Linear Regression in Finance

Riccardo Baudone

June 2025

1 Introduction

In this project, I applied Linear Regression to the financial domain, incrementally adding factors to the regression model to enhance its explanatory power and identify the key variables contributing to the prediction of the target variable, Apple's monthly risk premium. The goal is to evaluate how different factors influence the risk premiums of the stock and determine the most significant contributors.

The specific regressions applied in this study include:

A regression on the market's monthly risk premium, aimed at testing the validity of the Capital Asset Pricing Model (**CAPM**) as a model for capturing no-arbitrage stock returns;

A regression on the **three Fama-French factors**, which include the Market Risk Premium, SMB (Small Minus Big), and HML (High Minus Low), to investigate the model's explanatory power in asset pricing;

A regression on the **five Fama-French factors**, which extend the three-factor model by adding the RMW (Robust Minus Weak) and CMA (Conservative Minus Aggressive) factors;

A regression on the five Fama-French factors, supplemented with the monthly returns of **stock portfolios** clustered by sector, to assess the impact of sector-based diversification on the model's explanatory ability.

2 Data Analysis

2.1 Data Retrieval

The data for the **Market Risk Premium** (considered as monthly risk premiums for SP&500 returns), **Apple's Risk Premium**, and the **clustered returns** were all fetched via the `yahoo finance` API. The returns were calculated as the returns of successive price values for each month, ensuring the dataset reflects the monthly fluctuations of the asset prices. The risk-free rate, represented by the 13-week U.S. Treasury bill yield, was subsequently converted into the monthly equivalent yield rate.

The **five Fama-French factors** were retrieved from their official website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`), where they are made publicly available.

After gathering the necessary data, a final dataframe was constructed, which comprehensively includes all the variables month by month, enabling a structured and consistent analysis of the relationships between the variables across time.

2.2 Data Analysis

First, we display the density plots of Apple's Risk Premium (in blue) and the Market Risk Premium (in red). As shown in the plots, it is evident that Apple's Risk Premium has been more positive than the Market Risk Premium in the observed period, but at the same time its shifts have been more extreme.

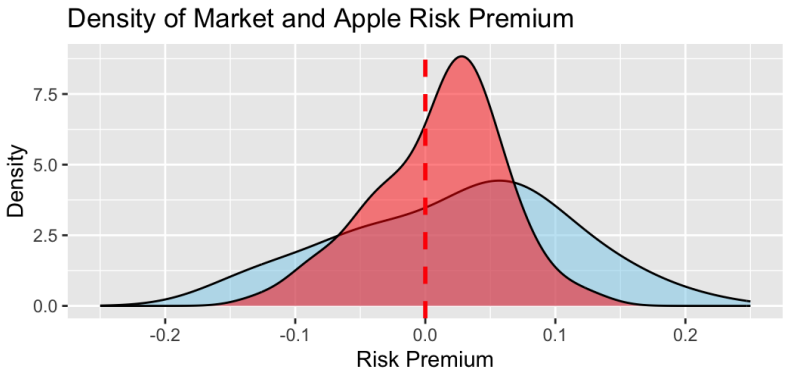


Figura 1: Density of Risk Premium Returns

Next, we display the density plots of the returns of the stock clusters and the cumulative returns of the various portfolios, alongside Market and Apple's. This allows for a clear comparison of how the returns of the individual clusters and the portfolios align with the broader market and Apple's performance.

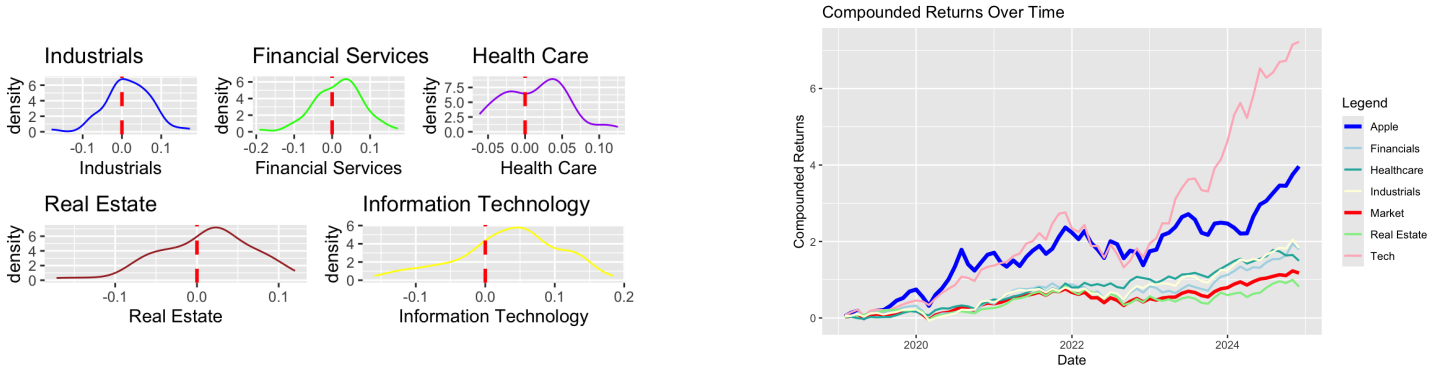


Figura 2: Stock Cluster Plots

Next, we investigate the correlation between the various data points, including the five Fama-French factors and the stock clusters.

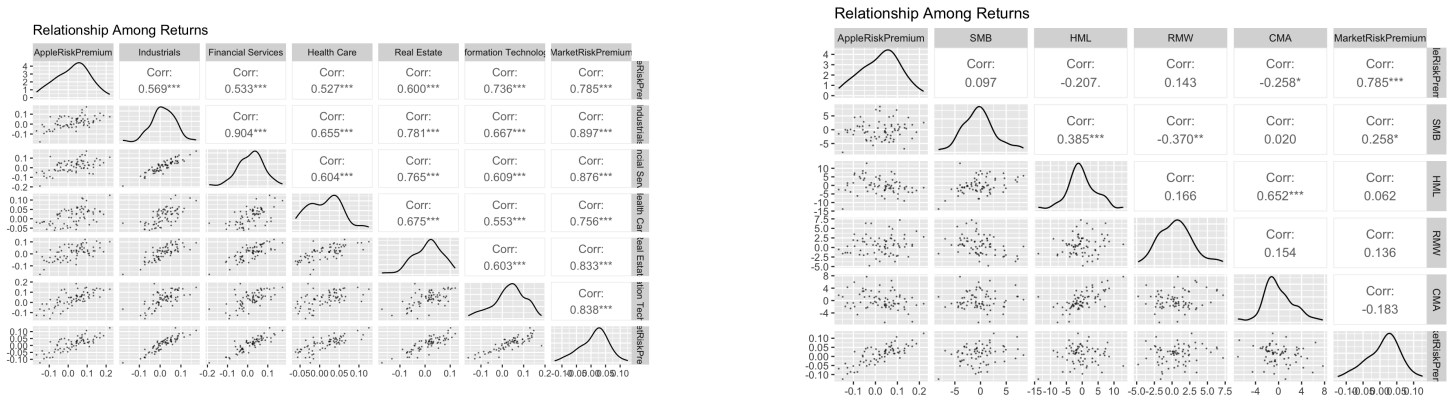


Figura 3: Correlation Plots

From a visual inspection of the plots, one hypothesis that can be tested is the low significance of the Fama-French factors, particularly in relation to the stock clusters. Furthermore, there seems to be reciprocal influence between the stock clusters, suggesting potential multicollinearity or overlapping information between them. This interaction between the stock clusters warrants further exploration and statistical testing.

3 Linear Regression: Mathematics & Methodology

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is calculated by minimizing the sum of squared residuals, which is the difference between the observed values and the predicted values. The simple linear regression model is expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where:

- y is the dependent variable,
- β_0 is the intercept,
- β_1 is the slope or coefficient of the independent variable x,
- ϵ is the error term.

For multiple linear regression, the model becomes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Where $x_1, x_2, \ldots, x_p$ are the independent variables, and $\beta_1, \beta_2, \ldots, \beta_p$ are the corresponding coefficients.

Interpretation of the Model

The R^2 (R-squared) value measures the proportion of the variance in the dependent variable that is explained by the independent variables. It is calculated as:

$$R^2 = 1 - \frac{\sum(\hat{e}_i)^2}{\sum(y_i - \bar{y})^2}$$

Where:

- y_i are the observed values,
- $\hat{e}_i$ are the residuals,
- $\bar{y}$ is the mean of the observed values.

The Adjusted R^2 accounts for the number of predictors in the model and is calculated as:

$$R^2_{\text{adj}} = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-p-1}$$

Where n is the number of observations, and p is the number of predictors.

Significance Testing of Parameters

The significance of the regression coefficients is tested using the t-test, where the null hypothesis is that a coefficient is equal to zero. The t-statistic is calculated as:

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Where:

- $\hat{\beta}_i$ is the estimated coefficient,
- $SE(\hat{\beta}_i)$ is the standard error of the estimated coefficient.

We can demonstrate that this quantity follows a Student's t-distribution with $n - r - 1$ degrees of freedom, allowing us to set the test region at a significance level of α. The p-value is used to test the null hypothesis that the coefficient is zero. If the p-value is smaller than the significance level (typically 0.05), the null hypothesis is rejected, indicating that the coefficient is statistically significant.

Improving the Model

Assumptions to Verify

- **Normality of Residuals:** The residuals should follow a normal distribution. This can be tested using the **Shapiro-Wilk test**, with the null hypothesis being that the residuals are normally distributed.

$$H_0 : \text{Residuals are normally distributed}$$

- **Homoscedasticity:** The variance of residuals should be constant across all levels of the independent variables. This can be tested using the **Breusch-Pagan test**, with the null hypothesis being that the residual variance is constant.

$$H_0 : \text{Constant variance of residuals}$$

- **Independence of Residuals:** The residuals should be independent. This can be tested using the Durbin-Watson test, with the null hypothesis being that there is no autocorrelation in the residuals of successive values, under the assumption of normality.

$$H_0 : \text{No autocorrelation of residuals for successive values}$$

Multicollinearity

Variance Inflation Factor (VIF) is used to detect multicollinearity, which occurs when independent variables are highly correlated. VIF is calculated as:

$$VIF(\beta_i) = \frac{1}{1 - R_i^2}$$

Where R_i^2 is the R-squared value obtained by regressing x_i on all other predictors.

