employment does not must mean positive contribution to wealth than unemployment. On the other side, the people who have high wealth may do not need to work.
- "marriage" and "childbirth" may leads to higher cost of life, so the wealth may decrease, so their coefficients are also negative. "divorce" has also negative coefficient, the wealth maybe divided by the divorced partner etc.
- "Familydeath" means a household member recently died, this also leads to increase of wealth. This may be caused by inheritance or insurance income.
- If we have a look at "pd", overall it still has negative contribution to wealth. The "pd" level increase 1 leads to 0.00059802 decrease of "wealth_log".

6. CLRM tests

In this study, the CLRM assumptions for errors have zero mean, linearity, homoskedasticity of the errors, statistical independence of the errors and normality of the error distribution are tested on Models 1, 5 and 6 [3]. Additionally we will also study the multicollinearity.

6.1 Assumption 1: The Average Values of Errors Is Zero

There is no diagnostic test for this assumption. To fulfil the assumption, the intercept must not be excluded from the regression Model. As the p-value of the intercept for Models 1, 5 and 6 was less than 0.0001, the intercept is significant and has to be kept in the Model. The inclusion of the intercept implies that the assumption of average error terms is zero has not been violated[4]. If we check the mean residual values of model 1, 5 and 6 which is shown in below chart, we can also find that the average residuals for three models are almost zero, so this assumption is not violated.

Analysis Variable : residual Residual				
	Mean	Std Dev	Minimum	Maximum
Model 1	-1.26557E-12	0.098514	-0.0984488	0.652424
Model 5	-1.26552E-12	0.096837	-0.1112126	0.6633876
Model 6	-1.26452E-12	0.0820704	-0.8774089	0.7350272

6.2 Assumption 2: There Is a Linear Relationship Between Dependent And Independent Variables

We assume that the regression models are linear in parameters. To test this assumption, we generate the correlation scores between the dependent variable and independent variables for model 1,5 and 6:

	NAME	time	pd	age	male	white	black	hispanic	other race	education	income	employed	divorce	marriage	childbirth	familydeath	laidoff	missed work	socio econ	pd_male
Model 1	Wealth_log		-0.15																	
Model 5	Wealth_log		-0.15		0.19															-0.04
Model 6	Wealth_log	0.01	-0.15	0.32	0.19	0.27		-0.03	0	0.28	0.38	-0.02	-0.08	-0.07	-0.08	0.02	-0.06	-0.03	0.38	-0.04

For model 1 and 5, all the correlation scores are below 0.2, so nearly no correlation between the independent and dependent variables.
For model 6, some correlation scores are between 0.2 to 0.4, such like “wealth_log” with “age” (0.32), with “white” (0.27), with “education” (0.28), with “income”(0.38), with “socioeconomic”(0.38); So the correlation scores of model 6 shows a weak correlation between dependent and some independent variables.

6.3 Assumption 3: The Disturbances have Constant Variance

White test is a statistical test that establishes whether the variance of the errors in a regression model is constant: that is for homoskedasticity[5]. The null hypothesis for White's general test is evaluated and used to determine the state of violation of assumption 3. The null and alternate hypothesis are as follows:
H0: $\sigma^2_i = \sigma^2$, the variances of errors/disturbances are equal, or the presence of homoskedasticity.
H1: $\sigma^2_i \neq \sigma^2$, the variances of errors are not equal or the presence of heteroskedasticity.
Results of White's Test:

	Model 1		Model 5		Model 6	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
White's General Test	12.19	0.0023	9.69	0.0459	7.81	0.6473

For white test, we see that for model 1 and 5, the p-values are both less than 0.05, so we reject the null hypothesis that the disturbances are homoscedastic. But for model 6, we see that the p-value is bigger than 0.05, so we accept the null hypothesis that the disturbances are homoscedastic;

6.4 Assumption 4: The Disturbances Are Not Correlated

Breusch-Godfrey's test is conducted on the detection of autocorrelation of up to the 10th order. The null and alternate hypotheses for the Breusch-Godfrey test are [4]:
H0: p1 = 0 and p2 = 0 and p10 =0, the errors are serially uncorrelated up to the 10th order.
H1: p1 ≠ 0 and p2 ≠ 0 and p10 ≠ 0, the errors are not serially uncorrelated up to the 10th order.

