Alexander Lindsey
Econ 326
Aug. 2, 2018
Assignment 3

The Effect of Education, Age, and Tenure on the Performance of Mutual Funds

Alexander Lindsey
43438150

Table of Contents

Analysis

I. Multiple Linear Regression Estimation and Analysis

This report seeks to analyze the effect of mutual fund manager's age, education, and tenure, on the overall performance of their funds. The data from 2,029 managers and the performance of their mutual funds were randomly sampled and collected. Two variables describe the manager's education: his/her SAT score and whether or not the manager has an MBA, with yes=1 and no=0. The overall performance of each fund was measured "by its risk-adjusted excess return", or the discrepancy between the actual return of the investment and the return which is considered "standard". The ages and tenure, or how many years the manager has been in charge, was also collected for each observation.

Stata was used to analyze the data. By performing a multiple regression of sat, mba, age, and tenure on return, the intercept, coefficient, standard error, and p-value for each independent variable was estimated, and is summarized in Table 1.

SAT scores (sat) were found to have a coefficient of .0050736, meaning that a 1 unit *ceteris paribus* increase in sat would, on average, result in an increase in return by aprox. .0051 units. A positive relationship between the two variables was expected, as a higher SAT score reflects a higher quality of education, which would positively impact the return on mutual funds. The p-value for this independent variable is 0.00, meaning that we reject the null hypothesis that H_0: $B_{SAT} = 0$ in favour of the alternative hypothesis, H_1: $B_{SAT} \neq 0$ at any significance level, as the p-value < any significance level. Most analysis usually use a significance level of 0.05.

The independent variable mba was found to have a coefficient of .6744004. For the same reasons as sat, a positive relationship was to be expected, but as we can see mba has a stronger effect on return. Therefore, having an mba, meaning mba=1 as opposed to mba=0, increases

return by .674 units, *ceteris paribus*. The p-value for this coefficient is 0.073, meaning we fail to reject H_0: $B_{mba} = 0$ at the 5% significance level.

The coefficient for age was -0.1405739, indicating a negative relationship between age and return, which is unexpected. One would expect that with age comes experience, and therefore a higher return. A *ceteris paribus* one unit increase in age will decrease return by aprox. 0.1406 units. The p-value is 0.001, meaning that we reject H_0: $B_{age} = 0$ at the 5% significance level in favour of the alternative hypothesis, H_1: $B_{age} \neq 0$.

The coefficient for tenure was found to be 0.0818061, which follows the expectation that more experience would increase return. A *ceteris paribus* 1 unit increase in tenure would increase the average return by aprox. .08181units. The p-value for tenure is 0.641, meaning that we fail to reject H_0: $B_{tenure} = 0$ at the 5% significance level as .641 > .05.

The intercept, _cons, was found to be -1.147635, however this is not important for our model. This intercept measures return if age, sat, mba, and tenure are all 0, which is impossible and therefore does not occur in our data.

We found the R-squared for the model to be 0.0151, meaning that only 1.51% of the variation in return can be explained by the multiple regression model. Furthermore, our p-value for our F value of 7.75 was found to be 0.00 for the model. When compared to any standard significance level, usually .01%, 1%, or 5%, we find that, as our p-value is less than each of these significance levels. As a result, we can conclude that sat, tenure, age, and mba all together reliably predict return. Lastly, the regression also gave us a 95% confidence interval for each parameter, meaning that there is a 95% chance that the true value of the population parameter lies within that interval. If an interval spans from negatives to positives, then were are less sure as to what the true direction of the parameter in the population. Furthermore, this also includes

the possibility that the parameter is 0 within the 95% interval. If the true parameter is 0, then we have over specified our model by including a variable that we think impacts return, but in reality does not. It is important to note that intervals that include 0 also fail to reject the null hypothesis that H_0: $B_j = 0$.

II. Could a More Simple Model be Used?

At the 5% level of significance, neither tenure nor mba are statistically significant as their respective p-values are greater than 0.05. Additionally, both confidence intervals include zero. We checked if they can be dropped from the model using an F test.