Model Optimization

The Akaike Information Criterion (AIC) is used for model optimization, balancing goodness-of-fit with the complexity of the model. It is calculated as:

$$AIC = 2k - 2\ln(L)$$

Where:

- k is the number of parameters in the model,
- L is the likelihood of the model.

Lower AIC values indicate a better model.

Outliers and Leverage Points

Leverage Points:
Leverage points are data points that have extreme values for the predictors, which can disproportionately influence the regression model. These points can be identified using the hat matrix, where a leverage value h_{ii} greater than $\frac{2(r+1)}{n}$ (where r is the number of predictors and n is the number of observations) indicates a leverage point. Specifically, h_{ii} is the element on the diagonal of the hat matrix.

Outliers:
Outliers are observations whose response values deviate significantly from the values predicted by the regression model. These can be identified by standardizing the residuals. Any residual with an absolute value greater than 2 is typically considered an outlier.

Cook's Distance:
Cook's distance measures the influence of each data point on the estimated regression coefficients. It combines both the leverage and residual for each observation. Cook's distance is calculated as:

$$D_i = \frac{(\beta_i - \beta)^T (Z^T Z)(\beta_i - \beta)}{S^2(r+1)}$$

Where:

- β_i is the regression coefficient vector when the i-th observation is excluded,
- β is the regression coefficient vector from the full model,
- $Z^T Z$ is the information matrix,
- S^2 is the mean squared error (MSE) of the regression,
- r is the number of predictors in the model.

Cook's distance helps identify influential points, with larger values suggesting greater influence on the model.

Transformations

The Box-Cox transformation is applied to the dependent variable to make the residuals more normal. The transformation is defined as:

$$y(\lambda) = \frac{y^\lambda - 1}{\lambda}, \text{ for } \lambda \neq 0$$

When $\lambda = 0$, the transformation is equivalent to the natural logarithm:

$$y(0) = \ln(y)$$

The Box-Cox transformation helps stabilize variance and make the data more normally distributed. It is defined only for positive values, but its validity has been demonstrated for negative values shifted to positive by a costant λ_2.

4 CAPM

The **Capital Asset Pricing Model (CAPM)** is a single-factor model used to explain the risk premium of an asset, based on several assumptions: a closed economy, a market with no transaction fees or costs, individuals who buy quantities of assets to maximize their expected utility, perfectly divisible and liquid assets, and, importantly, portfolios that lie on the efficient frontier. The CAPM links the asset's risk premium to the market's risk premium through the following formula:

$$E(R_i) - R_f = \beta_i \left(E(R_m) - R_f\right)$$

Where:

- $E(R_i) - Rf$ is the asset risk premium i;
- β_i is the asset's **beta**, representing the asset's sensitivity to the market's returns;
- $E(R_m) - R_f$ is the **market risk premium**, which is the excess return of the market over the risk-free rate.

The **beta** (β_i) is a measure of how much the asset's returns move relative to the market returns. A **beta greater than 1** indicates that the asset is more volatile than the market, while a **beta less than 1** indicates that the asset is less volatile. In other words, the beta represents the **systematic risk** that cannot be diversified away, directly linking the asset's risk to the broader market risk.

This relationship can also be empirically validated by performing a **linear regression** on the risk premium of the asset in question, using the market risk premium as the independent variable. In this case, the market risk premium is represented by the risk premium of the **S&P 500 index**, a broad basket of stocks that includes Apple, the stock under consideration in this analysis.

Applying such regression, we obtain the following results :

Variable	**Estimate**	**Standard Error**	**t-value**	**p-value**
Intercept	0.010503	0.006587	1.595	0.115
MarketRiskPremium	1.367443	0.130079	10.512	5.71e-16 ***

Tabella 1: Regression Results

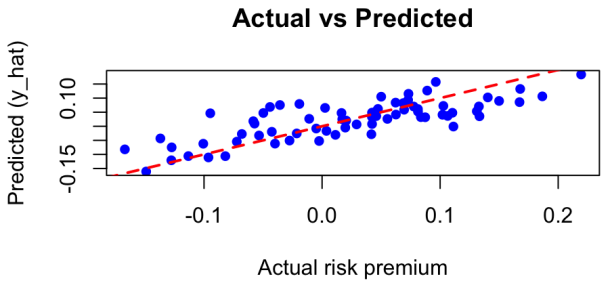


Figura 4: CAPM Regression Line

- Multiple R-squared: 0.6156
- Adjusted R-squared: 0.6101
- F-statistic 110.5 on 1 and 69 DF, p-value: 5.711e-16

We observe strong evidence to assume that the **Beta coefficient** from the regression (and thus the value of the covariance/variance) is significantly different from zero and, therefore, significant in our model. Furthermore, the estimated Beta is equal to **1.36**, indicating that the asset moves in the same direction as the market. The average value of the intercept, i.e., the **alpha** of the CAPM, is close to zero, but the statistical test does not provide sufficient evidence to reject the null hypothesis that the coefficient is equal to zero.

4.1 Residuals

Vediamo adesso se il modello verifica le ipotesi sui residui :

Test	**p-value**
Shapiro-Wilk Test	0.406
Breusch-Pagan Test	0.1887
Durbin-Watson Test (DW = 1.4139)	0.006044

Tabella 2: Results of Statistical Tests

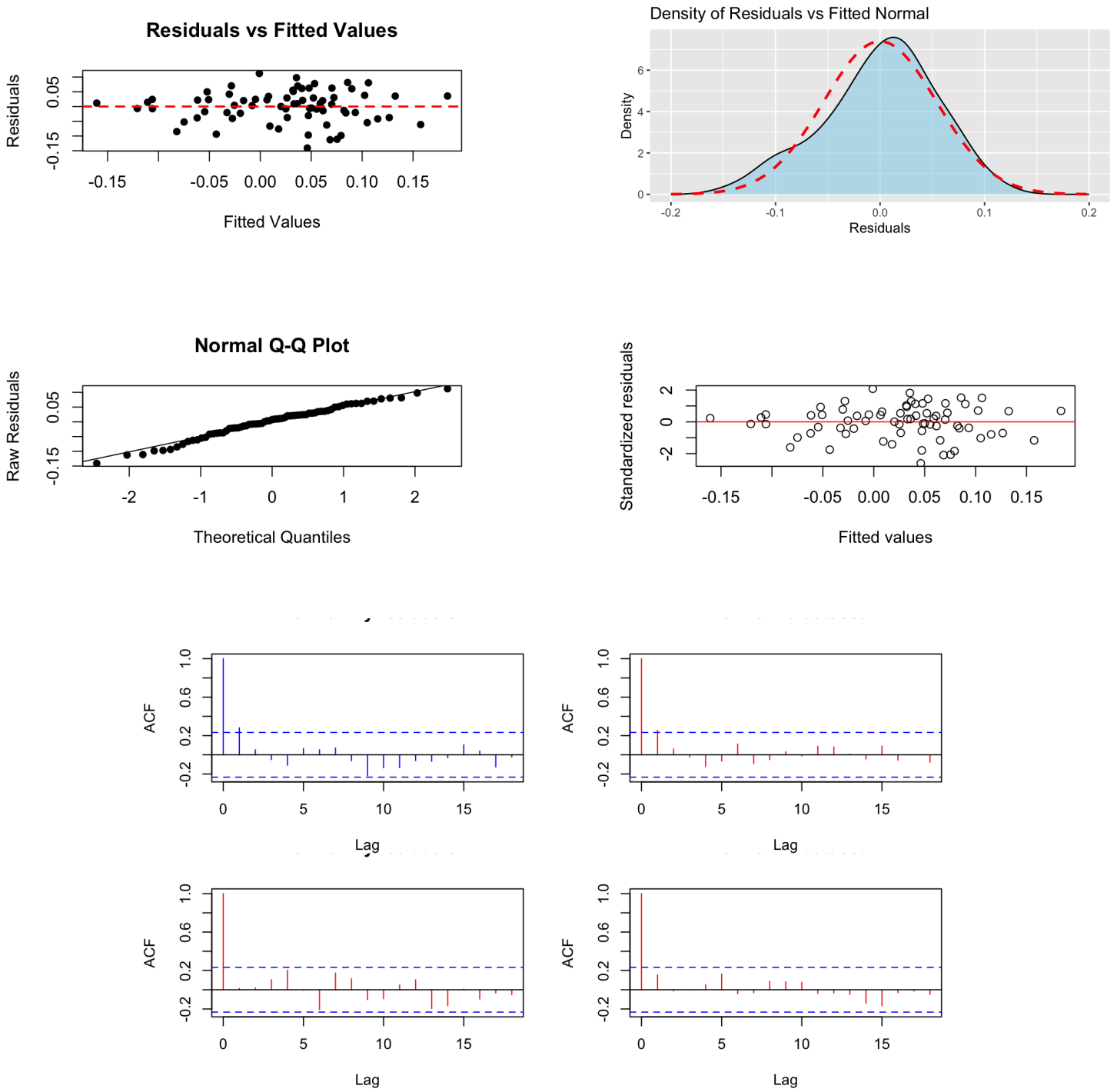


Figura 5: Residual Plots

From both a **visual** and **quantitative** analysis, we can conclude that:

- The data is **normally distributed**,
- The variance is **homogeneous** (homoscedasticity),
- However, there is a **serial correlation** in the data for a lag = 1, which is empirically observed even in the stock market, as precedent values influence successive ones and there is no full independence.