	Model 1		Model 5		Model 6	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Godfrey test AR(1)	1960.9383	<.0001	1532.4361	<.0001	83.883	<.0001
Godfrey test AR(2)	3238.063	<.0001	2612.498	<.0001	166.1072	<.0001
Godfrey test AR(3)	4094.5781	<.0001	3397.0528	<.0001	224.7823	<.0001
Godfrey test AR(4)	4655.7702	<.0001	3942.5838	<.0001	271.1962	<.0001
Godfrey test AR(5)	4961.9598	<.0001	4257.3773	<.0001	290.4235	<.0001
Godfrey test AR(6)	5298.4615	<.0001	4622.9176	<.0001	368.9614	<.0001
Godfrey test AR(7)	5527.1635	<.0001	4888.9564	<.0001	414.331	<.0001
Godfrey test AR(8)	5780.0746	<.0001	5185.1731	<.0001	487.2118	<.0001
Godfrey test AR(9)	5882.1379	<.0001	5319.5	<.0001	503.6706	<.0001
Godfrey test AR(10)	5939.1953	<.0001	5404.5362	<.0001	509.053	<.0001

In the Godfrey test allowing for 10 lags, for three models the p-values are all less than 0.001, so we reject the null hypothesis in which the residuals are serially uncorrelated up to the 10th order has to be rejected. Models 1, 5 and 6 violated the CLRM assumption of residuals are not serially correlated and the residuals in these Models are serially correlated.

6.5 Assumption 5: The Disturbances Are Normally Distributed

When analyze data, it's essential to understand its underlying distribution. One common distribution that arises in statistical analysis is the normal distribution. The Jarque-Bera test is a statistical test used to assess whether a dataset follows a normal distribution[6]. The null hypothesis of the Jarque-Bera test is that the residuals are normally distributed and the alternate hypothesis is that the residuals are not normally distributed.

	Model 1		Model 5		Model 6	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Jarque-Bera Test	385763.013	<.0001	386411.534	<.0001	404310.785	<.0001

The Jarque-Bera normality values for Models 1, 5and 6 are <0.001. The null hypothesis that the residuals term for Models 1, 5 and 6 are normally distributed is rejected. This means that the residuals for Models 1, 5 and 6 are not normally distributed.

6.6 No Presence of Multicollinearity

The multicollinearity test was conducted using the Variance Inflation Factor (VIF)[7,8]. Multicollinearity means that there is a high correlation between the predictive variables within the sample. The condition is that if VIF is found to be greater than 5, then there is a presence of multicollinearity[4].

VIF value			
	Model 1	Model 5	Model 6
Intercept	0	0	0
pd	1	2.57318	1.09431
time			1.04205
male		1.8069	1.28276
age			1.46512
white			1.30738
hispanic			1.05588
otherrace			1.10035
education			1.2818
income			1.77582
employed			1.46015
divorce			1.03446
marriage			1.07979
childbirth			1.10201
familydeath			1.01829
laidoff			1.02248
missedwork			1.01711
socioeconomic			1.33686
pd_male		3.0317	

All the VIF values are below 5, so there is no multicollinearity issue with these variables.

7. Summary

In this study we focus on the independent variable “pd” to study its influence on the dependent variable “wealth”. And we also added control variable “time” “age” “male” etc step by step for more analysis.

In data overall study and data wrangling part we generally studied all the data basics and distribution to get the overall understanding, and deleted value missing observations and used log for “wealth” to have better data distribution for analysis, then we also delete the top 1% and lowest 1% “wealth_log” data to eliminate the discrete dots in the “the least squares residuals” chart.

In the study we used 6 models to mainly research the “pd” influence on “wealth”. Models 1 to 3 used simple linear regression and models 4 to 6 used multiple regression.

