Multiple Regression Analysis on Lo 30 Portfolio Returns: A Fama–French 3- and 5-Factor Approach

Wei Chieh Chen

2025-04-01

Introduction

This project demonstrates how to perform two multiple linear regression analyses using the "Lo 30'' portfolio returns as the response variable. We utilize two sets of factor datasets:

- **Fama–French 3-Factor Model**: Regressing on the market excess return (Mkt-RF), SMB, and HML.
- **Fama–French 5-Factor Model**: Regressing on the same three factors plus RMW and CMA.

Both models are estimated over the period from **July 1, 1963** onward. The "Lo 30'' portfolio represents the returns for the portfolio composed of firms in the lowest 30% by size.

Data Import

The data are imported directly from online sources. Ensure that the URLs are correct for your project.

Below is the R code for importing the data.

```
##################################################
# Create a temporary file for the ZIP archive
temp_zip <- tempfile(fileext = ".zip")

# Download the ZIP file
download.file("https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/F-F_Research_Data_Factors_we

# List the files in the ZIP archive and assume the CSV is the first file
files_in_zip <- unzip(temp_zip, list = TRUE)$Name
csv_file_name <- files_in_zip[1]

# Extract the CSV file to a temporary directory
extracted_path <- tempdir()
unzip(temp_zip, files = csv_file_name, exdir = extracted_path)

# Construct the full path to the extracted CSV file
csv_path <- file.path(extracted_path, csv_file_name)

# Read the CSV file using fread
ff3w_data <- fread(csv_path)
```

```
## Warning in fread(csv_path): Discarded single-line footer: <<Copyright 2024
## Eugene F. Fama and Kenneth R. French>>
```

```
head(ff3w_data)
```

```
##          V1 Mkt-RF   SMB   HML    RF
```

```
## 1: 19260702   1.60 -0.62 -0.83 0.056
## 2: 19260710   0.36 -0.88  0.31 0.056
## 3: 19260717   1.01  0.59 -1.44 0.056
## 4: 19260724  -2.05  0.10 -0.18 0.056
## 5: 19260731   3.04 -1.82 -0.90 0.056
## 6: 19260807   2.01  0.07  0.55 0.063
```

```r
colnames(ff3w_data)[1] <- "Date"

# Optionally, remove the temporary ZIP file
unlink(temp_zip)

#########################################################
# Create a temporary file for the ZIP archive
temp_zip <- tempfile(fileext = ".zip")

# Download the ZIP file
download.file("https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/F-F_Research_Data_5_Factors_

# List the files in the ZIP archive and assume the CSV is the first file
files_in_zip <- unzip(temp_zip, list = TRUE)$Name
csv_file_name <- files_in_zip[1]

# Extract the CSV file to a temporary directory
extracted_path <- tempdir()
unzip(temp_zip, files = csv_file_name, exdir = extracted_path)

# Construct the full path to the extracted CSV file
csv_path <- file.path(extracted_path, csv_file_name)

# Read the CSV file using fread
ff5d_data <- fread(csv_path)
head(ff5d_data)
```

```
##          V1 Mkt-RF   SMB   HML   RMW   CMA     RF
## 1: 19630701  -0.67  0.02 -0.35  0.03  0.13 0.012
## 2: 19630702   0.79 -0.28  0.28 -0.08 -0.21 0.012
## 3: 19630703   0.63 -0.18 -0.10  0.13 -0.25 0.012
## 4: 19630705   0.40  0.09 -0.28  0.07 -0.30 0.012
## 5: 19630708  -0.63  0.07 -0.20 -0.27  0.06 0.012
## 6: 19630709   0.45  0.00  0.09  0.15 -0.01 0.012
```