4.2 Data Points

Through both visual and quantitative analysis, I have investigated the potential presence of outliers and leverage points that could have an anomalous impact on the residuals and the overall dataset, potentially altering its distribution.

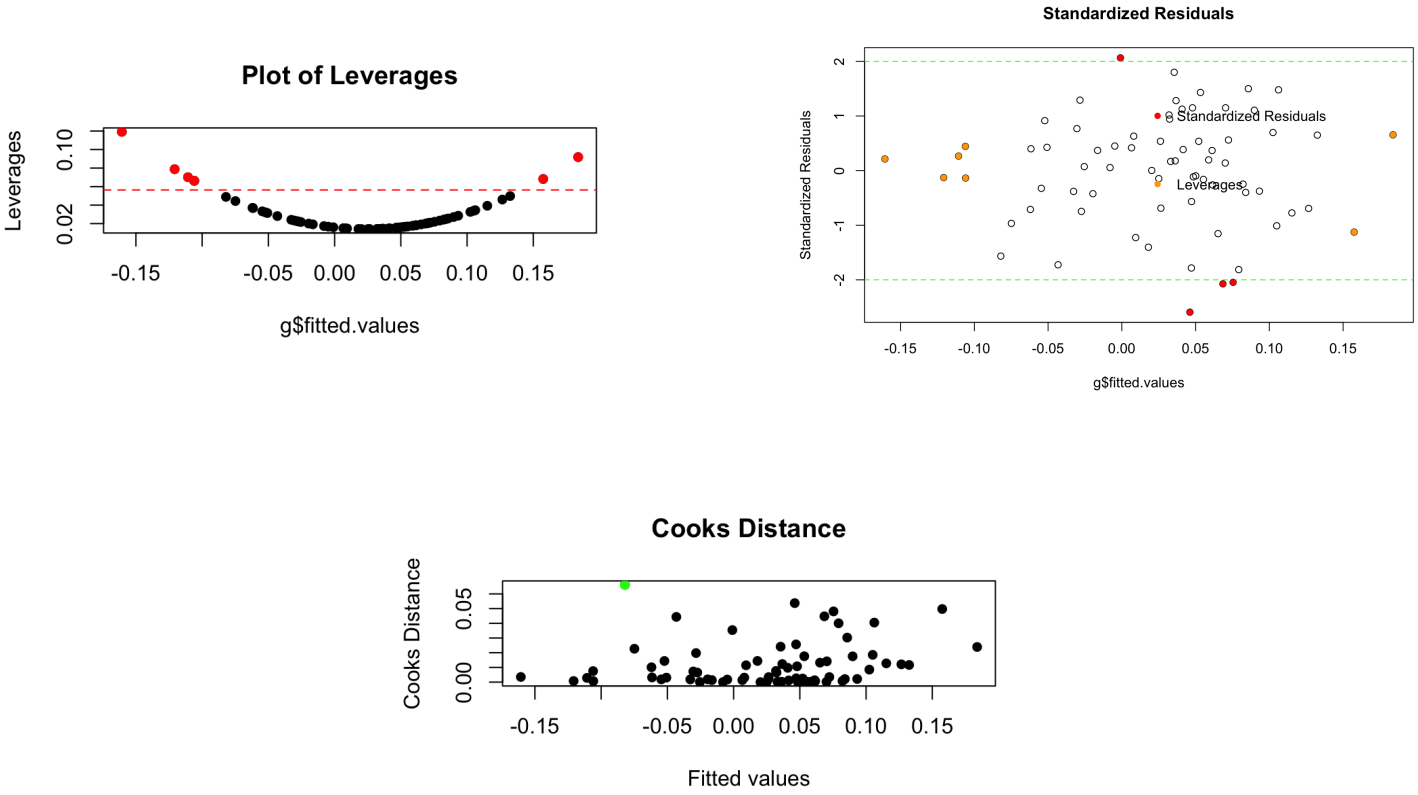


Figura 6: Leverage & Outlier Points

Let's test the model without **leverage points** :

Variable	Estimate	Standard Error	t value	p-value
Intercept	0.009398	0.007600	1.237	0.221
MarketRiskPremium	1.429308	0.185727	7.696	1.34e-10

Tabella 3: Regression Results

- Multiple R-squared: 0.4886
- Adjusted R-squared: 0.4803
- F statistic : 59.22 on 1 and 62 DF, p-value: 1.335e-10

We observe that the model has lost statistical significance, indicating that the omitted data points were essential for the model's construction. Specifically, the R-squared value has decreased by approximately 30%, suggesting that the excluded variables contributed substantially to the model's explanatory power.
Let's test the model taking off **outliers** :

Variable	Estimate	Standard Error	t-value	p-value
Intercept	0.013603	0.005856	2.323	0.0233 *
MarketRiskPremium	1.439310	0.113854	12.642	$< 2e-16$ ***

Tabella 4: Regression Results

- Multiple R-squared: 0.7109
- Adjusted R-squared: 0.7064
- F-statistic: 159.8 on 1 and 65 DF, p-value: $< 2.2e\text{-}16$

By removing the outlier points, we observe a significant increase in the R-squared value, and a greater confidence is provided in asserting that the coefficient of the intercept (alpha) is significantly different from zero. This suggests the presence of a surplus relative to the equilibrium condition. Nonetheless, the residuals still seem to show some autocorrelation for successive values.

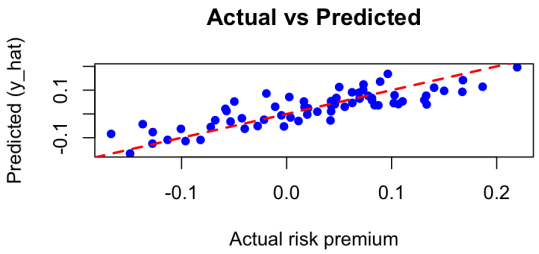


Figura 7: CAPM Regression Line without Outliers

Test	**p-value**
Shapiro-Wilk Test	0.2854
Breusch-Pagan Test	0.2198
Durbin-Watson Test (DW = 1.4866)	0.01705

Tabella 5: Results of Statistical Tests

5 Three Factor Analysis

The **Fama-French three-factor model** is an extension of the CAPM, which aims to explain the return on a stock by incorporating three key factors:

- **Market Risk Premium (MRP)**: This factor represents the excess return of the market over the risk-free rate. It is the same single factor seen in the CAPM.
- **Size (SMB)**: The "Small Minus Big" factor, which captures the difference in returns between small-cap stocks and large-cap stocks. Historically, small-cap stocks have been observed to have higher returns than large-cap stocks. The **SMB** factor is calculated as the average return on the three small portfolios minus the average return on the three big portfolios:

$$\text{SMB} = \frac{1}{3}(\text{Small Value} + \text{Small Neutral} + \text{Small Growth}) - \frac{1}{3}(\text{Big Value} + \text{Big Neutral} + \text{Big Growth})$$

- **Value (HML)**: The "High Minus Low" factor, which represents the difference in returns between high book-to-market ratio (value) stocks and low book-to-market ratio (growth) stocks. Value stocks tend to outperform growth stocks over the long term. The **HML** factor is calculated as the average return on the two value portfolios minus the average return on the two growth portfolios:

$$\text{HML} = \frac{1}{2}(\text{Small Value} + \text{Big Value}) - \frac{1}{2}(\text{Small Growth} + \text{Big Growth})$$

These three factors were introduced by **Eugene Fama** and **Kenneth French** to improve the explanatory power of asset pricing models by capturing returns based on **size** and **value** characteristics in addition to the **market risk**. Monthly and yearly data are publicly available on their website.

The **Fama-French three-factor model** can be expressed mathematically as:

$$E(R_i) - R_f = \beta_1 \cdot (E(R_m) - R_f) + \beta_2 \cdot \text{SMB} + \beta_3 \cdot \text{HML}$$

Where:

- $E(R_i)$ is the expected return of the asset i,
- R_f is the risk-free rate,
- $E(R_m) - R_f$ is the **Market Risk Premium**, i.e., the excess return of the market over the risk-free rate,
- **SMB** is the **Size factor** (Small Minus Big), which measures the return difference between small-cap and large-cap stocks,
- **HML** is the **Value factor** (High Minus Low), which measures the return difference between value and growth stocks,
- $\beta_1, \beta_2, \beta_3$ are the coefficients (loadings) that measure the sensitivity of the asset's return to each of the three factors.

The **SMB** and **HML** factors are constructed based on the characteristics of stocks. To calculate these factors, stocks are grouped into portfolios based on size and value, and the returns of these portfolios are used to calculate the respective factor values.

Applying such regression, we obtain the following results :

Variable	**Estimate**	**Standard Error**	**t-value**	**p-value**
Intercept	0.0097211	0.0061246	1.587	0.117172
MarketRiskPremium	1.3890995	0.1235387	11.244	$< 2e-16$ ***
SMB	0.0004637	0.0020893	0.222	0.825021
HML	-0.0050088	0.0013432	-3.729	0.000398 ***

Tabella 6: Regression Results

- Multiple R-squared: 0.6818
- Adjusted R-squared: 0.6675
- F-statistic: 47.85 on 3 and 67 DF, p-value: $<$ 2.2e-16

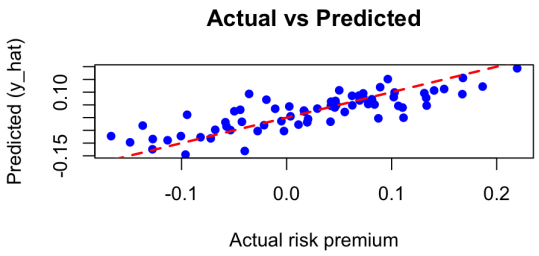


Figura 8: 3 Factor Regression Line

From the results obtained, we observe that the regression coefficient for the market risk premium remains at similar average values to those obtained in the CAPM regression, and the coefficients of the added factors do not significantly alter the model's response. However, we have statistical evidence that the HML factor, although very low, is significantly different from zero. This same result does not apply to the SMB variable.