- In model 1, we directly study the “pd” and “wealth_log”, to get an overall impression that “pd” has an negative influence on wealth, although there are also other factors that can influence the wealth.
- Model 2 is based on model 1, to discuss about the “pd” influence on wealth in different years from 2003 to 2019. This covers a long period of time and the society and economic background has also changed a lot in the 16 years. In general we see that “pd” has a more and more influence on wealth along the time. That means recent years people should care more and more about the psychological regulation to keep in good mode of “pd” level and this is a basis for wealth.
- In model 3, we further study the “pd” influence on wealth in different ages, one interesting finding is the turning point at around 65 years old, which is obviously related to the retirement age. So people should pay higher attention to “pd” level before the retirement, as its influence on wealth is becoming bigger and bigger when the “age” grows before the retirement.
- In model 4, we use multiple regression we add gender into consideration. We find that for same level “pd” on average the “male” has 0.03797 units higher “wealth_log” than “female”.
- In model 5, we want to connect “pd” and gender for more discussion, so we add one new variable “pd*male”. We find that, on average for “male”one unit increase of “pd” level will leads to -0.00422 units decrease of “wealth_log”, while for “female” only -0.00141 units decrease of “wealth_log”. So on average compared to “female”, for “male” the “pd” level increase has higher negative influence on “wealth_log”. So “male” should care much more about their “pd” level for accumulating more wealth.
- In mode 6, we add nearly all the other parameters to improve the model and primarily study other variables’ influences on wealth and “pd” influence on wealth in this model. The R-square is much improved to 0.3210, which is much higher than other models.

Further more, for model 1, 5 and 6, we have done CLRM test for the 5 assumptions and checked the multicollinearity:

- Assumption 1: The Average Values of Errors Is Zero. Model 1, 5 and 6 do not violate the assumption.
- Assumption 2: There Is a Linear Relationship Between Dependent And Independent Variables. Model 1 and 5 violate the assumption. Model 6 shows a weak correlation between dependent and some independent variables.
- Assumption 3: The Disturbances have Constant Variance. For model 1 and 5, the p-values are both less than 0.05, so we reject the null hypothesis that the disturbances were homoscedastic. For model 6, we see that the p-value is bigger than 0.05, so we accept the null hypothesis that the disturbances are homoscedastic.
- Assumption 4: The Disturbances Are Not Correlated. Models 1, 5 and 6 violate the assumption that the residuals are not serially correlated and the residuals in these Models were serially correlated.
- Assumption 5: The Disturbances Are Normally Distributed. The distributions of residuals for Models 1, 5 and 6 are not normally distributed.
- No Presence of Multicollinearity. For Models 1, 5 and 6 there is no multicollinearity issue.

In general the model 6 is best in consideration of more variables, with highest R-square and fulfil more CLRM assumptions.

Reference:

[1] 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

[2] Hampton, Y. R. (2023). The relationship between psychological distress and wealth: an analysis of household wealth changes from 2007-2019 in the United States by race, gender, and retirement status. https://doi.org/10.32469/10355/94063

[3] Balloch, A., Engels, C., & Philip, D. (2022). When it rains it drains: psychological distress and household net worth. Social Science Research Network. https://doi.org/10.2139/ssrn.3521323

[4] Airbnb case study 1-3 for applied economic lecture on the stream of Massey University

[5] Wikipedia contributors. (2024, February 10). White test.
Wikipedia. https://en.wikipedia.org/wiki/White_test

[6] Khadka, N., & Khadka, N. (2023, December 6). Jarque-Bera Test: Guide to Testing Normality with Statistical Accuracy - Dataaspirant. Dataaspirant - A Data Science Portal For
Beginners. https://dataaspirant.com/jarque-bera-test/

[7] Bobbitt, Z. (2021, October 21). A Guide to Multicollinearity & VIF in Regression.
Statology. https://www.statology.org/multicollinearity-regression/

