Econometrics - Collaborative Review Task 2

1. Analysis of the given models:

Part I:

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Constant	0.06906	*0.001575*	0.07629	*0.000336*	0.06969	*0.00148*	0.07697	*0.000312*	0.07374	*0.000307*	0.07472	*0.000264*
x_1	-0.08039	*0.000135*	~	~	-0.08118	*0.000124*	~	~	-0.0813	*<0.0001*	-0.0825	*<0.0001*
x_2	0.23038	*<0.0001*	0.3267	*<0.0001*	0.23188	*<0.0001*	0.32874	*<0.0001*	0.33752	*<0.0001*	0.34071	*<0.0001*
x_3	~	~	0.23239	*<0.0001*	~	~	0.23374	*<0.0001*	0.23387	*<0.0001*	0.2359	*<0.0001*
x_4	~	~	~	~	0.01192	*0.577191*	0.01208	*0.559985*	~	~	0.01796	*0.367086*
R^2	0.3775	~	0.4148	~	0.3785	~	0.4158	~	0.4638	~	0.466	~
F-statistic	42.55	*<0.0001*	69.81	*<0.0001*	39.79	*<0.0001*	46.5	*<0.0001*	56.51	*<0.0001*	42.55	*<0.0001*

From the table, the coefficient is the measure of the mathematical relationship of a particual regressor (x_1, x_2, x_3, x_4) and the dependent variable (y).

The p-value of the coefficient is used to determine if the observed relationship in sample exists in the larger population. The p-value tests the null hypothesis that no correlation exist between the independent and dependent variable against an alternative hypothesis, that the correlation exists.

If the p-value is less than the significant level of 0.05, the independent variable has a relationship with the dependent variable in the population level. If the p-value is greater than the significant level, there is no evidence that there is a relationship with the dependent variable and the coefficient is probably due to a random chance.

The R^2 measures, how good the model explains the dependent variable. A high R^2 means, more information about the dependent variable is captured by the model.

The F-statistic tests the null hypothesis that the model fits better with the intercept alone. The alternative hypothesis states the null hypothesis is not true.

After getting the sense of all the provided information, let's analyse the models to identify the best model.

The F-statistic for all the models provided are way more than the threshold to reject the null hypothesis (p-value < 0.0001, which is less than the significant level of 0.05).

The models with R^2 value < 0.4 (Model 1 and Model 3) can be rejected as they capture less information about the dependent variable.

Feature x_4 has a higher p-value than the significant level in every model to which it is added. (Model 4 and 6). This means that x_4 probably will not be able to generalize the model on a population level and is worth removing from the model. So Model 4 and Model 6 are not the best models.

On comparing Model 2 and Model 5, Model 5 has a 5% increase in the R^2 value on adding x_3.

Note that Model 6 might have a better R^2 than Model 5, but R^2 tends to increase with every new variable added. The alternative is to check the Adjusted R^2, which does not change on adding a variable that does not capture significant information about the dependent variable and sometimes the adjusted R^2 score even penalize when a new variable with no significant information is introduced.

So, we can conclude **Model 5** to be the best model that explains 46.38% of the change in y.

Part II: Likely correlation among the explanatory variables:

x_2 and x_3 are likely to be correlated. It can be inferred from the coefficients of x_2 and x_3 in Model 3 and Model 4. Coefficient of x_2 increases from 0.23188 to 0.32874, when x_3 is introduced to the model. This significant increase in the coefficient is likely to be due to the correlation between the two variables x_2 and x_3.

2. Importing the data provided

```
# importing necessary package
library(tidyverse)
options(pillar.sigfig = 5)

# Loading the CSV and changing the data types
fama_factors_2019 = read.csv(
  "D:/MScFE/610 - Econometrics/Module 2/CRT2/fama_factors_2019.csv",
  col.names = c("date", "MKT-RF", "SMB", "HML", "RF")
                             )
fama_factors_2019$date = as.Date(fama_factors_2019$date)
head(fama_factors_2019)
```

```
##         date MKT.RF  SMB   HML   RF
## 1 2019-01-02   0.23 0.56  1.10 0.01
## 2 2019-01-03  -2.45 0.40  1.21 0.01
## 3 2019-01-04   3.55 0.41 -0.70 0.01
## 4 2019-01-07   0.94 0.97 -0.77 0.01
## 5 2019-01-08   1.01 0.53 -0.64 0.01
## 6 2019-01-09   0.56 0.45  0.09 0.01
```

```
dim(fama_factors_2019)
```

```
## [1] 252   5
```

3. Fama-French 3 factor model equation:

$$E(R_i) = R_f + \beta_{MKT} E(R_m - R_f) + \beta_{SMB} E(SMB) + \beta_{HML} E(HML)$$

where,

$E(R_i)$ = Expected return on the stock

R_f = Risk-free return

$E(R_m - R_f)$ = Difference between expected return of market and the risk-free rate

$E(SMB)$ = Small minus big

$E(HML)$ = High minus low

$\beta_{MKT}, \beta_{SMB}, \beta_{HML}$ = beta factors

4.How the Fama-French model improves upon CAPM.

The Fama-French model is a multi-factor model, which is expanded upon the CAPM model. The Fama-French model adds a size premium (SMB) and a value premium (HML). SMB measures the historic excess of small-cap companies over the large-cap companies. HML represents the spread in returns between the companies that have high book-to-market ratio and the companies that have low book-to-market ratio.

The CAPM model has only one β, which measures the systematic factors, but the Fama-French model has three β coefficients which measures the size factors and value factors along with the systematic factors, thus being more efficient than the CAPM model.