5.1 Residuals

Let's now check if the model satisfies the assumptions on the residuals:

Test	**p-value**
Shapiro-Wilk Test	0.5956
Breusch-Pagan Test	0.4351
Durbin-Watson Test (DW = 1.582)	0.03679

Tabella 7: Results of Statistical Tests

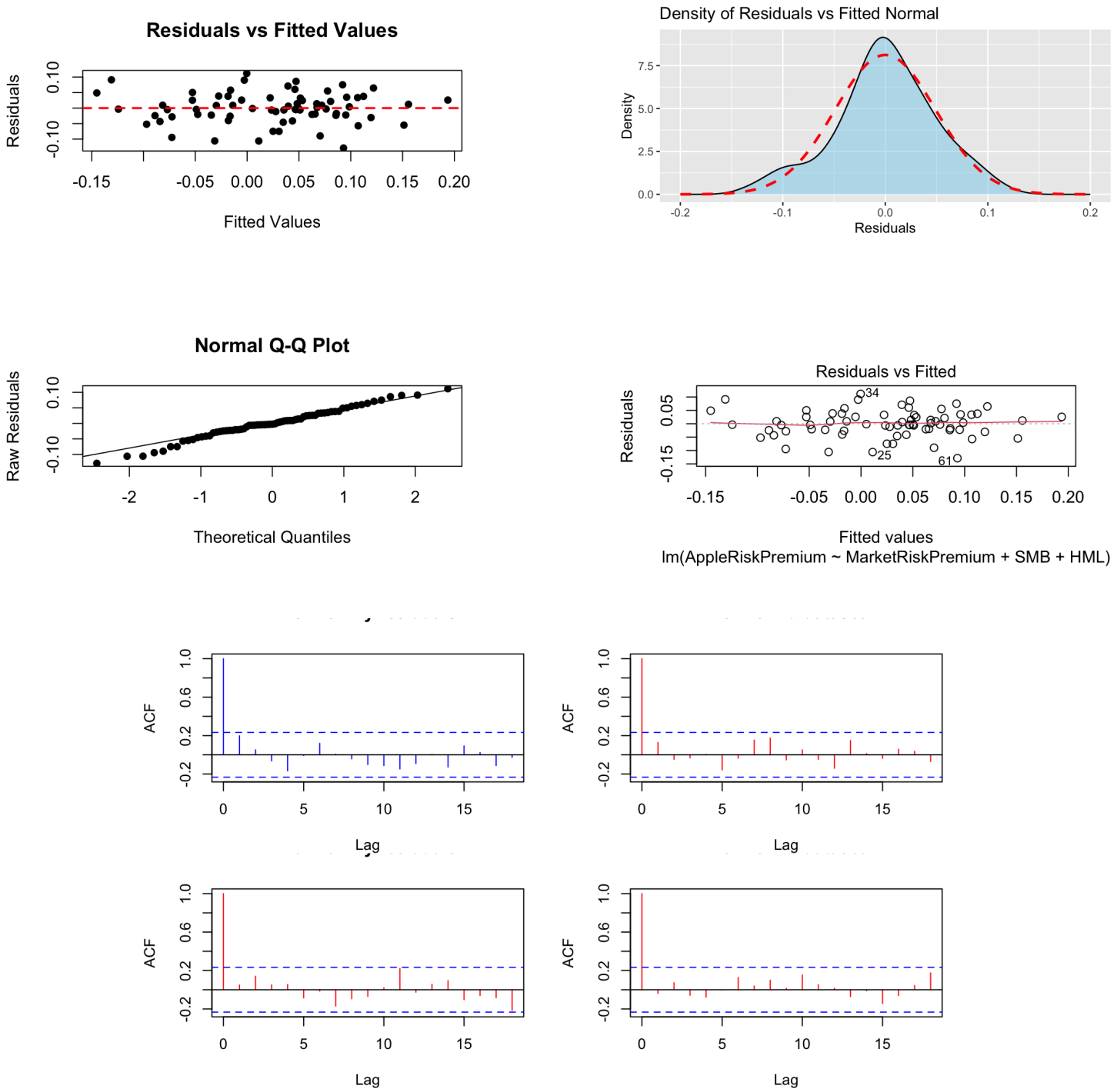


Figura 9: Residual Plots

From both a **visual** and **quantitative** analysis, we can conclude that:

- The data is **normally distributed**,
- The variance is **homogeneous** (homoscedasticity),
- However, there is a **serial correlation** in the data for a lag = 1, which is empirically observed even in the stock market, as precedent values influence successive ones and there is no full independence.

It is worth highlighting that the three-factor model resulted in higher confidence levels for the hypothesis concerning the residuals, suggesting that this model is more reliable.

5.2 Data Points

Through both visual and quantitative analysis, I have investigated the potential presence of outliers and leverage points that could have an anomalous impact on the residuals and the overall dataset, potentially altering its distribution.

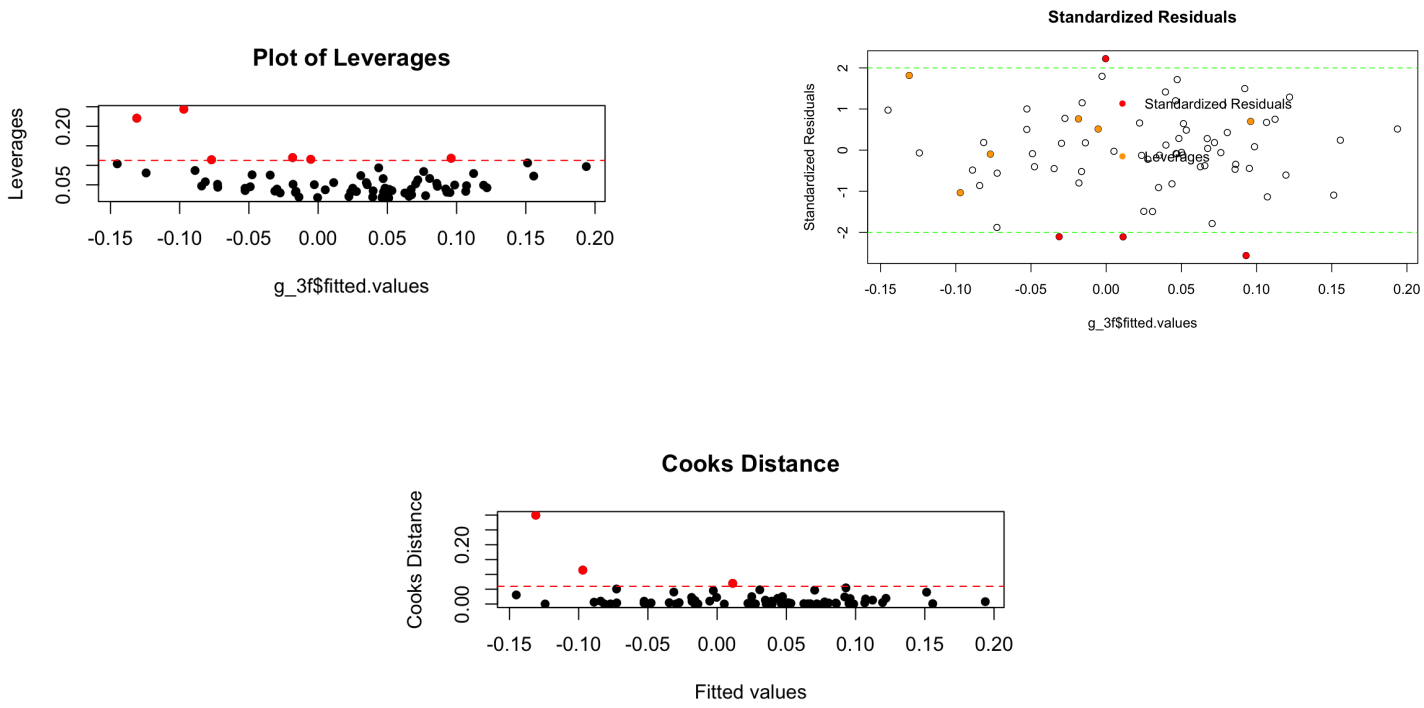


Figura 10: Leverage & Outlier Points

Let's test the model without leverage points :

Variable	Estimate	Standard Error	t-value	p-value
Intercept	0.0078896	0.0065404	1.206	0.232
MarketRiskPremium	1.3733969	0.1366260	10.052	1.46e-14 ***
SMB	0.0006728	0.0024826	0.271	0.787
HML	-0.0071158	0.0016589	-4.289	6.50e-05 ***

Tabella 8: Regression Results

- Multiple R-squared: 0.6822
- Adjusted R-squared: 0.6665
- F-statistic: 43.64 on 3 and 61 DF, p-value: 3.448e-15

Without the leverage points, our model's metrics didn't change much, the only observation that could be done is that the significance of 'HML' and 'MarketRiskPremium' increased.
Let's test the model taking off outliers :

Variable	Estimate	Standard Error	t value	p-value
Intercept	0.013155	0.005402	2.435	0.0177 *
MarketRiskPremium	1.399777	0.107099	13.070	$< 2e-16$ ***
SMB	0.001063	0.001795	0.592	0.5557
HML	-0.004980	0.001178	-4.229	7.74e-05 ***

Tabella 9: Regression Results

- Multiple R-squared: 0.7567
- Adjusted R-squared: 0.7451
- F-statistic: 65.32 on 3 and 63 DF, p-value: $< 2.2e\text{-}16$

In this case, we observe that the model's explanatory performance has improved significantly, as evidenced by the substantial increase in both the R-squared and adjusted R-squared values. However, what is particularly noteworthy is the increase in the significance of the intercept after removing the outliers. The intercept, often referred to as alpha, represents the excess return of the asset that cannot be explained by the market, size, or value factors included in the model.