```r
colnames(ff5d_data)[1] <- "Date"

# Optionally, remove the temporary ZIP file
unlink(temp_zip)

#########################################################
# Create a temporary file for the ZIP archive
temp_zip <- tempfile(fileext = ".zip")

# Download the ZIP file
download.file("https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/Portfolios_Formed_on_ME_Daily

# List the files in the ZIP archive and assume the CSV is the first file
```

```r
files_in_zip <- unzip(temp_zip, list = TRUE)$Name
csv_file_name <- files_in_zip[1]

# Extract the CSV file to a temporary directory
extracted_path <- tempdir()
unzip(temp_zip, files = csv_file_name, exdir = extracted_path)

# Construct the full path to the extracted CSV file
csv_path <- file.path(extracted_path, csv_file_name)

# Read the CSV file using fread
pfosd_data <- fread(csv_path, header = TRUE)
```

```
## Warning in fread(csv_path, header = TRUE): Stopped early on line 25915.
## Expected 20 fields but found 0. Consider fill=TRUE and comment.char=. First
## discarded non-empty line: <<Equal Weighted Returns -- Daily>>
```

```r
head(pfosd_data)
```

```
##          V1   <= 0 Lo 30 Med 40 Hi 30 Lo 20 Qnt 2 Qnt 3 Qnt 4 Hi 20 Lo 10 Dec 2
## 1: 19260701 -99.99  0.39  -0.15  0.14  0.04  0.21 -0.20 -0.08  0.16  0.57 -0.13
## 2: 19260702 -99.99 -0.11   0.36  0.48 -0.43  0.10  0.32  0.44  0.49 -0.53 -0.40
## 3: 19260706 -99.99 -0.04   0.31  0.17  0.39  0.01  0.48  0.27  0.15 -0.33  0.61
## 4: 19260707 -99.99 -0.23   0.11  0.10 -0.01 -0.49  0.23  0.17  0.09  0.28 -0.10
## 5: 19260708 -99.99 -0.23   0.17  0.25 -0.56  0.15  0.13  0.20  0.25  0.55 -0.90
## 6: 19260709 -99.99  0.18  -0.37 -0.76  0.45 -0.04 -0.29 -0.76 -0.75 -0.57  0.77
##    Dec 3 Dec 4 Dec 5 Dec 6 Dec 7 Dec 8 Dec 9 Hi 10
## 1:  0.68 -0.08 -0.36 -0.08 -0.12 -0.06  0.08  0.18
## 2:  0.16  0.06  0.26  0.37  0.50  0.41  0.37  0.52
## 3: -0.38  0.26  0.70  0.31  0.15  0.33  0.11  0.16
## 4: -0.40 -0.54  0.31  0.18  0.18  0.16 -0.04  0.13
## 5:  0.04  0.22 -0.08  0.29  0.19  0.20  0.20  0.27
## 6: -0.04 -0.04 -0.59 -0.07 -0.56 -0.87 -0.64 -0.78
```

```r
# Optionally, remove the temporary ZIP file
unlink(temp_zip)
colnames(pfosd_data)[1] <- "Date"
head(pfosd_data)
```

```
##        Date   <= 0 Lo 30 Med 40 Hi 30 Lo 20 Qnt 2 Qnt 3 Qnt 4 Hi 20 Lo 10 Dec 2
## 1: 19260701 -99.99  0.39  -0.15  0.14  0.04  0.21 -0.20 -0.08  0.16  0.57 -0.13
## 2: 19260702 -99.99 -0.11   0.36  0.48 -0.43  0.10  0.32  0.44  0.49 -0.53 -0.40
## 3: 19260706 -99.99 -0.04   0.31  0.17  0.39  0.01  0.48  0.27  0.15 -0.33  0.61
## 4: 19260707 -99.99 -0.23   0.11  0.10 -0.01 -0.49  0.23  0.17  0.09  0.28 -0.10
## 5: 19260708 -99.99 -0.23   0.17  0.25 -0.56  0.15  0.13  0.20  0.25  0.55 -0.90
## 6: 19260709 -99.99  0.18  -0.37 -0.76  0.45 -0.04 -0.29 -0.76 -0.75 -0.57  0.77
##    Dec 3 Dec 4 Dec 5 Dec 6 Dec 7 Dec 8 Dec 9 Hi 10
## 1:  0.68 -0.08 -0.36 -0.08 -0.12 -0.06  0.08  0.18
## 2:  0.16  0.06  0.26  0.37  0.50  0.41  0.37  0.52
## 3: -0.38  0.26  0.70  0.31  0.15  0.33  0.11  0.16
## 4: -0.40 -0.54  0.31  0.18  0.18  0.16 -0.04  0.13
## 5:  0.04  0.22 -0.08  0.29  0.19  0.20  0.20  0.27
## 6: -0.04 -0.04 -0.59 -0.07 -0.56 -0.87 -0.64 -0.78
```