Under the null hypothesis, assume that the coefficient on tenure and mba is zero. The alternative states that at least one of these coefficients is not zero.

The F-test gives a p-value of 0.1835, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that we can safely drop mba and tenure from the regression model. This original regression has turned out to be over specified.

We then estimate the new and more simple regression of return on sat and age only, and find that the adjusted R-squared for the model with all 4 variables is 0.0131 and the adjusted R-squared with just sat and age is 0.0125. The original model fits slightly better, even with the degrees of freedom correction included in the adjusted R-squared measure.

In the new model, summarized in Table 2, the coefficients for sat and age are 0.005098 and -1.303327, respectively. The intercept, _cons, does not matter in this model for the same reasons as before: there are no observations where sat=0 age age=0 as this is impossible. Having observations with these values would indicate errors during the sample collection process. The p-values for sat and age are 0.00 and 0.001, respectively, meaning that we reject the null

hypothesis that either parameter is equal to 0. Further supporting this claim, the 95% confidence interval for both age and sat do not include 0.

III. Validity of the Model

The validity of the multiple linear regression model and the Ordinary Least Squares (OLS) we derived depend on satisfying the Gauss-Markov assumptions, which are stated as follows:

1. the dependent variable can be expressed as a linear function, $y = B_0 + B_1 x_1 + \cdots + B_k x_k + u$;
2. the observations were randomly sampled;
3. the expected value of the error term is 0;
4. there is no perfect collinearity, meaning that there is no perfect linear relationship among any of the independent variables;
5. homoskedasticity exists, meaning that the error terms all have the same constant variance;
6. the u terms for each observation is normally distributed with a mean of 0 and a constant variance.

We previously found in part II that the coefficients for tenure and mba are likely 0, meaning that our population model has been revised from return = $B_0 + B_{sat}sat + B_{mba}mba + B_{age}age + B_{tenure}tenure + u$ to return = $B_0 + B_{sat}sat + B_{age}age + u$. Furthermore, to satisfy the second assumption, we must assume that the observations were randomly sampled.

To verify the third, fifth, and sixth assumptions, we created a kernel density estimate using the residuals for each observation. By graphing the kernel density against a normal

distribution, we found an almost perfect fit. This shows that the residuals follow a normal distribution with a mean of 0 and constant variance. This is shown in Figure 1.

To further verify homoskedasticity, we created a scatter plot of the residuals, which are the difference between the predicted return and observed return, against the linear prediction, or fitted values, for each observation. The scatter plot does not show a discernible pattern in the data. This is shown in Figure 2.

We can use a Bruesch-Pagan Test to quantify the relationship seen in the scatter plot. The null hypothesis assumed homoskedasticity, while the alternative hypothesis assumes heteroskedasticity. The test gives us a p-value of 0.0002, which is well below any standard significance level. There is insufficient evidence to reject homoskedasticity, and we conclude that the data is homoscedastic.

To test multicollinearity, we constructed a correlation matrix with sat and age. The correlation between the two variables was found to be -0.0018, meaning that there is a very small negative relationship. Therefore, there is no evidence of multicollinearity. To further support this, the vif score in stata showed values of 1 for both sat and age, further refuting multicollinearity. The correlation matrix is shown in Figure 3.

To test linearity and exogeneity, which is to test that u is not dependent on sat or age, we use scatter plots to identify patterns between residuals and the independent variables, as seen in Figure 3 and 4. Neither graph showed a discernible pattern. If we had an identifiable pattern, it would suggest that the error term is in fact endogenous, which is a violation of the third assumption, and would lead to bias estimates of the OLS estimators.