Before removing the outliers, the intercept might have been distorted by extreme data points, making it less reliable and less significant. These outliers can "pull" the regression line toward them, thereby inflating or deflating the intercept, which could lead to misleading conclusions about the abnormal return (alpha). After eliminating the outliers, the model fits better to the central tendency of the data, leading to a more accurate and precise estimate of the intercept.

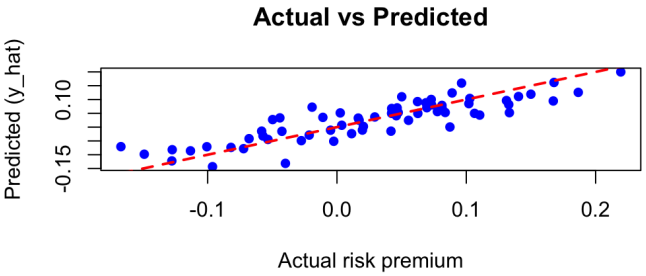


Figura 11: 3 Factors Outliers Regression Line

Test	p-value
Shapiro-Wilk Test	0.8239
Breusch-Pagan Test	0.04136
Durbin-Watson Test (DW = 1.6291)	0.06339

Tabella 10: Results of Statistical Tests

6 Five-Factor Fama-French Model

The **Fama-French five-factor model** extends the three-factor model by adding two additional factors:

- **RMW (Robust Minus Weak)**. This factor captures the difference in returns between firms with high operating profitability (Robust) and firms with low operating profitability (Weak). The formula for **RMW** is:

$$\text{RMW} = \frac{1}{2}(\text{Small Robust} + \text{Big Robust}) - \frac{1}{2}(\text{Small Weak} + \text{Big Weak})$$

  Where:

  - Small Robust and Big Robust represent small and large firms with high operating profitability,
  - Small Weak and Big Weak represent small and large firms with low operating profitability.

- **CMA (Conservative Minus Aggressive)**.This factor captures the difference in returns between firms that adopt conservative investment strategies and those that adopt aggressive investment strategies. The formula for **CMA** is:

$$\text{CMA} = \frac{1}{2}(\text{Small Conservative} + \text{Big Conservative}) - \frac{1}{2}(\text{Small Aggressive} + \text{Big Aggressive})$$

  Where:

  - Small Conservative and Big Conservative represent small and large firms with conservative investment strategies,
  - Small Aggressive and Big Aggressive represent small and large firms with aggressive investment strategies.

  Also for these factors, monthly and yearly data are publicly available on their website.

  The **Fama-French five-factor model** can be expressed mathematically as:

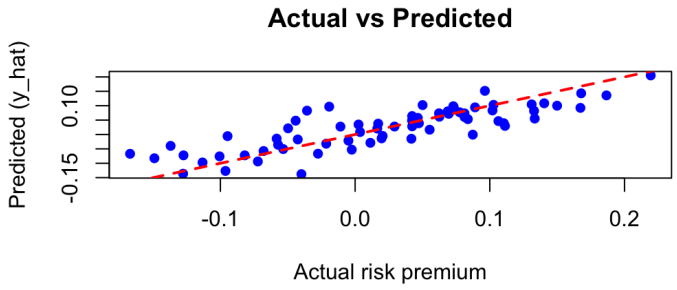
$$E(R_i) - R_f = \beta_1 \cdot (E(R_m) - R_f) + \beta_2 \cdot \text{SMB} + \beta_3 \cdot \text{HML} + \beta_4 \cdot \text{RMW} + \beta_5 \cdot \text{CMA}$$

Applying such regression, we obtain the following results :

Variable	Estimate	Standard Error	t-value	p-value
Intercept	0.007368	0.006217	1.185	0.240313
MarketRiskPremium	1.385979	0.132035	10.497	1.25e-15 ***
SMB	0.002053	0.002498	0.822	0.414311
HML	-0.007385	0.002132	-3.464	0.000947 ***
RMW	0.003897	0.002955	1.319	0.191880
CMA	0.003693	0.003083	1.198	0.235249

Tabella 11: Regression Results

- Multiple R-squared: 0.6958
- Adjusted R-squared: 0.6724
- F-statistic: 29.73 on 5 and 65 DF, p-value: 1.392e-15



In addition to the HML factor, which remains statistically significant, the other factors in the five-factor model do not appear to contribute significantly to the explainability of the model. As observed in the three-factor model analysis, the SMB, RMW, and CMA factors show relatively weak significance, with high p-values suggesting that their inclusion does not provide substantial explanatory power over and above the MarketRiskPremium and HML factors.

Furthermore, the R-squared values obtained for the five-factor model do not differ substantially from those calculated for the three-factor model, indicating that the additional factors do not significantly improve the model's fit to the data. This suggests that, in this particular case, the three-factor model might suffice in capturing the majority of the variability in the risk premium, and the added complexity of the five-factor model does not offer a significant improvement in predictive power or explanatory capacity.

6.1 Residuals

Let's now test the hypothesis on the residuals :

Test	p-value
Shapiro-Wilk Test	0.1326
Breusch-Pagan Test	0.3299
Durbin-Watson Test (DW = 1.6)	0.0442

Tabella 12: Results of Statistical Tests

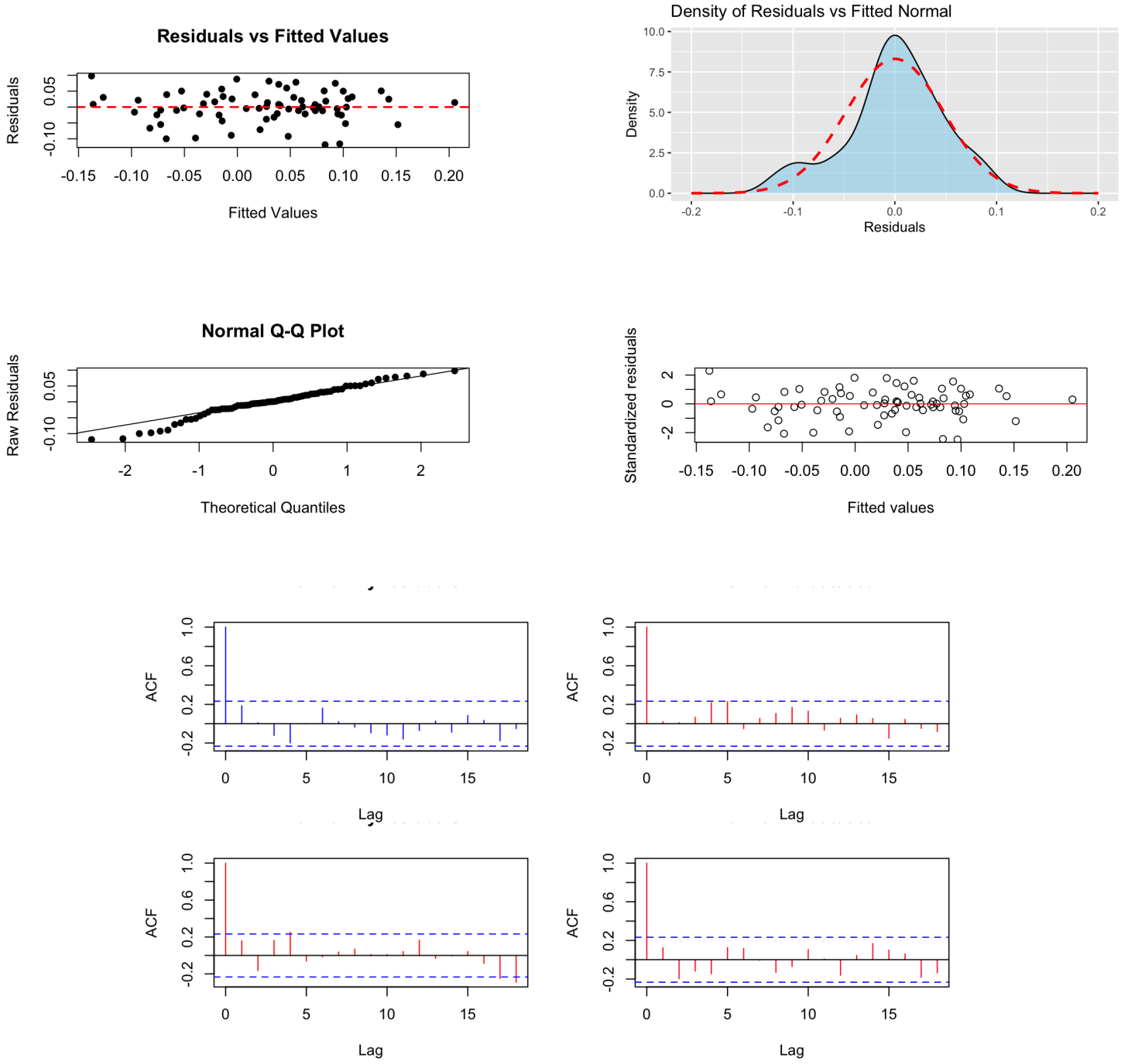


Figura 12: Multiple Plots in One Frame

The hypotheses appear to be satisfied, except for the one concerning the independence of successive data points. However, regarding the Shapiro-Wilk test, we observe a significantly smaller p-value, indicating that the residuals might deviate from normality.

6.2 Data Points

Let's perform the usual data points inspection.

