Analytics for Value Investing Assignment 1

Name: Lu Kaiqi
ID: 01351798

Q1

a) Source of data:

Source: Yahoo Finance
Data: S&P500 (^GSPC) monthly adjusted close price index
Period: Dec 1985 to Dec 2020 (421 months)

bii) Compute descriptive statistics of those monthly returns

After droping NA:

count	420.000000
mean	0.007854
std	0.043864
min	-0.217630
25%	-0.017368
50%	0.011959
75%	0.035292
max	0.131767

biii) Test the hypotheses that the returns for the month of January are higher or more positive, while those for the month of September are lower or more negative, than those for other months,

c) Briefly explain your statistical results and their implications for the hypotheses. Your answer should also include an explanation of what the intercept (β") of the above regression implies.

***t*-test of difference in mean returns between January and non-January months**	
Test conducted	Step 1: To determine if the population is equal, we conduct a Levene's test at 5% significance level with the hypotheses that: H0: Population variance of Jan return = Population variance of non-Jan return H1: Population variance of Jan return != Population variance of non-Jan return Step 2: Using scipy.stats.levene, we obtain p-value=0.361. This means that the evidence is insufficient to reject the null hypothesis that Population variance of Jan return = Population variance of non-Jan return.

	Step 3: Hence, we assume the population variances are unknown but equal: H0: Average January Return – Average non-January Return <= 0 H1: Average January Return – Average non-January Return > 0 Step 4: Perform an independent one-tailed t-test of difference in means at 5% significance level.
Results obtained	Result: Degree of freedom= 418 t statistics= 0.286964965739269 Critical Value= 1.648507149425946 p-value= 0.3871406696022036
Conclusion	Given that the p-value of 0.387 is greater than the significance level of 0.05, there is insufficient evidence to support the claim that the average return in January is higher than that of the other months.

- ***t*-test of difference in mean returns between September and non-September months;**

***t*-test of difference in mean returns between September and non-September months**	
Test conducted	Step1: To determine if the population is equal, we conduct a Levene's test at 5% significance level with the hypotheses that: H0: Population variance of Sep return = Population variance of non-Sep return H1: Population variance of Sep return != Population variance of non-Sep return Step2: Using scipy.stats.levene, we obtain p-value=0.280. This means that the evidence is insufficient to reject the null hypothesis that Population variance of Sep return = Population variance of non-Sep return. Step 3: Hence, we assume the population variances are unknown but equal H0: Average September Return – Average non-September Return >= 0 H1: : Average September Return – Average non-September Return < 0 Step4: Perform an independent one-tailed t-test of difference in means at 5% significance level.

Results obtained	Degree of freedom=418 t statistics=-1.9200309830619031 Critical Value=1.648507149425946 p-value=0.0277671060025958
Conclusion	Given that the p-value of 0.0278 is smaller than the significance level of 0.05, there is sufficient evidence to support that claim that the average return in September is smaller than that of the other months.

- **an OLS regression of monthly S&P500 returns as dependent variable and two dummy independent variables**

$$R_t = \beta_0 + \beta_1 D_{Jan} + \beta_2 D_{Sep} + \varepsilon_t$$

	Results
OLS Regression	OLS Regression Results == Dep. Variable: mktret R-squared: 0.009 Model: OLS Adj. R-squared: 0.004 Method: Least Squares F-statistic: 1.845 Date: Fri, 26 Feb 2021 Prob (F-statistic): 0.159 Time: 03:15:30 Log-Likelihood: 719.59 No. Observations: 420 AIC: -1433. Df Residuals: 417 BIC: -1421. Df Model: 2 Covariance Type: nonrobust == coef std err t P>|t| [0.025 0.975] -- Intercept 0.0090 0.002 3.850 0.000 0.004 0.014 djan 0.0009 0.008 0.114 0.909 -0.014 0.016 dsep -0.0147 0.008 -1.899 0.058 -0.030 0.001 == Omnibus: 60.413 Durbin-Watson: 1.960 Prob(Omnibus): 0.000 Jarque-Bera (JB): 137.381 Skew: -0.750 Prob(JB): 1.47e-30 Kurtosis: 5.366 Cond. No. 3.85 ============================== Warnings: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. The model is given by: $R_t = \beta_0 + \beta_1 D_{Jan} + \beta_2 D_{Sep} + \varepsilon_t$ Where: Beta0 = 0.0090 Beta1 = 0.0009 Beta2 = 0.0147