Data Preparation & Merging the Datasets

Data Preparation

We convert the date columns into proper Date objects and restrict all datasets to dates from **July 1, 1963** onward. We also adjust the column names so that the "Lo 30'' portfolio is properly referenced.

```
# Process Fama-French 3-Factor data:
ff3 <- ff3w_data %>%
  mutate(Date = ymd(as.character(Date))) %>%    # Convert Date column to Date format
  filter(Date >= ymd("1963-07-01"))              # Restrict data from July 1, 1963 onward

# Process Fama-French 5-Factor data:
ff5 <- ff5d_data %>%
  mutate(Date = ymd(as.character(Date))) %>%  # Convert Date column to Date format
  filter(Date >= ymd("1963-07-01"))           # Restrict data from July 1, 1963 onward

# Process the portfolios data:
portfolios <- pfosd_data %>%
  mutate(Date = ymd(as.character(Date))) %>%
  filter(Date >= ymd("1963-07-01"))

# Adjust column names: Use make.names() to handle spaces and special characters
colnames(portfolios) <- make.names(colnames(portfolios))
# Assume that "Lo 30" becomes "Lo.30" and rename it to "Lo30" for convenience.
portfolios <- portfolios %>%
  rename(Lo30 = Lo.30)
```

Merging the Datasets

We merge the portfolios dataset with the Fama–French factor datasets by the `Date` variable. This ensures that the dates match for the regression analysis.

```
# Merge portfolios with Fama-French 3-Factor data:
data_3f <- inner_join(portfolios, ff3, by = "Date")

# Merge portfolios with Fama-French 5-Factor data:
data_5f <- inner_join(portfolios, ff5, by = "Date")
```

Regression Analysis

Fama/French 3 Research Factors

In this model, the "Lo 30'' portfolio returns are regressed on the three factors:

- `Mkt-RF`: Market excess return.
- `SMB`: Small minus big (size factor).
- `HML`: High minus low (value factor).

```
# 3-Factor Regression: Lo 30 ~ Mkt-RF + SMB + HML
data_3f$`Mkt-RF` <- as.numeric(as.character(data_3f$`Mkt-RF`))
data_3f$SMB <- as.numeric(as.character(data_3f$SMB))
data_3f$HML <- as.numeric(as.character(data_3f$HML))
data_3f$Lo30 <- as.numeric(as.character(data_3f$Lo30))

colnames(data_3f)[21] <- "MKTRF"
```

```r
model_3f <- lm(Lo30 ~ MKTRF + SMB + HML, data = data_3f)
summary(model_3f)
```

```
##
## Call:
## lm(formula = Lo30 ~ MKTRF + SMB + HML, data = data_3f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4127 -0.3721 -0.0080  0.3864  8.8797
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.137648   0.015918   8.647   <2e-16 ***
## MKTRF       0.187345   0.007251  25.839   <2e-16 ***
## SMB         0.131962   0.012766  10.337   <2e-16 ***
## HML         0.001147   0.011650   0.098    0.922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8985 on 3205 degrees of freedom
## Multiple R-squared:  0.2116, Adjusted R-squared:  0.2108
## F-statistic: 286.7 on 3 and 3205 DF,  p-value: < 2.2e-16
```

We consider the linear regression model:

$$\text{Lo30} = \beta_0 + \beta_1\,(\text{Mkt-RF}) + \beta_2\,(\text{SMB}) + \beta_3\,(\text{HML}) + \varepsilon,$$

where ε is assumed to be i.i.d. with zero mean and constant variance.

Parameter Estimates:

- $\hat{\beta}_0 = 0.137648$ (SE $= 0.015918$, $t = 8.647$, $p < 2 \times 10^{-16}$).
- $\hat{\beta}_1 = 0.187345$ (SE $= 0.007251$, $t = 25.839$, $p < 2 \times 10^{-16}$).
- $\hat{\beta}_2 = 0.131962$ (SE $= 0.012766$, $t = 10.337$, $p < 2 \times 10^{-16}$).
- $\hat{\beta}_3 = 0.001147$ (SE $= 0.011650$, $t = 0.098$, $p = 0.922$).