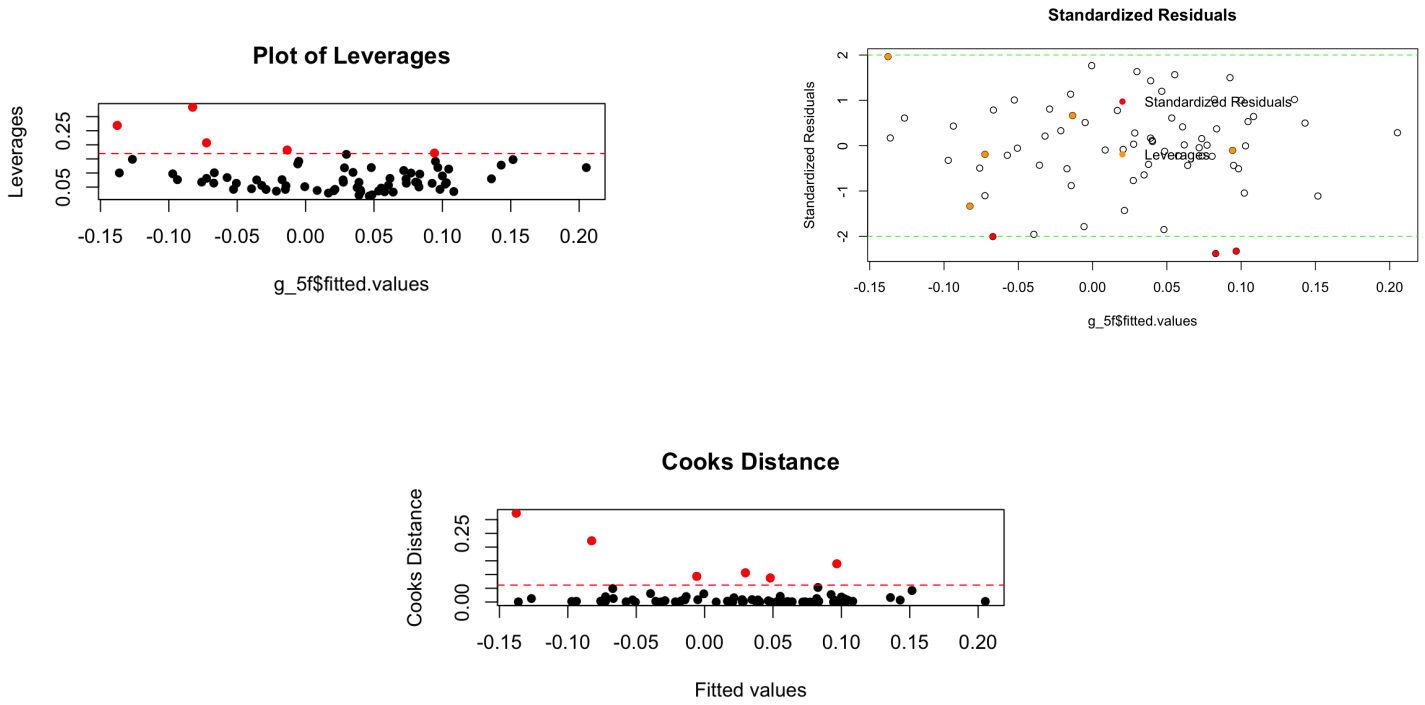


Figura 13: Leverage & Outlier Points

Let's test the model without leverage points :

Variable	Estimate	Standard Error	t value	Pr(>|t|)
Intercept	0.006023	0.006406	0.940	0.3508
MarketRiskPremium	1.301578	0.147782	8.807	2.08e-12 ***
SMB	0.004513	0.002905	1.554	0.1255
HML	-0.011308	0.002596	-4.357	5.25e-05 ***
RMW	0.006411	0.003324	1.929	0.0585 .
CMA	0.004702	0.003344	1.406	0.1649

Tabella 13: Regression Results

- Multiple R-squared: 0.7109
- Adjusted R-squared: 0.6868
- F-statistic: 29.5 on 5 and 60 DF, p-value: 5.475e-15

Let's test the model taking off outliers :

Variable	Estimate	Standard Error	t value	Pr(>|t|)
Intercept	0.011518	0.005614	2.052	0.0444 *
MarketRiskPremium	1.413838	0.119675	11.814	$< 2e-16$ ***
SMB	0.001988	0.002226	0.893	0.3753
HML	-0.008467	0.001914	-4.424	3.98e-05 ***
RMW	0.004638	0.002725	1.702	0.0937 .
CMA	0.005425	0.002788	1.946	0.0562 .

Tabella 14: Regression Results

- Multiple R-squared: 0.7516
- Adjusted R-squared: 0.7316

- F-statistic: 37.52 on 5 and 62 DF, p-value: $< 2.2e\text{-}16$

By removing the outliers, we observe a more significant improvement in the model's goodness-of-fit, as reflected by a notable increase in the R-squared value. Additionally, the standard errors of the variables decrease, ensuring greater precision in the model and providing a clearer isolation of causal relationships.

Regarding the factors, in both models, we observe an increase in the relevance of the HML factor, which, alongside the MarketRiskPremium, seems to be the only significant factor. In the model without outliers, however, we notice that the intercept (alpha) takes on a value significantly different from zero, with a much higher level of statistical significance. This highlights the impact of extreme points in skewing the surplus of the asset, underlining the importance of outlier removal in obtaining more accurate and meaningful results.

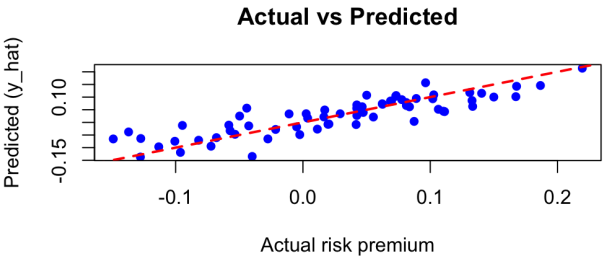


Figura 14: 5 Factors Outliers Plot

Test	**p-value**
Shapiro-Wilk Test	0.6117
Breusch-Pagan Test	0.1441
Durbin-Watson Test (DW = 1.6291)	0.1823

Tabella 15: Results of Statistical Tests

7 Multifactorial Analysis

I have decided to extend our model by adding five additional factors, each representing one of the five most important **GICS sectors**. The chosen factors are: **Industrials**, **Information Technology**, **Real Estate**, **Health Care**, and **Financial Services**.

The formula used to calculate the monthly returns for each sector is as follows:

$$r_i = \sum (r_s \cdot \frac{w_s}{w_N})$$

Where:

- r_s is the monthly return of the stock belonging to sector i,
- w_s is the market capitalization of stock s at the current date,
- w_N is the sum of the market capitalizations of all stocks in the sector.

It is reasonable to assume that this model will exhibit high collinearity, as the market return itself is partially constituted by the returns of the aforementioned sectors.

The formula for this extended model can be expressed as:

$$E(R_i) - R_f = \beta_1 \cdot (E(R_m) - R_f) + \beta_2 \cdot \text{SMB} + \beta_3 \cdot \text{HML} + \beta_4 \cdot \text{RMW} + \beta_5 \cdot \text{CMA} + \sum_{i=1}^{5} \beta_i \cdot R_i$$

Applying such regression, we obtain the following results :

Variable	Estimate	Standard Error	t-value	p-value
Intercept	0.028014	0.007209	3.886	0.000257 ***
MarketRiskPremium	3.912009	0.519729	7.527	3.15e-10 ***
Industrials	-0.749531	0.250156	-2.996	0.003970 **
'Financial Services'	-0.563266	0.271009	-2.078	0.041957 *
'Health Care'	-0.800028	0.216648	-3.693	0.000481 ***
'Real Estate'	-0.213836	0.167105	-1.280	0.205593
'Information Technology'	-0.327864	0.175972	-1.863	0.067337 .
SMB	0.005291	0.002257	2.344	0.022406 *
HML	-0.006569	0.002545	-2.581	0.012307 *
RMW	0.003605	0.002772	1.300	0.198406
CMA	0.006264	0.003243	1.932	0.058116 .

Tabella 16: Regression Results

- R-squared : 0.8066
- Adjusted R-squared: 0.7743
- F-statistic: 25.02 on 10 and 60 DF, p-value: $<$ 2.2e-16

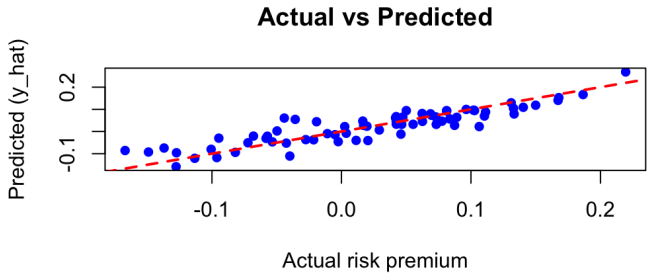


Figura 15: MultiFactorial Regression Line

We observe a negative impact on the coefficients for the returns of all stock clusters, including the Information Technology (IT) sector to which Apple belongs. However, the standard errors for these clusters are relatively high, suggesting that the results should be interpreted with caution. The test indicates a decreased significance of the HML factor, while the Health Care factor shows a very high coefficient with extremely high statistical significance.

Furthermore, the intercept is significantly different from zero and remains positive, reinforcing the apparent surplus return recorded by Apple's stock in recent years. This suggests that, despite some variation in the factors, Apple continues to exhibit excess returns beyond what is explained by the sector-based factors.