	Implications for the hypotheses The intercept beta0 has a value of 0.009 which is the return when the variables djan and dsep are 0 i.e. when the month is not January or September. Beta0 also has a low p-value of 0.00 which means the evidence is sufficient to reject the null hypothesis that beta0 coefficient is 0. The coefficients of djan and dsep are 0.0009 and -0.0147 respectively. For January return, this supports the hypothesis because the market return would be 0.009 + 0.0009*1 = 0.0099 which is higher than returns for other months. For September return, this supports the hypothesis because market return would be 0.009 – 0.0147*1 = -0.0057 which is lower than the returns for other months However, the p-values of djan and dsep are 0.909 and 0.058 which are higher than 0.05. This suggests that the evidence is not sufficient to reject the null hypotheses that the coefficients of these two variables are 0. This causes the coefficient estimates of djan and dsep to be unreliable. **R-squared** The adjusted R-squared is only 0.009 which means only 0.9% of the variance in the S&P500 monthly return is explained by the two dummy variables. **Biases:** The OLS coefficient estimates remain unbiased. However, without adjusting for heteroskedasticity, there is risk of t-statistics of coefficient estimates being biased upwards, causing a risk of rejecting the null hypothesis when it is in fact true. There also exists a slightly positive autocorrelation among the error terms as shown by the Durbin-Watson value of 1.96. Hence, the t-statistics of coefficient estimates are biased upwards, causing a risk of rejecting the null hypothesis when it is in fact true.
White test for heteroskedasticity	White test for heteroskedasticity: 0 1 2 3 0 LM-Statistic LM p-value F-Statistic F-Test p-value 1 0.283866 0.867679 0.141014 0.868518 The White test has a p-value of 0.869 which is larger than the significance level of 0.05. It means that evidence is insufficient to reject the null hypothesis that the error terms have constant variances. Hence there is likely homoskedasticity.

OLS using HC3

OLS regression of mktret against dummy variables under heteroskedasticity-consistent std errors:

OLS Regression Results

Dep. Variable:	mktret	R-squared:	0.009
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	1.574
Date:	Fri, 26 Feb 2021	Prob (F-statistic):	0.208
Time:	04:23:20	Log-Likelihood:	719.59
No. Observations:	420	AIC:	-1433.
Df Residuals:	417	BIC:	-1421.
Df Model:	2		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.0090	0.002	3.900	0.000	0.004	0.014
djan	0.0009	0.008	0.106	0.916	-0.015	0.017
dsep	-0.0147	0.008	-1.758	0.079	-0.031	0.002

Omnibus:	60.413	Durbin-Watson:	1.960
Prob(Omnibus):	0.000	Jarque-Bera (JB):	137.381
Skew:	-0.750	Prob(JB):	1.47e-30
Kurtosis:	5.366	Cond. No.	3.85

Warnings:
[1] Standard Errors are heteroscedasticity robust (HC3)

The OLS estimator with HC3 produces similar result as normal OLS regression as above. However, the z-scores of the coefficients of djan and dsep are adjusted downward compared to normal OLS regression (from 0.114 and -1.899 to 0.106 and -1.758).

This is because HC3 takes into accounts of heteroskedasticity which could inflate the t-statistics of the coefficient estimates.

Despite the fact that the White test ascertains that there is sufficient evidence for homoskedasticity, the variances of the error terms might still be slightly inconsistent. HC3 adjusts for the small heteroskedasticity to deflate the t-statistics of djan and dsep coefficient estimates.

Implication for hypotheses

However, the p-values of djan and dsep are 0.916 and 0.079 which are higher than 0.05. This suggests that the evidence is not sufficient to reject the null hypotheses that the coefficients of these two variables are 0. This causes the coefficient estimates of djan and dsep to be unreliable.

