Checking Assumptions of the CLRM

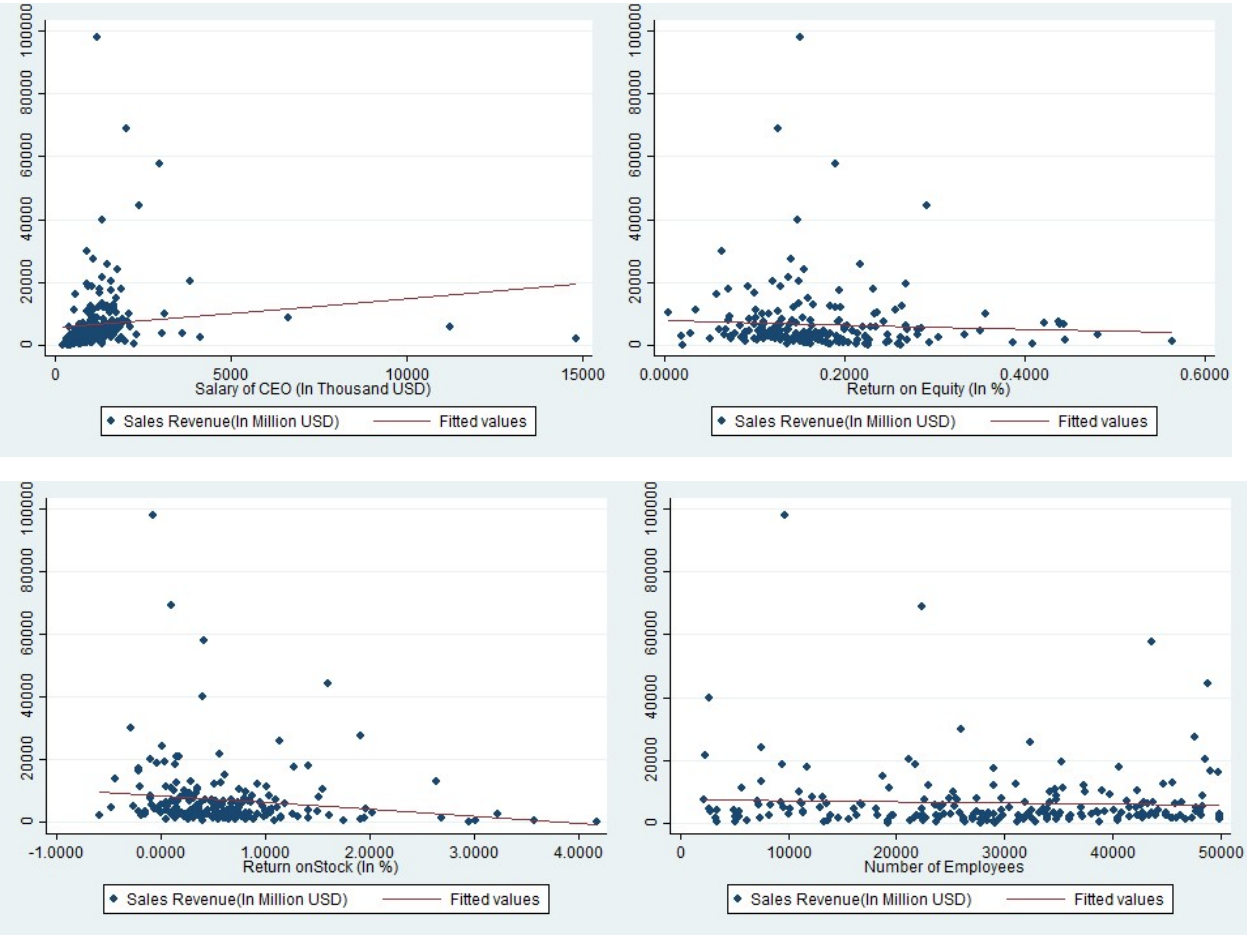
Before conducting regression analysis, it's crucial to check the assumptions underlying the Ordinary Least Squares (OLS) regression model to ensure that the estimated coefficients are unbiased, efficient, and have the smallest variance among all unbiased estimators, known as the Best Linear Unbiased Estimator (BLUE) property. Violations of these assumptions can lead to biased estimates and inefficient estimators, destroying the BLUE property of the OLS estimators.

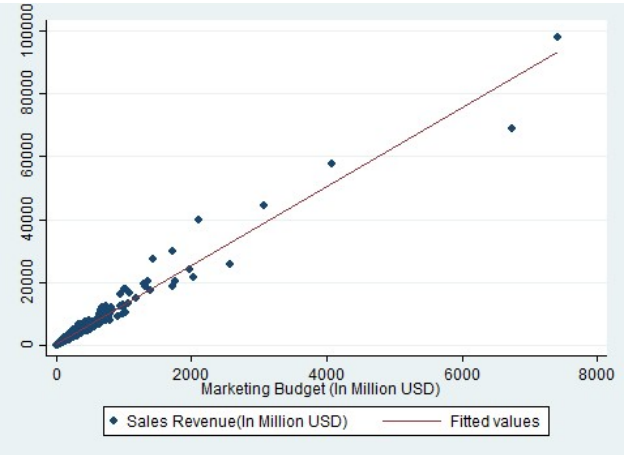
We will now proceed to validate these assumptions.