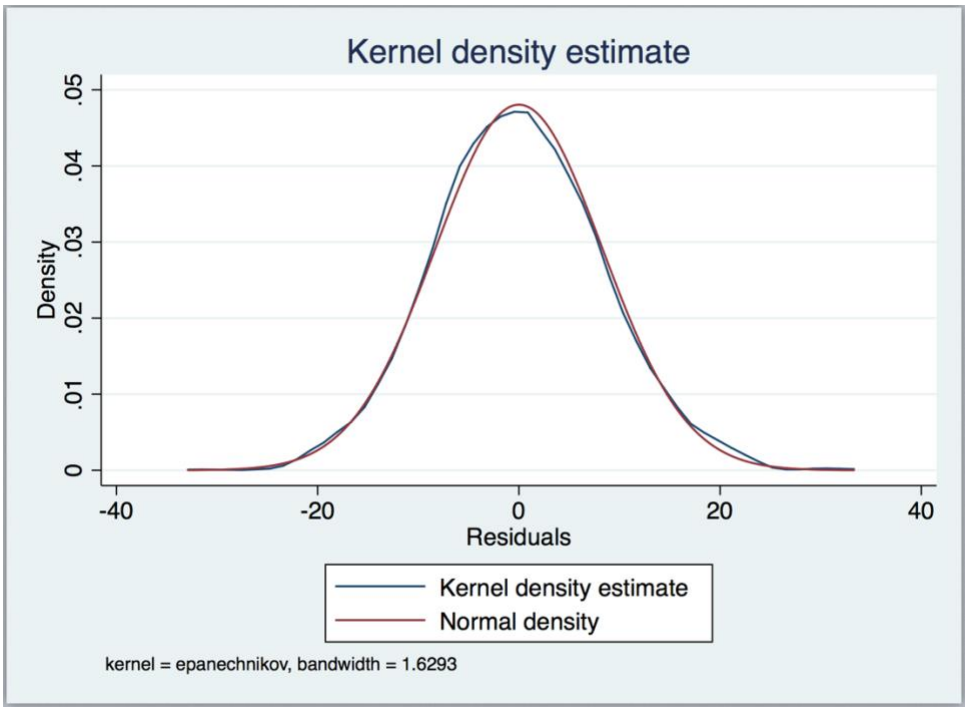
IV. Adding logarithmic or quadratic terms

To check for additional functional forms, a scatter plot was created to check for a quadratic or logarithmic relationship between return and age, as age is often included as a quadratic term. This scatter plot is seen in Figure 5. While there was no clear pattern to suggest any such relationship exists, we still estimate a regression model including a quadratic term for age to verify that this new term's coefficient would be 0.

Generating a new term, age-squared, and regressing age, age-squared, and sat on return, we found that both coefficients for age and age-squared became negative and their p-values became .945 and .844, respectively. This shows that the estimations are very insignificant in this new model.

To check for logarithitmic functions, we generated lnsat as the ln of sat. By regressing lnsat and age on return, we found that lnsat had coefficient of aprox. 5.84 and p-value of 0.00, which is significant by any confidence level. This means that a 1% increase in sat scores is predicted to increase return by 0.0584 units.