OLS using HAC

OLS regression of mktret against dummy variables under autocorrelation-consistent std errors:

OLS Regression Results

Dep. Variable:	mktret	R-squared:	0.009
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	1.638
Date:	Fri, 26 Feb 2021	Prob (F-statistic):	0.196
Time:	04:23:20	Log-Likelihood:	719.59
No. Observations:	420	AIC:	-1433.
Df Residuals:	417	BIC:	-1421.
Df Model:	2		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0090	0.002	3.842	0.000	0.004	0.014
djan	0.0009	0.008	0.109	0.913	-0.015	0.017
dsep	-0.0147	0.008	-1.790	0.074	-0.031	0.001

Omnibus:	60.413	Durbin-Watson:	1.960
Prob(Omnibus):	0.000	Jarque-Bera (JB):	137.381
Skew:	-0.750	Prob(JB):	1.47e-30
Kurtosis:	5.366	Cond. No.	3.85

Warnings:
[1] Standard Errors are heteroscedasticity and autocorrelation robust (HAC) using 1 lags and without small sample correction

---

OLS regression of mktret against dummy variables under autocorrelation-consistent std errors:

OLS Regression Results

Dep. Variable:	mktret	R-squared:	0.009
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	1.739
Date:	Fri, 26 Feb 2021	Prob (F-statistic):	0.177
Time:	04:23:20	Log-Likelihood:	719.59
No. Observations:	420	AIC:	-1433.
Df Residuals:	417	BIC:	-1421.
Df Model:	2		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0090	0.002	4.062	0.000	0.005	0.013
djan	0.0009	0.008	0.109	0.913	-0.015	0.017
dsep	-0.0147	0.008	-1.853	0.065	-0.030	0.001

Omnibus:	60.413	Durbin-Watson:	1.960
Prob(Omnibus):	0.000	Jarque-Bera (JB):	137.381
Skew:	-0.750	Prob(JB):	1.47e-30
Kurtosis:	5.366	Cond. No.	3.85

Warnings:
[1] Standard Errors are heteroscedasticity and autocorrelation robust (HAC) using 3 lags and without small sample correction

---

OLS regression of mktret against dummy variables under autocorrelation-consistent std errors:

OLS Regression Results

Dep. Variable:	mktret	R-squared:	0.009
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	1.843
Date:	Fri, 26 Feb 2021	Prob (F-statistic):	0.160
Time:	04:23:20	Log-Likelihood:	719.59
No. Observations:	420	AIC:	-1433.
Df Residuals:	417	BIC:	-1421.
Df Model:	2		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0090	0.002	4.284	0.000	0.005	0.013
djan	0.0009	0.008	0.110	0.912	-0.015	0.017
dsep	-0.0147	0.008	-1.905	0.057	-0.030	0.000

Omnibus:	60.413	Durbin-Watson:	1.960
Prob(Omnibus):	0.000	Jarque-Bera (JB):	137.381
Skew:	-0.750	Prob(JB):	1.47e-30
Kurtosis:	5.366	Cond. No.	3.85

Warnings:
[1] Standard Errors are heteroscedasticity and autocorrelation robust (HAC) using 6 lags and without small sample correction

```
--------------------------------------------------------------------------------


OLS regression of mktret against dummy variables under autocorrelation-
consistent std errors:
                            OLS Regression Results
==============================================================================
Dep. Variable:                 mktret   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     1.898
Date:                Fri, 26 Feb 2021   Prob (F-statistic):              0.151
Time:                        04:23:20   Log-Likelihood:                719.59
No. Observations:                 420   AIC:                            -1433.
Df Residuals:                     417   BIC:                            -1421.
Df Model:                           2
Covariance Type:                  HAC
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0090      0.002      4.369      0.000       0.005       0.013
djan           0.0009      0.008      0.110      0.912      -0.015       0.017
dsep          -0.0147      0.008     -1.913      0.056      -0.030       0.000
==============================================================================
Omnibus:                       60.413   Durbin-Watson:                   1.960
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              137.381
Skew:                          -0.750   Prob(JB):                     1.47e-30
Kurtosis:                       5.366   Cond. No.                         3.85
==============================================================================

Warnings:
[1] Standard Errors are heteroscedasticity and autocorrelation robust (HAC)
using 12 lags and without small sample correction

--------------------------------------------------------------------------------
```

The t-statistics of all coefficient estimates increase from 1 to 12 lags.

The OLS estimator using HAC adjusts for both heteroskedasticity and autocorrelation of the error terms. It seems that HAC with lag=1 is more robust/aggressive in eliminating heteroskedasticity and autocorrelation because its t-statistics of the coefficient estimates are the highest, which means there is the lowest risk of rejecting the null hypothesis.