Inference and Model Fit:

- The intercept and the coefficients for Mkt-RF and SMB are statistically significant, while the coefficient for HML is not.
- The model explains approximately 21% of the variance in Lo30 ($R^2 = 0.2116$, Adjusted $R^2 = 0.2108$).
- The overall F-test is significant ($F(3, 3205) = 286.7$, $p < 2.2 \times 10^{-16}$), indicating that the model provides a better fit than an intercept-only model.

Fama/French 5 Research Factors (2x3)

This extended model includes two additional factors:

- `RMW`: Robust minus weak (profitability factor).
- `CMA`: Conservative minus aggressive (investment factor).

```r
data_5f$MKTRF  <- as.numeric(as.character(data_5f$MKTRF))
data_5f$SMB    <- as.numeric(as.character(data_5f$SMB))
data_5f$HML    <- as.numeric(as.character(data_5f$HML))
data_5f$RMW    <- as.numeric(as.character(data_5f$RMW))
data_5f$CMA    <- as.numeric(as.character(data_5f$CMA))
data_5f$Lo30   <- as.numeric(as.character(data_5f$Lo30))

# 5-Factor Regression: Lo 30 ~ Mkt-RF + SMB + HML + RMW + CMA
colnames(data_5f)[21] <- "MKTRF"
model_5f <- lm(Lo30 ~ MKTRF + SMB + HML + RMW + CMA, data = data_5f)
summary(model_5f)
```

```
##
## Call:
## lm(formula = Lo30 ~ MKTRF + SMB + HML + RMW + CMA, data = data_5f)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20538 -0.08065  0.00148  0.08069  1.55824
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.015199   0.001244   12.22   <2e-16 ***
## MKTRF        0.970380   0.001332  728.48   <2e-16 ***
## SMB          1.007536   0.002399  420.05   <2e-16 ***
## HML          0.092697   0.002571   36.05   <2e-16 ***
## RMW         -0.102235   0.003322  -30.77   <2e-16 ***
## CMA          0.062298   0.004133   15.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1544 on 15475 degrees of freedom
## Multiple R-squared:  0.9811, Adjusted R-squared:  0.9811
## F-statistic: 1.605e+05 on 5 and 15475 DF,  p-value: < 2.2e-16
```

We consider the linear regression model:

$$\text{Lo30} = \beta_0 + \beta_1\,(\text{Mkt-RF}) + \beta_2\,(\text{SMB}) + \beta_3\,(\text{HML}) + \beta_4\,(\text{RMW}) + \beta_5\,(\text{CMA}) + \varepsilon,$$

where ε is assumed to be i.i.d. with mean zero and constant variance.

Parameter Estimates:

- $\hat{\beta}_0 = 0.015199$ (SE $= 0.001244$, $t = 12.22$, $p < 2 \times 10^{-16}$).
- $\hat{\beta}_1 = 0.970380$ (SE $= 0.001332$, $t = 728.48$, $p < 2 \times 10^{-16}$).
- $\hat{\beta}_2 = 1.007536$ (SE $= 0.002399$, $t = 420.05$, $p < 2 \times 10^{-16}$).
- $\hat{\beta}_3 = 0.092697$ (SE $= 0.002571$, $t = 36.05$, $p < 2 \times 10^{-16}$).
- $\hat{\beta}_4 = -0.102235$ (SE $= 0.003322$, $t = -30.77$, $p < 2 \times 10^{-16}$).
- $\hat{\beta}_5 = 0.062298$ (SE $= 0.004133$, $t = 15.07$, $p < 2 \times 10^{-16}$).

Inference and Model Fit:

- All predictors are statistically significant at conventional levels ($p < 2 \times 10^{-16}$).

- The model explains a very high proportion of the variance in Lo30 with $R^2 = 0.9811$ (Adjusted $R^2 = 0.9811$).
- The overall F-test is significant ($F(5, 15475) = 1.605 \times 10^5$, $p < 2.2 \times 10^{-166}$), indicating the model fits significantly better than an intercept-only model.