5. Formulate Fama-French regression using the stock's returns, all Fama-French factors and benchmark returns.

We will use Microsoft as the preferred stock to formulate the Fama-French factors. The Fama factors for the year 2019 are already loaded in task 2. Now, let's get the returns of Microsoft stock.

```
# importing necessary libraries
library(tidyquant)
library(data.table)
# Fetching the adjusted closing price of Microsoft stock for the year 2019
msft = data.frame(tq_get("MSFT",
                         from="2019-01-01",
                         to="2020-01-01")[c("date","adjusted")])
head(msft)
```

```
##         date  adjusted
## 1 2019-01-02  97.58066
## 2 2019-01-03  93.99085
## 3 2019-01-04  98.36230
## 4 2019-01-07  98.48776
## 5 2019-01-08  99.20185
## 6 2019-01-09 100.62040
```

```
dim(msft)
```

```
## [1] 252   2
```

```
# Computing daily returns
msft$daily.returns = round(
  (msft$adjusted / shift(msft$adjusted,1) - 1),
  4)
head(msft)
```

```
##         date  adjusted daily.returns
## 1 2019-01-02  97.58066            NA
## 2 2019-01-03  93.99085       -0.0368
## 3 2019-01-04  98.36230        0.0465
## 4 2019-01-07  98.48776        0.0013
## 5 2019-01-08  99.20185        0.0073
## 6 2019-01-09 100.62040        0.0143
```

```
# Joining the Microsoft data with Fama factors
df = na.omit(merge.data.frame(x=msft,
                              y=fama_factors_2019,
                              by="date"))
head(df)
```

```
##         date  adjusted daily.returns MKT.RF  SMB   HML   RF
## 2 2019-01-03  93.99085       -0.0368  -2.45 0.40  1.21 0.01
## 3 2019-01-04  98.36230        0.0465   3.55 0.41 -0.70 0.01
## 4 2019-01-07  98.48776        0.0013   0.94 0.97 -0.77 0.01
## 5 2019-01-08  99.20185        0.0073   1.01 0.53 -0.64 0.01
```

```
## 6 2019-01-09 100.62040        0.0143   0.56 0.45  0.09 0.01
## 7 2019-01-10  99.97385       -0.0064   0.42 0.03 -0.44 0.01
```

Fama French model:

$$r_i - rf = \alpha + \beta_{mkt}(r_m - r_f) + \beta_{smb}(\text{SMB}) + \beta_{hml}(\text{HML}) + \epsilon$$

We will add a new column "$r_i - rf$" to the df and remove the "adjusted" column.

```
# Adding Ri - Rf to the df
df$Ri.Rf = df$daily.returns - df$RF
# Removing columns
df = subset(df, select = -c(date, adjusted, daily.returns, RF))
head(df)
```

```
##   MKT.RF  SMB   HML   Ri.Rf
## 2  -2.45 0.40  1.21 -0.0468
## 3   3.55 0.41 -0.70  0.0365
## 4   0.94 0.97 -0.77 -0.0087
## 5   1.01 0.53 -0.64 -0.0027
## 6   0.56 0.45  0.09  0.0043
## 7   0.42 0.03 -0.44 -0.0164
```

Performing linear regression with $r_i - r_f$ as the dependent variable.

1. CAPM model

```
capm_model = lm(Ri.Rf ~ MKT.RF, data = df)
summary(capm_model)
```

```
##
## Call:
## lm(formula = Ri.Rf ~ MKT.RF, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.031673 -0.003591  0.000289  0.003748  0.032579
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0077814  0.0004834  -16.10   <2e-16 ***
## MKT.RF       0.0121238  0.0005835   20.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007602 on 249 degrees of freedom
## Multiple R-squared:  0.6342, Adjusted R-squared:  0.6328
## F-statistic: 431.8 on 1 and 249 DF,  p-value: < 2.2e-16
```

The equation of the CAPM regression model is

$$r_i - r_f = -0.0078 + 0.0121(r_m - r_f)$$

2. Fama-French model

```
ff_model = lm(Ri.Rf ~ ., data = df)
summary(ff_model)
```

```
##
## Call:
## lm(formula = Ri.Rf ~ ., data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.036596 -0.003180 -0.000090  0.003187  0.028327
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0081190  0.0004156 -19.533  < 2e-16 ***
## MKT.RF       0.0124816  0.0005249  23.781  < 2e-16 ***
## SMB         -0.0048571  0.0009273  -5.238 3.48e-07 ***
## HML         -0.0052448  0.0007143  -7.343 3.03e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00651 on 247 degrees of freedom
## Multiple R-squared:  0.7339, Adjusted R-squared:  0.7307
## F-statistic: 227.1 on 3 and 247 DF,  p-value: < 2.2e-16
```

The equation of the Fama-French regression model is

$$r_i - r_f = -0.0081 + 0.0124(r_m - r_f) - (0.0048 \times \text{SMB}) - (0.0052 \times \text{HML})$$

6. Greek letter used in front of each factor

β_{mkt} is the market beta, which is 1.2% for the model.

β_{smb} is the size beta, which is -0.4% for the model.

β_{hml} is the value beta, which is -0.5% for the model.

7. Which model performed better?

Fama-French model performed better.

The systematic risk alone was able to explain 63.42% (r^2 value) of the returns, but with the addition of size and value factors, the r^2 value increased to 73.39%.

The negative co-efficient of SMB suggests that the expected return on Microsoft stock decreases, as big companies are safer option to invest and is congruent with the fact that the expected returns decreases with less risk. Also, the negative co-efficient of HML means that the expected returns tend to decrease for the growth stocks.

The low p-value implies that the study is statistically significant.