7.1 Residuals

Vediamo adesso se il modello verifica le ipotesi sui residui :

Test	p-value
Shapiro-Wilk Test	0.4525
Breusch-Pagan Test	0.5269
Durbin-Watson Test (DW = 1.42)	0.0034

Tabella 17: Results of Statistical Tests

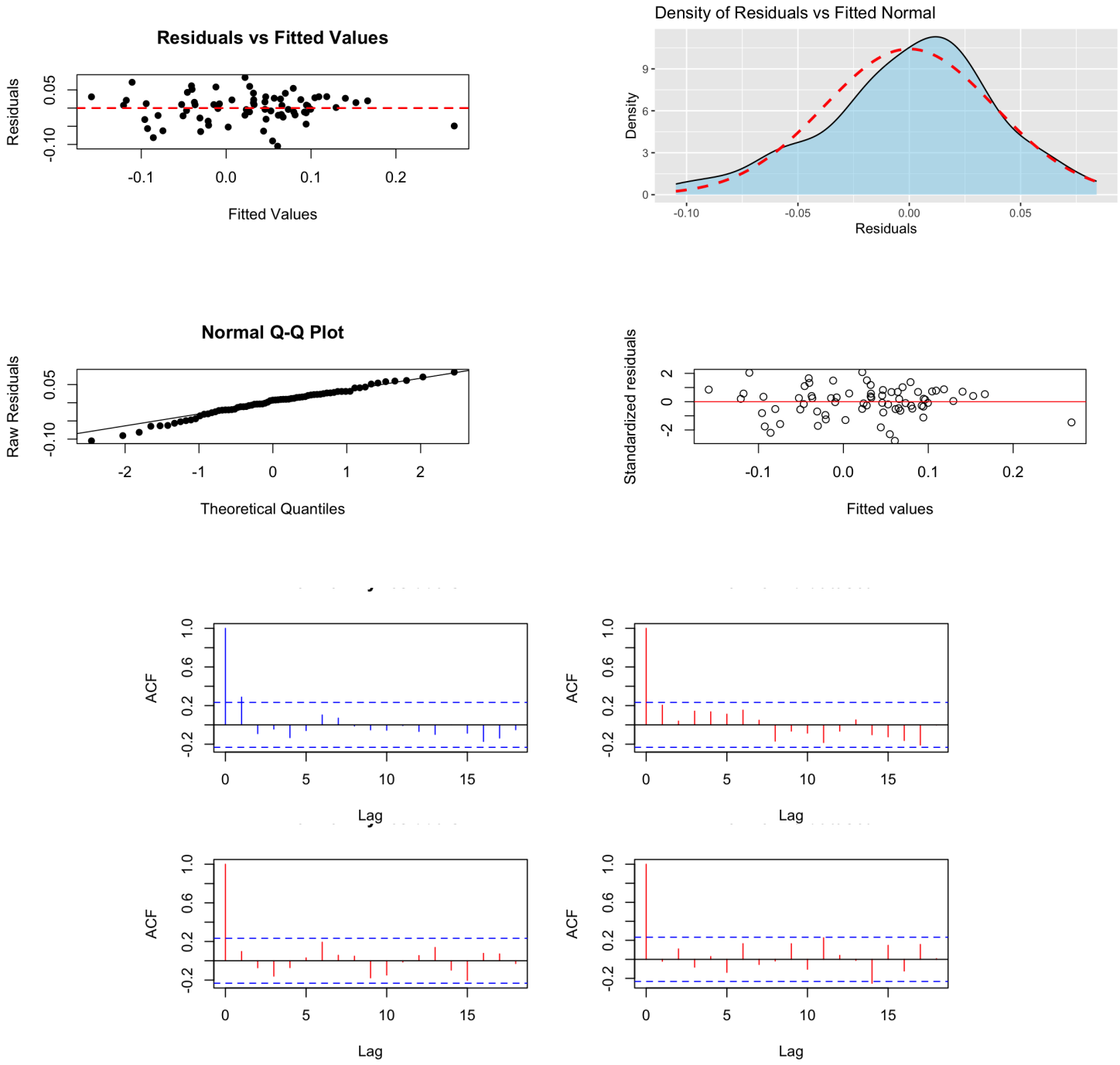


Figura 16: Residual Plots

The model appears to satisfy the assumptions regarding the residuals, with the exception of the assumption concerning the independence of successive values.

7.2 Data Points

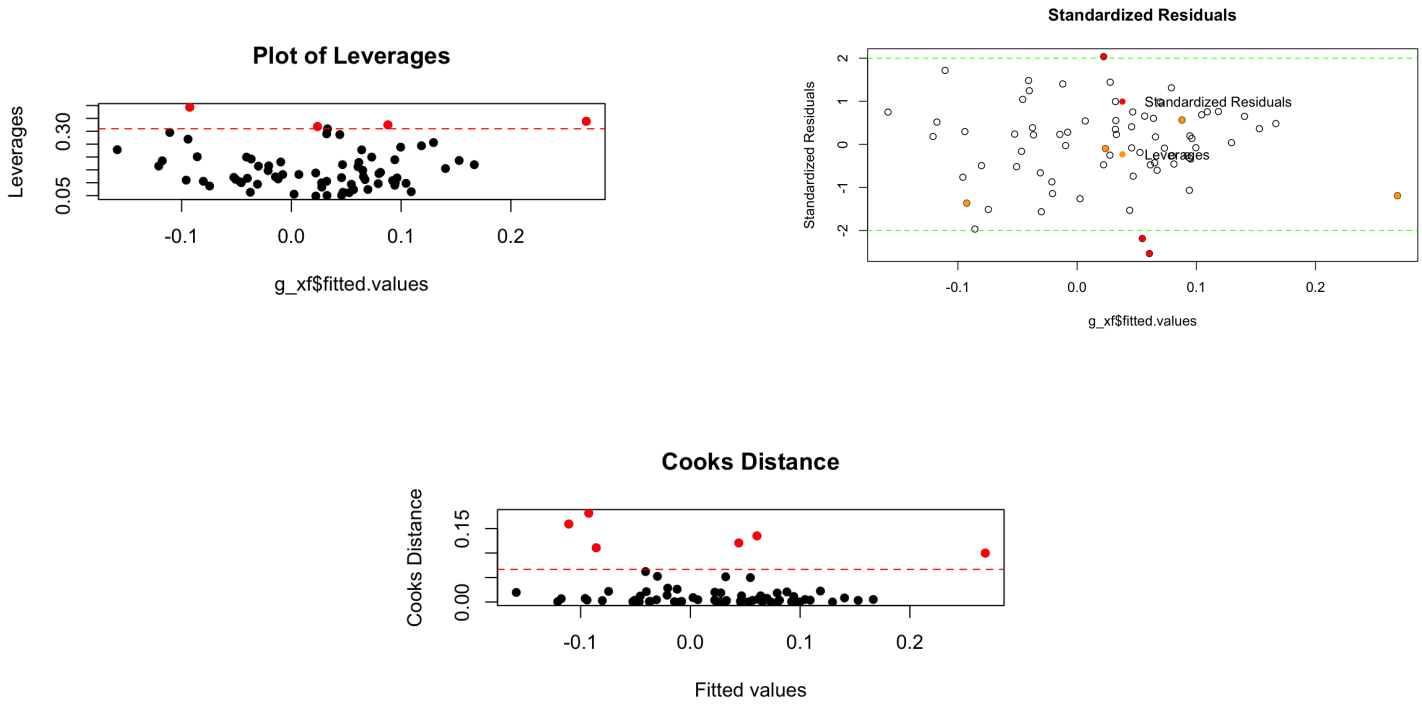


Figura 17: Leverage & Outlier Points

Let's test the model without leverage points :

Variable	Estimate	Standard Error	t-value	p-value
Intercept	0.031696	0.007516	4.217	9.12e-05 ***
MarketRiskPremium	4.152878	0.570493	7.279	1.19e-09 ***
Industrials	-0.930799	0.283632	-3.282	0.00178 **
Financial Services	-0.510534	0.278415	-1.834	0.07201 .
Health Care	-0.703240	0.223434	-3.147	0.00264 **
Real Estate	-0.321371	0.172203	-1.866	0.06725 .
Information Technology	-0.393620	0.189253	-2.080	0.04213 *
SMB	0.006166	0.002592	2.379	0.02080 *
HML	-0.008184	0.002742	-2.985	0.00420 **
RMW	0.003585	0.003061	1.171	0.24641
CMA	0.007580	0.003564	2.127	0.03785 *

Tabella 18: Regression Results

- Multiple R-squared: 0.7946
- Adjusted R-squared: 0.7579
- F-statistic: 21.66 on 10 and 56 DF, p-value: 8.412e-16

Let's test the model taking off outliers :

Variable	Estimate	Standard Error	t-value	p-value
Intercept	0.030910	0.006314	4.896	8.43e-06 ***
MarketRiskPremium	4.082674	0.455910	8.955	1.83e-12 ***
Industrials	-0.544279	0.222831	-2.443	0.01771 *
Financial Services	-0.779696	0.239895	-3.250	0.00194 **
Health Care	-0.895625	0.191682	-4.672	1.86e-05 ***
Real Estate	-0.186987	0.145550	-1.285	0.20410
Information Technology	-0.371767	0.154157	-2.412	0.01913 *
SMB	0.004179	0.001979	2.112	0.03911 *
HML	-0.005623	0.002218	-2.535	0.01402 *
RMW	0.003284	0.002462	1.334	0.18757
CMA	0.006206	0.002821	2.200	0.03187 *

Tabella 19: Regression Results

- Multiple R-squared: 0.8576
- Adjusted R-squared: 0.8326
- F-statistic: 34.34 on 10 and 57 DF, p-value: $<$ 2.2e-16

We observe that the Adjusted R-squared of the model improved significantly, reaching a peak of 0.83, the highest value achieved so far in our analysis. Testing the hypothesis on the residuals, we find that the model is both robust and statistically relevant.

The Health Care factor and the intercept remain highly significant in both models, both with and without outliers and leverage points. This further reinforces their relevance in explaining the risk premium of the asset.

Regarding the Fama-French factors, they do not appear to be very significant in this context. While the size and value factors still provide some explanatory power, their contribution is limited. Additionally, the standard errors for the factors related to the stock clusters remain relatively high, suggesting that these factors are still influenced by variability or noise in the data.

Test	p-value
Shapiro-Wilk Test	0.2203
Breusch-Pagan Test	0.1949
Durbin-Watson Test (DW = 1.7443)	p-value = 0.1384

Tabella 20: Results of Statistical Tests

This time, even the independence of successive residuals hypothesis seems to be confirmed statistically. Nonetheless, we observe that the standard error for the MarketRiskPremium is relatively high, as is the standard error for other asset clusters. This suggests the potential presence of multicollinearity, where some variables may not be completely independent of each other, making it difficult to isolate the individual contributions and effects of each variable in the model.