Assumptions of CLRM

Assumption 1: The regression model is linear in the parameters.

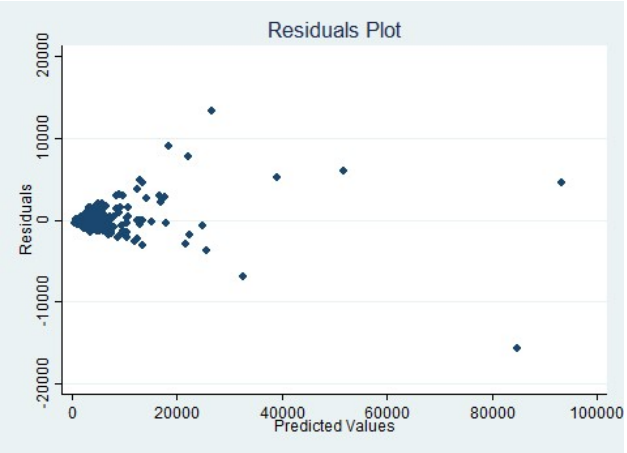
By creating scatterplots between the dependent variable and each independent variable, we can determine whether the model is linear or not.





From the above graphs, it is clear that the assumption of linearity holds true for our model.

Examine the residuals plot(s) to assess whether the relationship between the residuals and the predicted values or independent variables appears to be random and evenly spread out around zero.



So, a random scatter of points around zero suggests that the assumptions of linearity and constant variance are met.

Assumption 2: There should be enough variation in Xi to be qualified as explanatory variable.

By calculating summary statistics for each independent variable, we can determine whether the independent variables vary or not.

Salary of CEO (In Thousand USD)	
Mean	1281.119617
Standard Error	94.92710989
Median	1039
Mode	1368
Standard Deviation	1372.345308
Sample Variance	1883331.644
Kurtosis	58.97122942
Skewness	6.904576803
Range	14599
Minimum	223
Maximum	14822
Sum	267754
Count	209

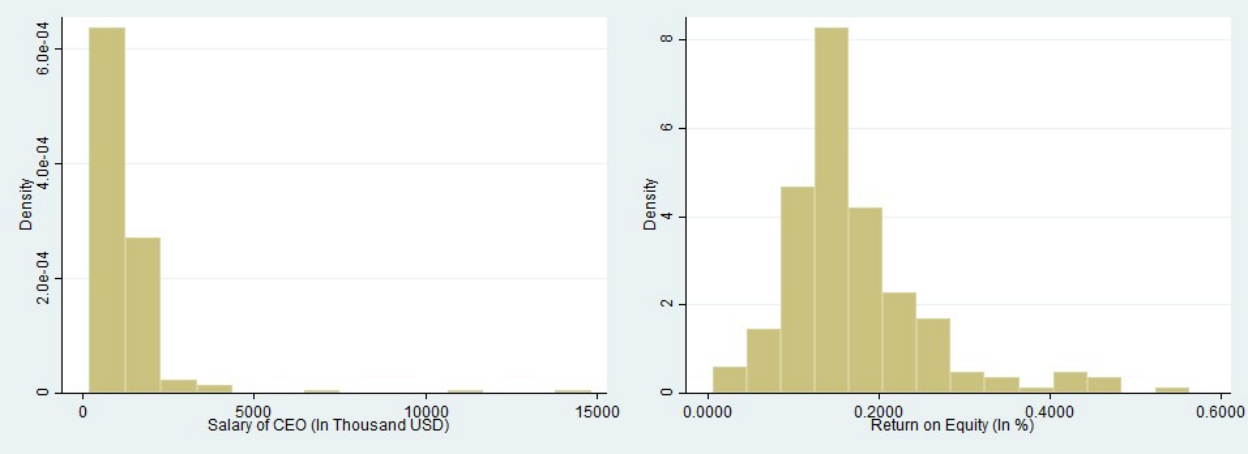
Return on Equity (In %)	
Mean	0.171842105
Standard Error	0.005892376
Median	0.155
Mode	0.151
Standard Deviation	0.085185087
Sample Variance	0.007256499
Kurtosis	3.797366054
Skewness	1.572125873
Range	0.558
Minimum	0.005
Maximum	0.563
Sum	35.915
Count	209