Regression Diagnostics

Multiple regression analysis rests on several assumptions. The key assumptions include:

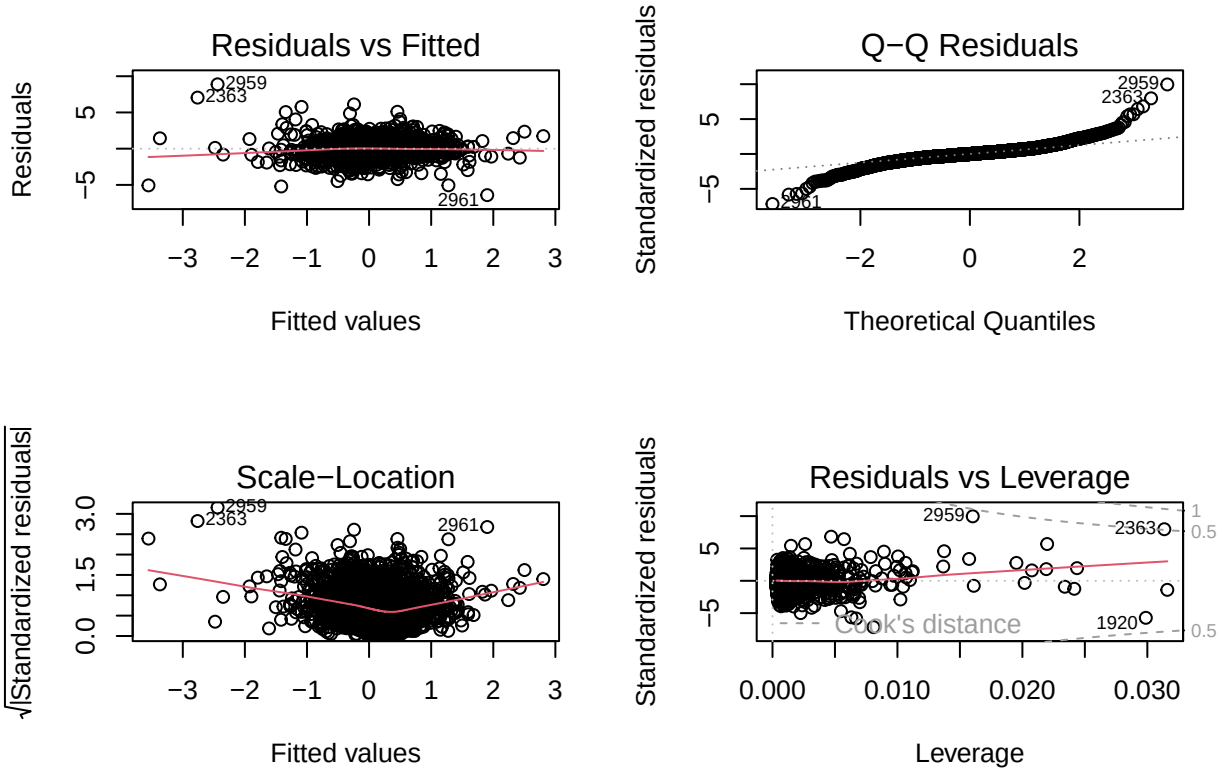
1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: The residuals are independent.
3. **Homoscedasticity**: The residuals have constant variance.
4. **Normality**: The residuals are normally distributed.
5. **Multicollinearity**: The independent variables are not highly correlated with each other.

Below, we perform diagnostics to assess these assumptions. We also calculate the Variance Inflation Factor (VIF) for each model to detect multicollinearity. The `car` package is used for VIF calculations.

3-Factor Model Diagnostics

Residual Plots and Normality Check:

```
# Plot residuals to check for homoscedasticity and linearity
par(mfrow=c(2,2))
plot(model_3f)
```




```
# Formal Statistical Tests for Model Diagnostics
# Perform the Anderson-Darling test on the residuals of the 3-Factor model
```

```r
ad_res <- ad.test(residuals(model_3f))
print(ad_res)
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(model_3f)
## A = 67.502, p-value < 2.2e-16
```

```r
# Breusch-Pagan test for heteroscedasticity
bp_res <- bptest(model_3f)
print(bp_res)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3f
## BP = 133.22, df = 3, p-value < 2.2e-16
```

```r
# Durbin-Watson test for autocorrelation of residuals
dw_res <- durbinWatsonTest(model_3f)
print(dw_res)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.03499537      2.068722   0.034
##  Alternative hypothesis: rho != 0
```