[8] Regression with SAS Chapter 2 – Regression Diagnostics.
(n.d.b). https://stats.oarc.ucla.edu/sas/webbooks/reg/chapter2/regressionwith-saschapter-2-regression-diagnostics/

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;

* input the excle file *;
PROC IMPORT OUT=WORK.Assessment1_data
    DATAFILE="&path\Assessment1_data.xlsx"
          DBMS=XLSX REPLACE;
          GETNAMES=Yes;
          Run;

* 2. Overall data study;
* 2.1 Overall data status;
proc contents data=WORK.Assessment1_data;
run;
proc means data=WORK.Assessment1_data;
 var time -- socioeconomic;

* 2.2 Data overall study findings;
proc sort data=Assessment1_data;
by male;
run;
proc means data=WORK.Assessment1_data;
by male;
 var pd wealth;

*  3. Primary data wrangling;
* 3.1 Delete the observations with missing values;
* delete the "pd" observations with no values;
 data Assessment2_data;
  set Assessment1_data;
if cmiss(pd) then delete;
run;
proc means data=WORK.Assessment2_data;
 var time -- socioeconomic;

* delete other observations with missing values;
   data Assessment3_data;
  set Assessment2_data;
if cmiss(age) then delete;
if cmiss(white) then delete;
if cmiss(black) then delete;
if cmiss(hispanic) then delete;
if cmiss(otherrace) then delete;
if cmiss(education) then delete;
if cmiss(income) then delete;
if cmiss(laidoff) then delete;
if cmiss(socioeconomic) then delete;
run;
proc means data=WORK.Assessment3_data;
 var time -- socioeconomic;

*“wealth” distribution ;
proc univariate data=WORK.Assessment3_data;
 var wealth;
 histogram / normal;
```

```
run;
ods output Histogram=hist_table;

* 3.2 Apply log for “wealth” data;
data Assessment4_data;
 set WORK.Assessment3_data;
 Wealth_log=log(wealth +3197000+1);
 run;
proc univariate data=WORK.Assessment4_data;
  var wealth_log;
  histogram / normal;
run;

* “wealth” distribution of different level of“pd”;
proc means data=Assessment4_data mean min max;
  var wealth;
  class pd;
  output out=summary_stats mean=mean_wealth min=min_wealth max=max_wealth;
run;
proc sgplot data=summary_stats;
  vbar pd / response=mean_wealth datalabel;
  title 'Mean wealth_log by pd';
run;


* 4. simple linear regression;
proc reg data=Assessment4_data;
model wealth_log = pd;
title 'wealth_log regression';
run;

*The least squares residuals;
proc reg data=Assessment4_data;
model wealth_log = pd;
plot residual.*pd;
plot wealth_log*pd;
title 'regression with plot options';
run;

proc univariate data=WORK.Assessment4_data;
  var wealth_log;
  histogram / normal;
run;

* delete the top 1% data and lowest 1% data ;
data work.Assessment5_data;
   set WORK.Assessment4_data;
   if 14.9451<wealth_log<15.6745;
run;
proc reg data=Assessment5_data;
model wealth_log = pd;
plot residual.*pd;
plot wealth_log*pd;
title 'regression with plot options';
run;
proc univariate data=WORK.Assessment5_data;
  var wealth_log;
  histogram / normal;
run;

* model 1 : linear regression;
```

```
proc reg data=Assessment5_data;
model wealth_log = pd;
title 'wealth_log regression';
run; run;

* model 2: study the pd trend in different time;
proc sort data=Assessment5_data;
by time;
run;
ods output ParameterEstimates=reg_estimates;
proc reg data=Assessment5_data;
by time;
model wealth_log = pd;
title 'Wealth Regression by Time';
run;
proc sql;
create table pd_estimates as
select time, Estimate as pd_estimate
from reg_estimates
where Variable = 'pd';
quit;
proc print data=pd_estimates;
title 'Estimates of pd by Time';
run;
proc sgplot data=pd_estimates;
title 'Plot of pd_estimate by Time';
series x=time y=pd_estimate / lineattrs=(color=blue);
xaxis values=(0 to 8 by 1);
run;