List of Figures



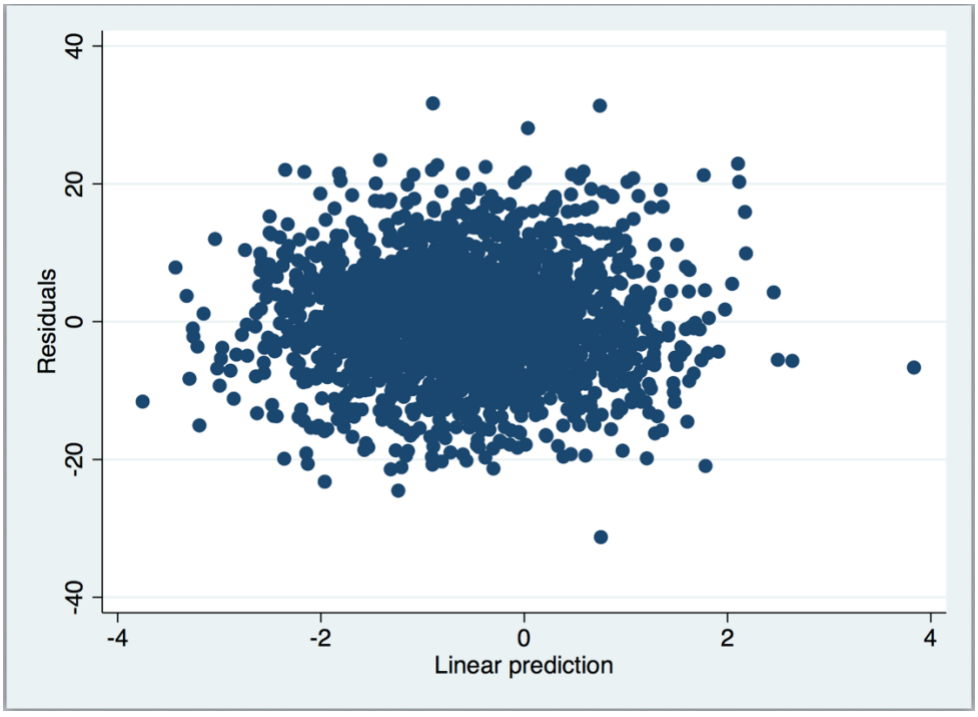


Figure 3: *Scatter plot of residuals and age*

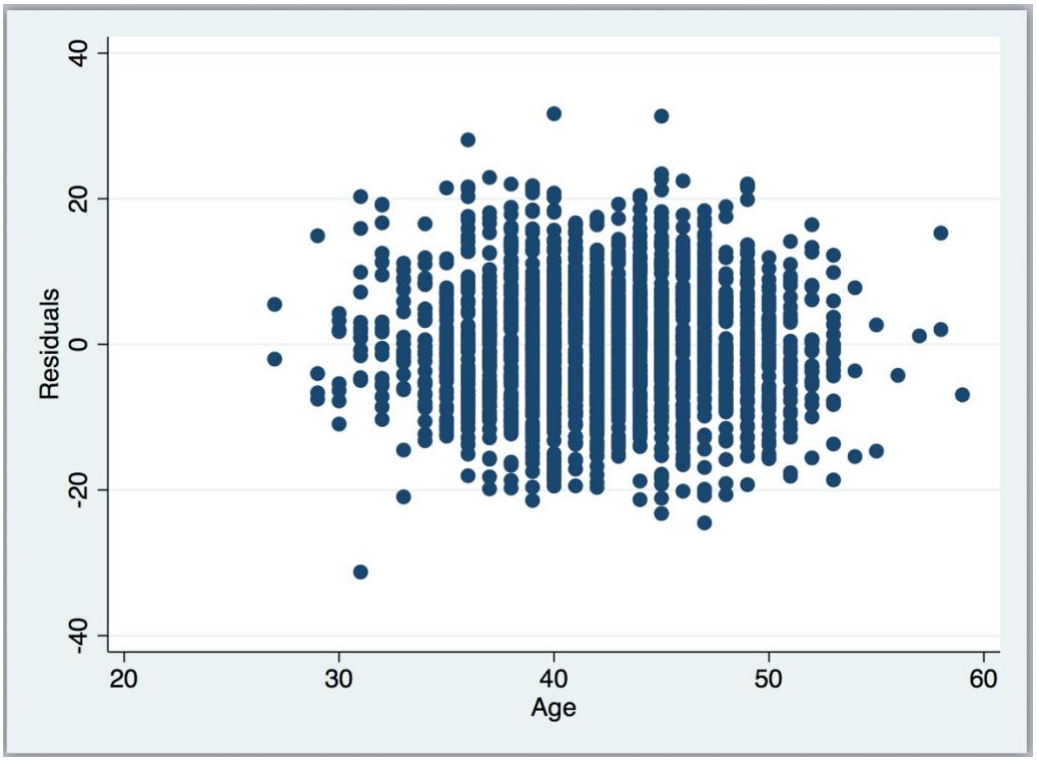