Implication for hypotheses

However, under HAC with lag=1, the p-values of djan and dsep are 0.913 and 0.074 which are higher than 0.05. This suggests that the evidence is not sufficient to reject the null hypotheses that the coefficients of these two variables are 0. This causes the coefficient estimates of djan and dsep to be unreliable.

Conclusion from OLS regression analyses:

The OLS regression analyses do not support the hypotheses even after adjusting for heteroskedasticity and autocorrelation under HC3 and HAC.

The p-values of djan and dsep from all the models above are greater than 0.05, which suggests that the coefficient estimates are statistically insignificant i.e. they are very likely to be 0 which is the null hypothesis. Hence, there is insufficient evidence to support the hypotheses.

Q2

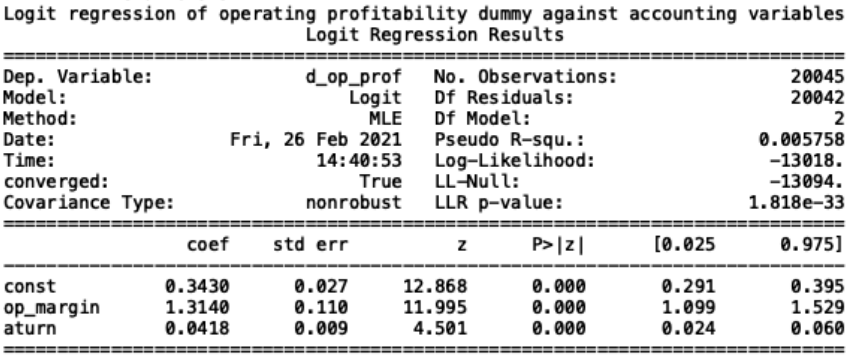
a) Source of data

Source: Compustat
Data: Annual Financial Statements of US Companies
Conditions: “gsecotr” != 40 AND “mktvalt” >= 1,000
Period: Dec 2003 to Dec 2019 (16 years)

b) Test the hypothesis that operating profitability for the year is positively correlated with the previous- year operating profit margin and operating asset turnover, using the three regression models below:

	Result
Pooled OLS	Pooled OLS regression of future operating profitability against accounting variables with HC std errors OLS Regression Results Dep. Variable: op_prof R-squared: 0.003 Model: OLS Adj. R-squared: 0.003 Method: Least Squares F-statistic: 17.99 Date: Fri, 26 Feb 2021 Prob (F-statistic): 1.56e-08 Time: 14:39:31 Log-Likelihood: -20283. No. Observations: 20045 AIC: 4.057e+04 Df Residuals: 20042 BIC: 4.060e+04 Df Model: 2 Covariance Type: HC3 coef std err t P>|t| [0.025 0.975] const 0.0383 0.012 3.165 0.002 0.015 0.062 op_margin 0.2726 0.046 5.966 0.000 0.183 0.362 aturn 0.0090 0.003 2.842 0.004 0.003 0.015 Omnibus: 1970.232 Durbin-Watson: 1.949 Prob(Omnibus): 0.000 Jarque-Bera (JB): 13853.759 Skew: 0.184 Prob(JB): 0.00 Kurtosis: 7.056 Cond. No. 18.5 Warnings: [1] Standard Errors are heteroscedasticity robust (HC3) The model produced is: $$\%\Delta EBIT_{i,t} = \beta_0 + \beta_1 \frac{EBIT_{i,t-1}}{Rev_{i,t-1}} + \beta_2 \frac{Rev_{i,t-1}}{NOA_{i,t-1}} + \varepsilon_{it}$$ Where: Beta0 = 0.0383 Beta1= 0.2726 Beta2= 0.0090 **<u>Implication of hypothesis:</u>** Op_margin and aturn have coefficients of 0.2726 and 0.0090 which means it supports the hypothesis that operating profitability for the year is positively correlated with the previous- year operating profit margin and operating asset turnover.