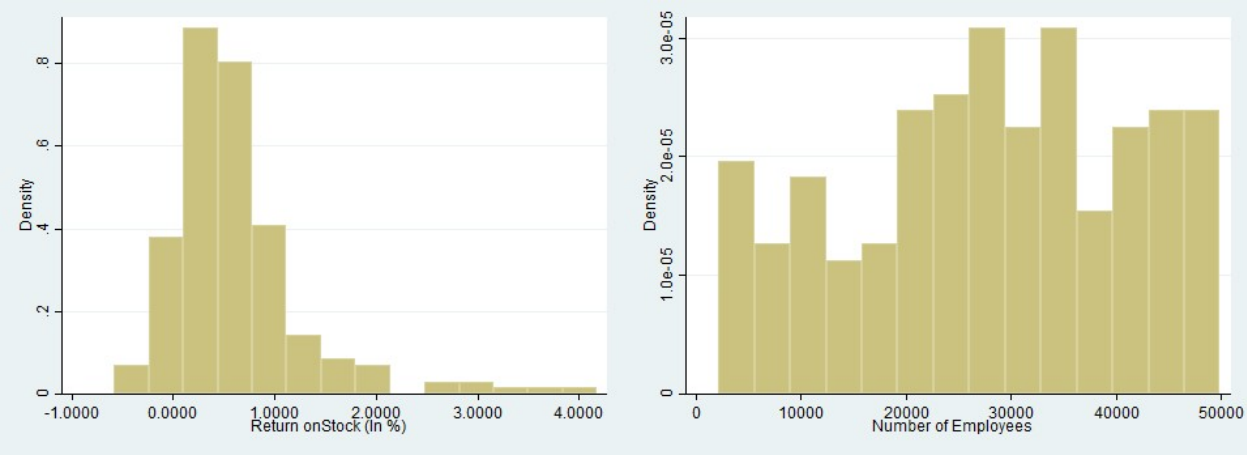
Return on Stock (In %)	
Mean	0.618038278
Standard Error	0.047159053
Median	0.52
Mode	0.37
Standard Deviation	0.681770516
Sample Variance	0.464811037
Kurtosis	6.585102165
Skewness	2.094631134
Range	4.76
Minimum	-0.58
Maximum	4.18
Sum	129.17
Count	209

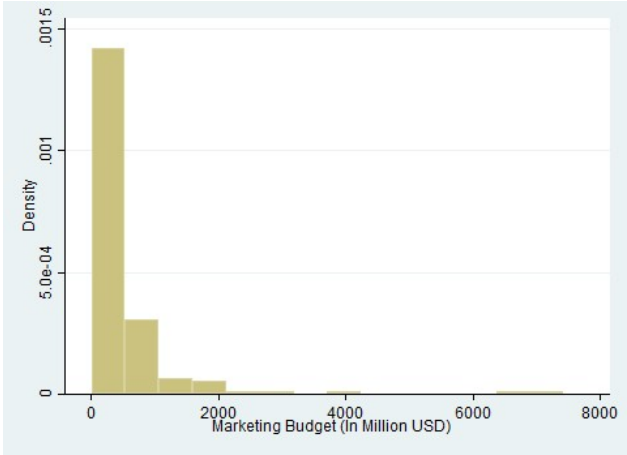
Number of Employees	
Mean	28056.60287
Standard Error	918.6547344
Median	28587
Mode	34166
Standard Deviation	13280.83743
Sample Variance	176380642.9
Kurtosis	-0.93243448
Skewness	-0.22796284
Range	47745
Minimum	2166
Maximum	49911
Sum	5863830
Count	209

Marketing Budget (In Million USD)	
Mean	514.4784848
Standard Error	57.60310168
Median	277.7911334
Mode	#N/A
Standard Deviation	832.7583806
Sample Variance	693486.5205
Kurtosis	38.45188229
Skewness	5.526168879
Range	7414.697905
Minimum	11.59584612
Maximum	7426.293751
Sum	107526.0033
Count	209

We can also visualize the distribution of each independent variable using histograms or density plots.







It is evident from the above discussion that there is enough variation in the values of minimum, maximum, mean, and standard deviation, indicating sufficient variability in the distribution. This suggests that the independent variable is strong.