Figure 4: *Scatter plot of residuals and sat*

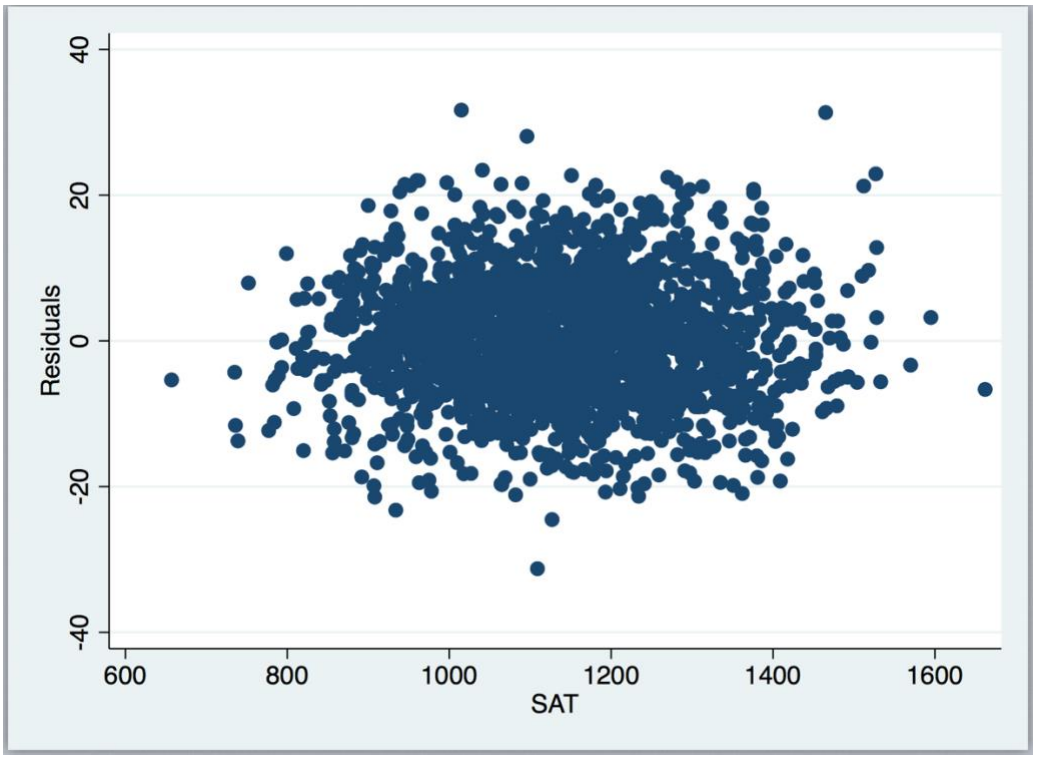
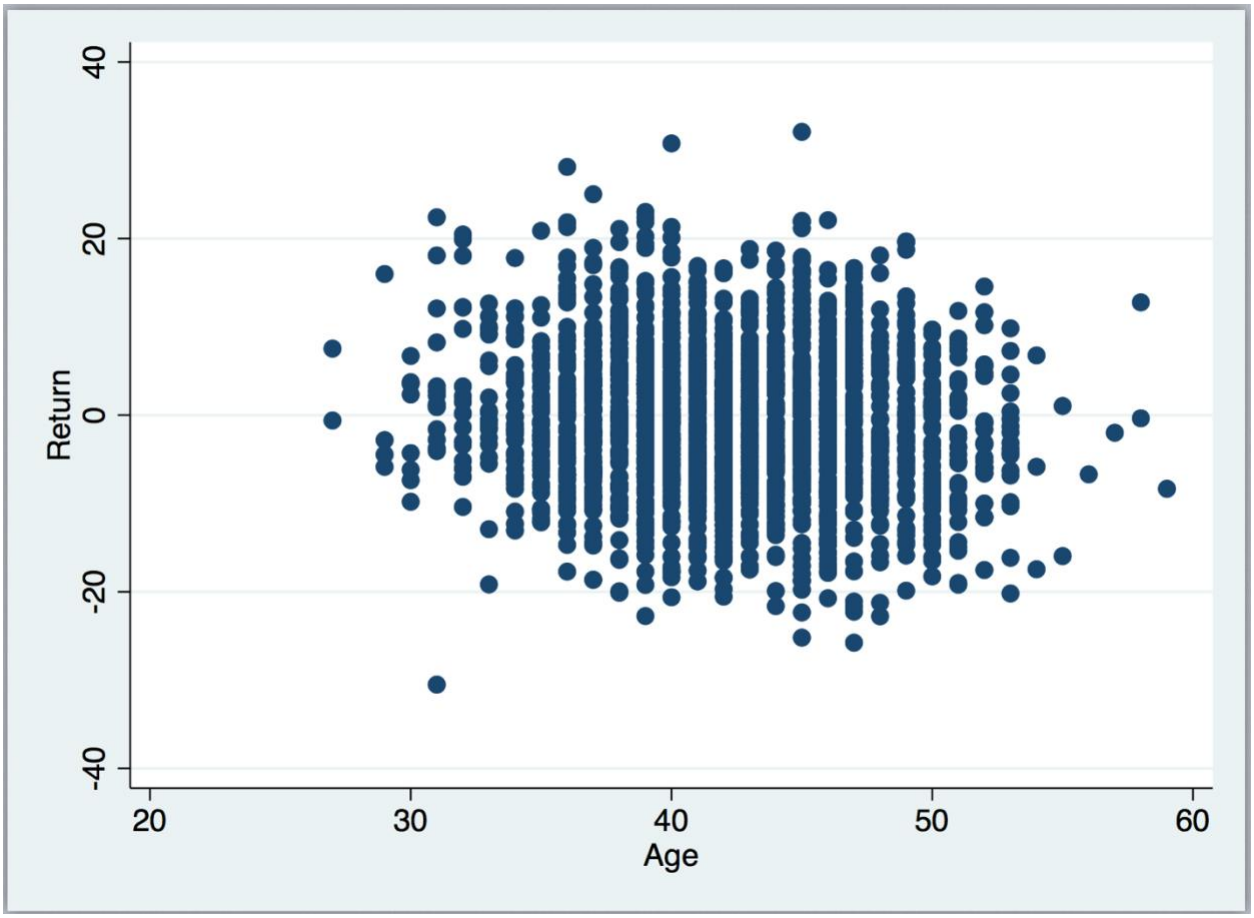