To further investigate this, we will calculate the Variance Inflation Factor (VIF) values to check whether this hypothesis of multicollinearity is valid.

7.3 VIF

Variable	VIF
MarketRiskPremium	27.780834
Industrials	9.036870
Financial Services	12.648789
Health Care	3.648747
Real Estate	3.781909
Information Technology	6.957951
SMB	2.267613
HML	5.142590
RMW	1.771597
CMA	3.229065

Tabella 21: Variance Inflation Factors (VIF) for the Model Variables

As we can see from the VIF values, we have high collinearity among the stock classes and the Market Premium, especially with Financial Services and Industrials. This makes sense as the whole market movement are the combined results of all the other stock groups behaviours. Let's try take off 'Industrials' and 'Financial Services' :

Variable	VIF
MarketRiskPremium	14.204612
Health Care	3.628116
Real Estate	3.766993
Information Technology	6.450859
SMB	1.994594
HML	3.541704
RMW	1.454162
CMA	2.884595

Tabella 22: Variance Inflation Factors (VIF) for the Model Variables

We observe lowered values for the Market Premium, but we can still try to improve it by taking off 'Information Technology' variable:

Variable	VIF
MarketRiskPremium	5.400732
Health Care	3.430490
Real Estate	3.480160
SMB	1.990397
HML	3.030704
RMW	1.454160
CMA	2.879182

Tabella 23: Variance Inflation Factors (VIF) for Final Model Variables

This model appears to be generally robust; however, upon testing the residuals, we obtain a Shapiro-Wilk test p-value of **0.02841**, which leads us to reject the normality assumption. This result undermines the statistical relevance of the model, as the violation of normality could affect the validity of inference and hypothesis testing within the model.

7.4 AIC

Let me attempt an alternative approach by applying AIC model optimization methods, followed by efforts to reduce the VIF of the resulting model.

By applying backward model selection to the original set of variables, I obtain the following results:

Variable	Estimate	Standard Error	t-value	p-value
Intercept	0.029389	0.006955	4.226	7.96e-05 ***
MarketRiskPremium	3.803101	0.472233	8.053	3.20e-11 ***
Industrials	-0.683608	0.245302	-2.787	0.007055 **
Financial Services	-0.720561	0.245918	-2.930	0.004738 **
Health Care	-0.785529	0.217775	-3.607	0.000619 ***
Information Technology	-0.299815	0.169380	-1.770	0.081632 .
SMB	0.003404	0.001891	1.800	0.076713 .
HML	-0.004806	0.002297	-2.093	0.040482 *
CMA	0.005165	0.003180	1.624	0.109364

Tabella 24: Regression Results

- Multiple R-squared: 0.7966
- Adjusted R-squared: 0.7704

We observe that the model is likely to suffer from some collinearity. However, this time, the Shapiro-Wilk test yields a p-value of 0.3934, which provides strong evidence in favor of normality of the residuals, ensuring statistical confidence in the assumption of normality.

The VIF values for this model are as follows:

Variable	VIF
MarketRiskPremium	23.261721
Industrials	8.517004
Financial Services	10.495425
Health Care	3.607492
Information Technology	6.437174
SMB	1.561488
HML	4.170887
CMA	3.047507

Tabella 25: Variance Inflation Factors (VIF) for the Model Variables

Let's now try to lower the VIF of the model, by taking off Financial Services :

Variable	VIF
MarketRiskPremium	16.400001
Industrials	7.786946
Health Care	3.559250
Information Technology	6.021136
SMB	1.550644
HML	3.412862
CMA	2.858244

Tabella 26: Variance Inflation Factors (VIF) for the Final Model Variables

We can still do better, let's try taking off 'Information Technology' too :

Variable	VIF
MarketRiskPremium	8.505591
Industrials	6.928617
Health Care	3.305518
HML	2.724826
CMA	2.699688

Tabella 27: Variance Inflation Factors (VIF) for the Selected Model Variables

We finally obtained accettable values for the VIF of our variable, although the model might still present some collinearity among the first two variables. Let's test the hypothesis of such this model and see how it performs :

Variable	Estimate	Standard Error	t-value	p-value
Intercept	0.018266	0.005366	3.404	0.00117 **
MarketRiskPremium	2.651206	0.288541	9.188	3.54e-13 ***
Industrials	-0.639311	0.223172	-2.865	0.00569 **
Health Care	-0.709675	0.208679	-3.401	0.00118 **
HML	-0.006196	0.001847	-3.355	0.00136 **
CMA	0.007548	0.002950	2.559	0.01297 *

Tabella 28: Regression Results

- Multiple R-squared: 0.7974
- Adjusted R-squared: 0.7811
- F-statistic: 48.8 on 5 and 62 DF, p-value: $<$ 2.2e-16

The Shapiro-Wilk test and the Breusch-Pagan test yield p-values of 0.225 and 0.4047, respectively, which provide higher confidence in the validity of the results. These values suggest that the assumptions of normality and homoscedasticity are reasonably satisfied, reinforcing the robustness of the model.
This one seems to be the **best model** obtained so far throughout our analysis.

7.5 Alternative Studies

I also explored different approaches for variable selection. The first method involved removing variables with higher **p-values**, based on the assumption that these variables are likely to be less relevant. The second approach involved applying the **Box-Cox** transformation to improve model fit.

The p-value transformation resulted in the following model:

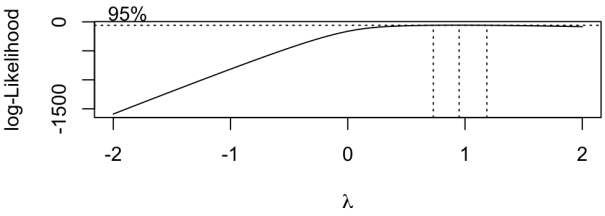
Variable	Estimate	Standard Error	t-value	p-value
Intercept	0.012798	0.005505	2.325	0.023363 *
MarketRiskPremium	2.017088	0.194660	10.362	3.71e-15 ***
Health Care	-0.758630	0.217712	-3.485	0.000911 ***
HML	-0.008745	0.001712	-5.109	3.33e-06 ***
RMW	0.003809	0.002249	1.694	0.095295 .
CMA	0.008819	0.003026	2.915	0.004948 **

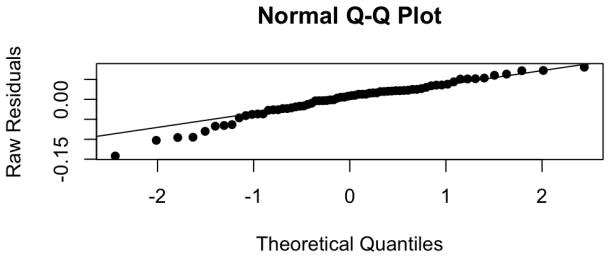
Tabella 29: Regression Results

- Multiple R-squared: 0.7807
- Adjusted R-squared: 0.763

We see that almost all the variables are significantly different from zero, but the shapiro test resulted in a p-value of 0.02412 , rejecting the hypothesis of normality of residuals.
The **Box-Cox** method was applied to the target variable, shifted upwards by the minimum value of the dataset to ensure positive values, which is a fundamental assumption for the applicability of the model. The method resulted in a lambda value very close to one, indicating that no transformation improves the normality of the residuals.

The results obtained using the lambda value of 0.95 yield a p-value of 0.0148 for the Shapiro-Wilk test, leading to the rejection of the normality assumption.





8 Results

- The value of the **beta** for the risk premium increased as more factors were added, in parallel with the rise in **standard error** and **R-squared** values. This suggests that as additional factors are included, the model's ability to minimize the squared distance between predicted and actual values improves. However, it is important to note that the relatively small size of the dataset may lead to potential **overfitting**, meaning that while the model performs well on the given data, its ability to generalize to new data could be limited.
- Upon adding the **stock cluster factors**, the **market beta** decreased, but when combined with the beta values for the individual stock groups, the result was approximately the same as in the previous regressions. This leads us to infer that the **correlation**, in the context of the model, is largely absorbed by the **market risk premium** and mitigated proportionally by the various sectors. However, the high **standard errors** make precise analysis challenging, likely due to the high **multicollinearity**.
- The **Fama-French factors** have little impact on the model, and their effect is definitely weaker than that of the **clustered stock returns**, as evidenced by the very low values and confidence intervals that include zero. This suggests that, in the context of this analysis, the **Fama-French model** may not be well-suited for explaining **linear correlations**.

- When outliers were removed, all models improved significantly, as did the assumptions regarding the residuals. This is consistent with the fact that outliers tend to **inflate the tails** of the residual distribution, altering its normality. An interesting next step would be to implement a method for **preemptively identifying outliers**, in order to construct a model that is as **statistically relevant** as possible, and then analyze the **alpha**, or the **surplus return** of the asset.
- The best-performing **multifactor model** shows a **non-zero intercept**, confirming that, relative to the factors, Apple's return has generated a **surplus**, consistent with **financial valuations**.

Model Name	**F-Test**	**Adj. R-Squared**	**Intercept Coefficient (p-value)**
CAPM	110.5	0.6101	0.01 (0.115)
3-Factors	47.85	0.6675	0.0097 (0.117)
5-Factors	29.73	0.6724	0.0074 (0.24)
Multifactor	25.02	0.774	0.03 (2.5e-4)

Tabella 30: Regression Model Results

Model Name	**F-Test**	**Adj. R-Squared**	**Intercept Coefficient (p-value)**
CAPM	159.8	0.7064	0.014 (0.023)
3-Factors	65.32	0.7451	0.013 (0.117)
5-Factors	37.52	0.7316	0.012 (0.04)
Multifactor	48.8	0.7811	0.018 (1.2e-3)

Tabella 31: Best Models Results