* mean of wealth_log by time;
proc means data=WORK.Assessment5_data mean noprint;
 by time;
 var wealth_log;
 output out=wealth_means mean=mean_wealth;
run;
proc sgplot data=wealth_means;
 title 'mean wealth in different time';
 series x=time y=mean_wealth / markers;
 xaxis label='time' values=(0 to 8);
 yaxis label='mean wealth_log';
run;

* mean of pd by time;
proc sort data=WORK.Assessment5_data;
 by time;
run;
proc means data=WORK.Assessment5_data mean noprint;
 by time;
 var pd;
 output out=pd_means mean=mean_pd;
run;
proc sgplot data=pd_means;
 title 'mean pd in different time';
 series x=time y=mean_pd / markers;
 xaxis label='time' values=(0 to 8);
 yaxis label='mean pd';
run;

* model 3: study the pd trend in different age;
proc sort data=Assessment5_data;
```

```
by age;
run;
ods output ParameterEstimates=reg_estimates;
proc reg data=Assessment5_data;
by age;
model wealth_log = pd;
title 'Wealth Regression by age';
run;
proc sql;
create table pd_estimates as
select age, Estimate as pd_estimate
from reg_estimates
where Variable = 'pd';
quit;
proc print data=pd_estimates;
title 'Estimates of pd by age';
run;
proc sgplot data=pd_estimates;
title 'Plot of pd_estimate by age';
series x=age y=pd_estimate / lineattrs=(color=blue);
xaxis values=(16 to 99 by 1);
run;

* mean of wealth_log by age;
proc sort data=WORK.Assessment5_data;
  by age;
run;
proc means data=WORK.Assessment5_data mean noprint;
  by age;
  var wealth_log;
  output out=wealth_means mean=mean_wealth;
run;
proc sgplot data=wealth_means;
  title 'mean wealth in different age';
  series x=age y=mean_wealth / markers;
  xaxis label='age' values=(16 to 99);
  yaxis label='mean wealth_log';
run;

* mean of pd by age;
proc sort data=WORK.Assessment5_data;
  by age;
run;
proc means data=WORK.Assessment5_data mean noprint;
  by age;
  var pd;
  output out=wealth_means mean=mean_pd;
run;
proc sgplot data=wealth_means;
  title 'Mean PD in Different Age Groups';
  series x=age y=mean_pd / markers;
  xaxis label='Age' values=(16 to 99);
  yaxis label='Mean PD';
run;

* 5. Multiple regression analysis;
* show the correlation of wealth, pd...socioecomomic;
proc corr data=WORK.Assessment5_data pearson nosimple noprob plots=none outp=corr_out;
 var pd -- wealth_log;
proc print data=corr_out noobs;
run;
```

```
proc sort data=WORK.Assessment5_data;
  by age;
run;
proc means data=WORK.Assessment5_data mean noprint;
  by age;
  var pd;
  output out=pd_means mean=mean_pd;
run;
proc sgplot data=pd_means;
  title 'mean pd in different age';
  series x=age y=mean_pd / markers;
  xaxis label='age' values=(16 to 99);
  yaxis label='mean pd';
run;

* show the correlation of wealth, pd...socioecomomic;
proc corr data=WORK.Assessment5_data pearson nosimple noprob plots=none outp=corr_out;
 var pd -- wealth_log;
proc print data=corr_out noobs;
run;

* model 4 study the pd trend in different gender;
proc reg data=Assessment5_data;
model wealth_log = pd male;
title 'wealth regression';
run;

* model 5 study the pd trend in different age in different gender;
data Assessment6_data;
 set WORK.Assessment5_data;
 pd_male=pd*male;
 run;
proc reg data=Assessment6_data;
model wealth_log = pd male pd_male;
title 'wealth regression';
run;

* model 6 study the pd trend in different age in different gender;
proc reg data=Assessment6_data;
model wealth_log = pd time male age white hispanic otherrace education income employed divorce marriage childbirth
familydeath laidoff missedwork socioeconomic;
title 'wealth regression';
run;