Figure 5: *Scatter plot of return and age*



List of Tables

Table 1: *Summary of regression of sat, mba, age, and tenure on return*

Independent Variable	Coefficient	p-value	[95% Confidence Interval]	
sat	0.0050736	0.000	0.0025606	0.0075866
mba	0.6744004	0.073	-0.062909	1.41171
age	-0.1405739	0.001	-0.223769	-0.0573789
tenure	0.0818061	0.641	-0.262442	0.4260539

Table 2: *Summary of regression of sat and age on return*

Independent Variable	Coefficient	p-value	[95% Confidence Interval]	
sat	0.005098	0.000	0.002585	0.007611
age	-0.130333	0.001	-0.2050429	-0.0556225

Table 3: *Correlation matrix between sat and age*

	sat	age
sat	1.0000	1.0000
age	-0.0018	1.0000

Appendix: Stata .log File

User: Alexander Lindsey

```
      name:  <unnamed>
       log:  /Users/alexanderlindsey/Desktop/econ326_a3_alex.smcl
  log type:  smcl
 opened on:   2 Aug 2018, 12:14:02

1 .
2 . import excel "/Users/alexanderlindsey/Desktop/C17-01.xlsx", sheet("Sheet1") firstrow

3 . rename Return return

4 . rename SAT sat

5 . rename MBA mba

6 . rename Age age

7 . rename Tenure tenure

8 . regress return sat mba age tenure

      Source |       SS           df       MS      Number of obs   =     2,029
-------------+----------------------------------   F(4, 2024)      =      7.75
       Model |  2137.16895         4  534.292237   Prob > F        =    0.0000
    Residual |   139560.99     2,024  68.9530582   R-squared       =    0.0151
-------------+----------------------------------   Adj R-squared   =    0.0131
       Total |  141698.159     2,028   69.870887   Root MSE        =    8.3038

------------------------------------------------------------------------------
      return |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sat |   .0050736   .0012814     3.96   0.000     .0025606    .0075866
         mba |   .6744004   .3759601     1.79   0.073    -.0629087     1.41171
         age |  -.1405739   .0424218    -3.31   0.001    -.2237689   -.0573789
      tenure |   .0818061   .1755348     0.47   0.641    -.2624417    .4260539
       _cons |  -1.147635   2.195444    -0.52   0.601    -5.453201    3.157931
------------------------------------------------------------------------------

9 . browse

10 . test mba tenure

 ( 1)  mba = 0
 ( 2)  tenure = 0

       F(  2,  2024) =    1.70
            Prob > F =    0.1835

11 . regress return sat age
```