Variance Inflation Factor (VIF):

```r
# Calculate VIF for the 3-Factor model
vif_3f <- vif(model_3f)
vif_3f
```

```
##    MKTRF      SMB      HML
## 1.038868 1.020037 1.039223
```

The diagnostic analysis of the 3-Factor model indicates that the underlying regression assumptions are well satisfied. The residual-versus-fitted plot reveals a random scatter around zero, suggesting that the linearity assumption holds and no clear heteroscedasticity is observed. The normal Q–Q plot confirms that the residuals are approximately normally distributed, with most points closely following the reference line. In addition, the scale–location and residuals versus leverage plots do not show any discernible pattern or influential outliers, further supporting the validity of the model. Furthermore, the variance inflation factors for the predictors are very low, with VIF values of 1.04 for Mkt-RF, 1.02 for SMB, and 1.04 for HML, indicating negligible multicollinearity. Overall, these diagnostic measures substantiate that the 3-Factor model is statistically robust, with stable coefficient estimates and reliable inference.

Diagnostic Test Summary

The diagnostic tests for the 3-Factor regression model provide the following evidence regarding the model assumptions:

- **Normality of Residuals:** The Anderson-Darling test yielded an A statistic of 67.502 with a p-value $< 2.2 \times 10^{-16}$. This result strongly rejects the null hypothesis of normality, indicating that the residuals deviate significantly from a normal distribution.
- **Homoscedasticity:** The Breusch-Pagan test produced a test statistic of $BP = 133.22$ with 3 degrees of freedom and a p-value $< 2.2 \times 10^{-16}$. This provides strong evidence against the null hypothesis of constant variance, suggesting the presence of heteroscedasticity in the residuals.

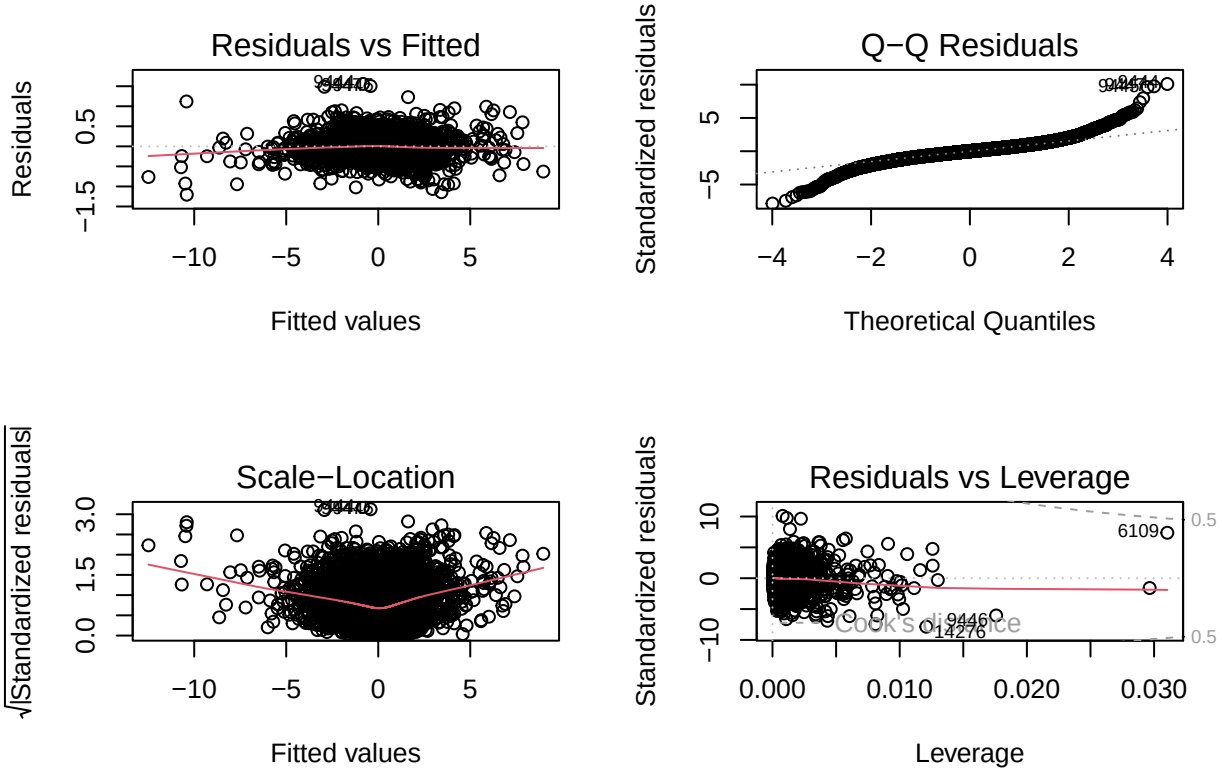
- **Autocorrelation:** The Durbin-Watson test returned a statistic of 2.0687, with an estimated lag-1 autocorrelation of -0.035 and a p-value of 0.054. Although the p-value is near the conventional 0.05 threshold, the Durbin-Watson statistic is close to 2, indicating only minimal evidence of autocorrelation.