* 6. CLRM test;
*6.1 Assumption 1: The Average Values of Errors Is Zero

proc reg data=Assessment6_data;
   model wealth_log = pd;
   output out=Residuals predicted=yhat residual=residual;
run;
proc means data=Residuals;
   var residual;
run;

proc reg data=Assessment6_data;
   model wealth_log = pd male pd_male;
   output out=Residuals predicted=yhat residual=residual;
run;
proc means data=Residuals;
```

```
   var residual;
run;

proc reg data=Assessment6_data;
   model wealth_log = pd time male age white hispanic otherrace education income employed divorce marriage childbirth
familydeath laidoff missedwork socioeconomic;
   output out=Residuals predicted=yhat residual=residual;
run;
proc means data=Residuals;
   var residual;
run;

* 6.2 Assumption 2: There Is a Linear Relationship Between Dependent And Independent Variables;
proc means data=WORK.Assessment6_data;
  run;
proc corr data=WORK.Assessment6_data pearson nosimple noprob plots=none outp=corr_out;
 var time -- pd_male;
proc print data=corr_out noobs;
run;

* 6.3 Assumption 3: The Disturbances have Constant Variance;
* White test;

proc reg data=WORK.Assessment6_data;
model wealth_log = pd;
PROC MODEL;
PARMS B0 B1;
wealth_log = B0 + B1*pd;
FIT wealth_log /WHITE;
RUN;

proc reg data=WORK.Assessment6_data;
model wealth_log = pd male pd_male;
PROC MODEL;
PARMS B0 B1 B2 B3;
wealth_log = B0 + B1*pd + B2*male + B3*pd_male;
FIT wealth_log /WHITE;
RUN;

proc reg data=WORK.Assessment6_data;
model wealth_log = pd time male age white hispanic otherrace education income employed divorce marriage childbirth
familydeath laidoff missedwork socioeconomic;
PROC MODEL;
PARMS B0 B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12 B13 B14 B15 B16 B17;
wealth_log
=B0+B1*pd+B2*time+B3*male+B4*age+B5*white+B6*hispanic+B7*otherrace+B8*education+B9*income+B10*employed+B11*di
vorce+B12*marriage+B13*childbirth+B14*familydeath+B15*laidoff+B16*missedwork+B17*socioeconomic;
FIT wealth_log /WHITE;
RUN;

*6.4 Assumption 4: The Disturbances Are Not Correlated;
* Godfrey test allowing for 10 lags;
proc autoreg data=WORK.Assessment6_data;
model wealth_log = pd/godfrey=10;
run;

proc autoreg data=WORK.Assessment6_data;
model wealth_log = pd male pd_male/godfrey=10;
run;

proc autoreg data=WORK.Assessment6_data;
```

```
model wealth_log = pd time male age white hispanic otherrace education income employed divorce marriage childbirth
familydeath laidoff missedwork socioeconomic/godfrey=10;
run;

*6.5 Assumption 5: The Disturbances Are Normally Distributed;
* Jarque-Bera Normality test;
proc autoreg  data=WORK.Assessment6_data;
model wealth_log = pd/NORMAL;
run;

proc autoreg  data=WORK.Assessment6_data;
model wealth_log = pd male pd_male/NORMAL;
run;

proc autoreg  data=WORK.Assessment6_data;
model wealth_log = pd time male age white hispanic otherrace education income employed divorce marriage childbirth
familydeath laidoff missedwork socioeconomic/NORMAL;
run;


*6.6 No Presence of Multicollinearity;

proc reg data=WORK.Assessment6_data;
 model wealth_log = pd / vif;
run;

proc reg data=WORK.Assessment6_data;
 model wealth_log = pd male pd_male/ vif;
run;

proc reg  data=WORK.Assessment6_data;
model wealth_log = pd time male age white hispanic otherrace education income employed divorce marriage childbirth
familydeath laidoff missedwork socioeconomic/ vif;
run;
```