	The p-values of op_margin and aturn are 0.000 and 0.004 which are lower than 0.05. This suggests that the evidence is sufficient to reject the null hypotheses that the coefficients of these two variables are 0. Hence, the coefficients are reliable. **R-squared** The adjusted R-squared is only 0.003 which means only 0.3% of the variance in the future operating profitability is explained by current operating profit margin and asset turn. **Biases:** However, without adjusting for heteroskedasticity, there is risk of t-statistics of coefficient estimates being biased upwards, causing a risk of rejecting the null hypothesis when it is in fact true. There also exists a positive autocorrelation among the error terms as shown by the Durbin-Watson value of 1.949. Hence, the t-statistics of coefficient estimates are biased upwards, causing a risk of rejecting the null hypothesis when it is in fact true.
Year-by-year regressions	Coefficients of year-by-year regressions const op_margin aturn fyear 2003 0.003484 0.878796 0.006357 2004 0.067969 0.896456 0.012460 2005 0.147332 0.400387 -0.004640 2006 0.085794 0.433101 0.001260 2007 0.064683 0.223379 0.007689 2008 -0.001878 0.409437 0.020506 2009 -0.111508 -0.023692 0.011437 2010 0.042498 1.093718 0.034851 2011 0.062079 0.571334 0.000691 2012 -0.059425 0.190401 0.031559 2013 0.048396 -0.047039 -0.007697 2014 0.118278 0.087991 -0.014723 2015 0.088395 -0.684804 0.013977 2016 0.031582 0.155108 -0.000517 2017 -0.066317 0.775148 0.024280 2018 0.107216 0.257314 -0.016585 2019 0.022601 -0.164744 0.010829 T-statistics and p-values of year-by-year coefficients: t-stat p-value const 2.2898 0.0360 op_margin 2.9806 0.0088 aturn 2.1686 0.0455 **Implication for hypothesis** The year-by-year OLS regression shows that the operating profitability has different correlations with operating profit margin and asset turn for each year. For example, in 2005, op_prof is negatively correlated with aturn with a coefficient of -0.00464 and

	this does not support the hypothesis that op_prof is positively correlated with aturn. Through conducting a 1-sample two-tailed t-test , the p-values of the samples of op_margin and aturn are 0.0088 and 0.0455 which are lower than 0.05. This suggests that the evidence is sufficient to reject the null hypotheses that the year-by-year coefficients estimates of these two variables are 0. The year-by-year sample of intercept const gives a p-value of 0.036. Given a significance level of 5%, it means that the evidence is sufficient to reject the null hypothesis that the year-by-year const coefficient estimates are 0.
Logistic regression	Optimization terminated successfully. Current function value: 0.649450 Iterations 5 Logit regression of operating profitability dummy against accounting variables Logit Regression Results  The model produced is: $$D_{i,t} = \beta_0 + \beta_1 \frac{EBIT_{i,t-1}}{Rev_{i,t-1}} + \beta_2 \frac{Rev_{i,t-1}}{NOA_{i,t-1}} + \varepsilon_{it}$$ Where beta0 = 0.343 Beta1 = 1.314 Beta2 = 0.0418 **<u>Implication for hypothesis</u>** Op_margin and aturn have coefficients of 1.31 and 0.0418 which means it supports the hypothesis that operating profitability for the year is positively correlated with the previous- year operating profit margin and operating asset turnover. The p-values of the coefficient estimates are all 0, which means that the evidence is sufficient to reject the null hypothesis that the coefficients estimates are 0.

Dep. Variable:		d_op_prof	No. Observations:			20045
Model:		Logit	Df Residuals:			20042
Method:		MLE	Df Model:			2
Date:		Fri, 26 Feb 2021	Pseudo R-squ.:			0.005758
Time:		14:40:53	Log-Likelihood:			-13018.
converged:		True	LL-Null:			-13094.
Covariance Type:		nonrobust	LLR p-value:			1.818e-33
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.3430	0.027	12.868	0.000	0.291	0.395
op_margin	1.3140	0.110	11.995	0.000	1.099	1.529
aturn	0.0418	0.009	4.501	0.000	0.024	0.060

	<u>Pseudo R squared:</u> The model has a low pseudo R-squared of 0.00576 which suggests a poor fit and would lead to inaccurate prediction. **<u>Biases:</u>** However**,** heteroskedasticity and auto-correlation on logistic regression model will cause the coefficient estimates to be biased and the impact on t-statistics is unknown. Inferences of statistical significance of coefficient estimates of logit regression model are therefore incorrect.