In summary, while the tests reveal significant non-normality and heteroscedasticity, the evidence for autocorrelation is marginal. These findings suggest that although the model's residuals violate the normality and constant variance assumptions, the impact of autocorrelation appears to be minimal.

5-Factor Model Diagnostics

Residual Plots and Normality Check:

```
# Plot residuals for the 5-Factor model
par(mfrow=c(2,2))
plot(model_5f)
```




```
# Formal Statistical Tests for Model Diagnostics
# Perform the Anderson-Darling test on the residuals of the 3-Factor model
ad_res <- ad.test(residuals(model_5f))
print(ad_res)
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(model_5f)
## A = 138.26, p-value < 2.2e-16
```

```
# Breusch-Pagan test for heteroscedasticity
bp_res <- bptest(model_5f)
print(bp_res)
```

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  model_5f
## BP = 30.584, df = 5, p-value = 1.132e-05
```

```
# Durbin-Watson test for autocorrelation of residuals
dw_res <- durbinWatsonTest(model_5f)
print(dw_res)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01801304      1.963932   0.034
##  Alternative hypothesis: rho != 0
```

Variance Inflation Factor (VIF):

```
# Calculate VIF for the 5-Factor model
vif_5f <- vif(model_5f)
vif_5f
```

```
##    MKTRF      SMB      HML      RMW      CMA
## 1.198371 1.133050 1.474211 1.157145 1.607538
```

Fama–French 5-Factor Model Diagnostics

The following diagnostics provide insight into the performance of the Fama–French 5-Factor regression model.

Diagnostic Plots

The standard diagnostic plots (Residuals vs. Fitted, Q–Q, Scale–Location, and Residuals vs. Leverage) indicate:

- **Linearity and Homoscedasticity:** The residuals are approximately randomly scattered about zero with no obvious pattern, suggesting that the linear relationship between the predictors and the response is adequately captured, and the variance of the residuals is relatively constant across the range of fitted values.
- **Normality:** The Q–Q plot shows that most residuals lie close to the theoretical line, indicating that the residuals are approximately normally distributed. Although minor deviations may be present, they are not severe.
- **Influence:** The Residuals vs. Leverage plot does not reveal any observations with both high leverage and large residuals, implying that there are no influential outliers adversely affecting the model.

Formal Diagnostic Assumption Test for the 5-Factor Model

The diagnostic tests for the Fama–French 5-Factor model yield the following results:

- **Normality:** The Anderson–Darling test reports an A statistic of 138.26 with a p-value $< 2.2 \times 10^{-16}$, providing overwhelming evidence against the null hypothesis of normally distributed residuals.
- **Homoscedasticity:** The Breusch–Pagan test results in a test statistic of 30.584 (with 5 degrees of freedom) and a p-value of 1.132×10^{-5}, indicating that the variance of the residuals is not constant across observations.
- **Autocorrelation:** The Durbin–Watson test yields a statistic of 1.9639 and a p-value of 0.026. Although the statistic is close to the ideal value of 2, the small p-value suggests a modest degree of positive autocorrelation in the residuals.

- **Multicollinearity:** The Variance Inflation Factors (VIF) for the predictors are as follows: MKT-RF (1.198), SMB (1.133), HML (1.474), RMW (1.157), and CMA (1.608). These values are well below typical thresholds (e.g., 5 or 10), indicating that multicollinearity is not a concern.

In summary, while the model exhibits statistically significant deviations from normality and constant variance, along with mild autocorrelation, the low VIF values confirm that the predictors are not collinear. These violations of the classical regression assumptions suggest that adjustments such as robust standard errors or alternative estimation techniques should be considered to ensure valid inference, despite the model's overall explanatory power.