User: Alexander Lindsey

Source	SS	df	MS	Number of obs	=	2,029
				F(2, 2026)	=	13.79
Model	1903.12753	2	951.563766	Prob > F	=	0.0000
Residual	139795.031	2,026	69.000509	R-squared	=	0.0134
				Adj R-squared	=	0.0125
Total	141698.159	2,028	69.870887	Root MSE	=	8.3067

return	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.005098	.0012814	3.98	0.000	.002585	.007611
age	-.1303327	.0380954	-3.42	0.001	-.2050429	-.0556225
_cons	-.8592848	2.187331	-0.39	0.694	-5.148937	3.430367

```
12 . predict r, resid

13 . browse

14 . kdensity r, normal

15 . predict f, xb

16 . browse

17 . browse

18 . scatter r f

19 . estat hettest r f

   Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
            Ho: Constant variance
            Variables: r f

            chi2(2)      =    17.18
            Prob > chi2  =   0.0002

20 . correl sat age
   (obs=2,029)
```

	sat	age
sat	1.0000	
age	-0.0018	1.0000

```
21 . vif
```

User: Alexander Lindsey

Variable	VIF	1/VIF
age	1.00	0.999997
sat	1.00	0.999997
Mean VIF	1.00	

```
22 . scatter r age

23 . scatter r sat

24 . scatter return age

25 . scatter return age

26 . gen age2 = age^2

27 . browse

28 . regress return sat age age2
```

Source	SS	df	MS		
Model	1905.78598	3	635.261994	Number of obs	= 2,029
Residual	139792.373	2,025	69.0332705	F(3, 2025)	= 9.20
Total	141698.159	2,028	69.870887	Prob > F	= 0.0000
				R-squared	= 0.0134
				Adj R-squared	= 0.0120
				Root MSE	= 8.3086

return	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.005102	.0012819	3.98	0.000	.0025881	.007616
age	-.0338363	.4932032	-0.07	0.945	-1.001075	.9334023
age2	-.0011424	.0058217	-0.20	0.844	-.0125596	.0102747
_cons	-2.874734	10.50084	-0.27	0.784	-23.4683	17.71884

```
29 . vif
```

Variable	VIF	1/VIF
age	167.53	0.005969
age2	167.53	0.005969
sat	1.00	0.999742
Mean VIF	112.02	

```
30 .
```

User: Alexander Lindsey

```
31 . drop age2

32 . gen lnreturn = ln(return)
   (1,081 missing values generated)

33 . drop lnreturn

34 . gen lnsat = ln(sat)

35 . regress return lnsat age
```

Source	SS	df	MS	Number of obs	=	2,029
				F(2, 2026)	=	14.13
Model	1948.80189	2	974.400943	Prob > F	=	0.0000
Residual	139749.357	2,026	68.9779649	R-squared	=	0.0138
				Adj R-squared	=	0.0128
Total	141698.159	2,028	69.870887	Root MSE	=	8.3053

return	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnsat	5.840276	1.437978	4.06	0.000	3.020207	8.660346
age	-.1300733	.0380893	-3.41	0.001	-.2047716	-.0553749
_cons	-36.11968	10.24732	-3.52	0.000	-56.21607	-16.02329

```
36 . log close
      name:  <unnamed>
       log:  /Users/alexanderlindsey/Desktop/econ326_a3_alex.smcl
  log type:  smcl
 closed on:   2 Aug 2018, 14:51:06
```

```
      name:  <unnamed>
       log:  /Users/alexanderlindsey/Desktop/econ326_a3_alex.smcl
  log type:  smcl
 opened on:   2 Aug 2018, 16:33:55

37 . putdocx
subcommand required
r(198);

38 .
```