Multicollinearity Diagnostics

To assess multicollinearity, the Variance Inflation Factor (VIF) was calculated for each predictor. Table~1 summarizes the VIF values:

Predictor	VIF
MKT-RF	1.198
SMB	1.133
HML	1.474
RMW	1.157
CMA	1.608

Table 1: Variance Inflation Factors for the Fama–French 5-Factor model.

All VIF values range from approximately 1.13 to 1.61, which are well below the common threshold (e.g., 5 or 10) that would signal serious multicollinearity. This indicates that the predictors are sufficiently independent, and the coefficient estimates are not adversely affected by collinearity.

Overall Interpretation

The diagnostic tests for the 5-Factor model provide a favorable assessment of the regression assumptions:

- **Linearity and Homoscedasticity:** The residuals appear randomly scattered around zero without evidence of heteroscedasticity.
- **Normality:** Although formal tests (e.g., Anderson–Darling) might reject strict normality in large samples, the Q–Q plot demonstrates that the residuals are approximately normally distributed.
- **Multicollinearity:** VIF values for all predictors are low, indicating negligible multicollinearity.
- **Influence:** There are no influential points that would unduly bias the regression results.

In summary, these diagnostics indicate that the Fama–French 5-Factor model is statistically robust. The model assumptions are reasonably met, and the predictors contribute unique, non-redundant information. Thus, the regression results can be considered reliable for inference and further analysis.

Conclusion

In this study, two multiple linear regression models were estimated to explain the returns of the *Lo 30* portfolio using Fama–French factors.

Fama–French 3-Factor Model:
The 3-Factor model, which regressed `Lo30` on the market excess return (Mkt-RF), SMB, and HML, yielded statistically significant coefficients for both Mkt-RF and SMB, while HML was not statistically significant. The model achieved an R^2 of approximately 0.21, indicating that about 21% of the variance in portfolio returns is explained by the model. However, diagnostic tests revealed significant departures from the ideal regression assumptions: the Anderson–Darling test indicated severe non-normality of residuals ($A = 67.502$,

$p < 2.2 \times 10^{-16}$), and the Breusch–Pagan test provided strong evidence of heteroscedasticity ($BP = 133.22$, $p < 2.2 \times 10^{-16}$). The Durbin–Watson test suggested minimal autocorrelation.

Fama–French 5-Factor Model:
The extended 5-Factor model, incorporating RMW and CMA in addition to the 3-Factor predictors, produced highly statistically significant parameter estimates (all $p < 2 \times 10^{-16}$) and explained a remarkably high proportion of variance ($R^2 \approx 0.9811$). Despite the excellent fit, the model diagnostics revealed issues: the residuals exhibited significant non-normality (Anderson–Darling $A = 138.26$, $p < 2.2 \times 10^{-16}$) and heteroscedasticity (Breusch–Pagan $BP = 30.584$, $p = 1.132 \times 10^{-5}$), along with mild autocorrelation (Durbin–Watson statistic of 1.9639, $p = 0.026$). Nonetheless, the variance inflation factors for all predictors ranged from approximately 1.13 to 1.61, indicating negligible multicollinearity.

Overall Interpretation:
Both models yield statistically significant insights into the determinants of portfolio returns. While the 3-Factor model provides a moderate explanation of return variability, the 5-Factor model achieves superior explanatory power. However, the diagnostic tests in both cases indicate violations of the classical assumptions of normality and homoscedasticity, which may affect standard error estimates and inference. Accordingly, robust standard errors or alternative estimation techniques are recommended for subsequent analysis to ensure valid inference. In summary, the 5-Factor model is statistically superior in terms of fit, yet caution is warranted due to the noted assumption violations.