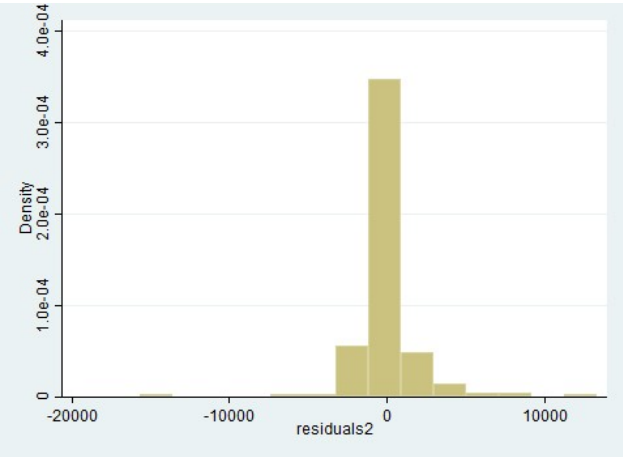
Assumption 3: The mean value of the stochastic error term is zero and it follows a normal distribution.

```
. summarize residuals2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
residuals2	209	-.0000109	2170.63	-15660.54	13365.75

```
. histogram residuals2
(bin=14, start=-15660.539, width=2073.3061)
```

In our case, the mean is approximately -0.0000109. Since it is very close to zero, it suggests that, on average, the residuals are centred around zero, which is in line with the assumption that the mean value of the error term is zero. Additionally, the spread of the residuals around the mean, as measured by the standard deviation, is quite large, indicating variability in the residuals.



The histogram of the residuals does not resemble a bell-shaped curve, indicating that the error term does not follow a normal distribution. We can confirm this by conducting the Shapiro-Wilk W test.

```
. swilk residuals2
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
residuals2	209	0.68711	48.505	8.951	0.00000

The **Shapiro-Wilk test** is a statistical test used to assess whether a sample of data comes from a normally distributed population. The null hypothesis is that the population from which the sample is drawn follows a normal distribution.

In this case, the p-value is 0.00000, we can reject the null hypothesis even at 1% level of significance. The very small p-value suggests that the residuals do not follow a normal distribution. **Therefore, the assumption of normality for the error terms is violated.**

Assumption 4: The values of X_i(s) are fixed over repeated sampling.

It may not be possible to definitively prove the assumption of fixed independent variables.

Assumption 5: The covariance between X_i and U_i is zero.

```
. predict residuals, residuals

. pwcorr SalaryofCEOInThousandUSD ReturnonEquityIn ReturnonStockIn NumberofEmployees MarketingBudgetInMillionUS residuals
```

	Salary~D	Retu~yIn	Retu~kIn	Number~s	Market~S	residu~s
SalaryofCE~D	1.0000					
ReturnonEq~n	0.1148	1.0000				
ReturnonSt~n	-0.0337	0.2749	1.0000			
NumberofEm~s	-0.0344	0.1136	0.0407	1.0000		
MarketingB~S	0.1185	-0.0555	-0.1383	-0.0569	1.0000	
residuals	0.0000	0.0000	0.0000	-0.0000	0.0000	1.0000

Based on the correlation matrix, it appears that there are no significant correlations between the independent variables and the residuals. This suggests that the assumption of zero covariance between the independent variables and the residuals holds in our regression model.

Assumption 6: Number of parameters to be estimated from the model should be much less than the total number of observations in the sample.

We have 209 observations and a total of 7 variables in our model. Since the number of parameters (7) is much less than the number of observations (209), this suggests that the assumption holds in our model.

Assumption 7: The econometric model should be correctly specified. That means there should not be any model misspecification.

Assumption 8: Homoscedasticity - The error variance should be constant.

1. **Breusch-Pagan / Cook-Weisberg test**: This test is used to determine whether there is evidence of heteroskedasticity in a regression model.

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of SalesRevenueInMillionUSD

         chi2(1)      =   968.86
         Prob > chi2  =   0.0000
```

   The null hypothesis for this test is that there is constant variance.
   In the provided results:
   chi2(1): The chi-squared statistic is 968.86.
   Prob > chi2: The p-value is very close to zero (0.0000), we can reject the null hypothesis even at 1% level of significance.
   Since the p-value is less than the significance level (usually 0.05), we reject the null hypothesis of constant variance. Therefore, there is evidence of heteroskedasticity in the regression model.

2. **White 's General Het. Test**

```
. imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(20)      =    130.37
         Prob > chi2   =    0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	130.37	20	0.0000
Skewness	24.94	5	0.0001
Kurtosis	2.53	1	0.1114
Total	157.84	26	0.0000

The null hypothesis is that the errors (or residuals) in the regression model are homoscedastic. In the provided results:
chi2(20): The chi-squared statistic is 130.37.
Prob > chi2: The p-value is very close to zero (0.0000), we can reject the null hypothesis even at 1% level of significance.

Overall, the results suggest strong evidence against homoskedasticity, indicating that the errors in the regression model exhibit unrestricted heteroskedasticity.

Assumption 9: There should not be auto correlation.

1. **Durbin-Watson Test**: The Durbin-Watson test statistic is commonly used to detect first-order autocorrelation in the residuals. The test statistic ranges from 0 to 4, where a value close to 2 suggests no autocorrelation, while values significantly below 2 (less than 1.5) or above 2 (greater than 2.5) indicate positive or negative autocorrelation, respectively.

```
. dwstat

Durbin-Watson d-statistic(  6,    209) =  1.920896
```

In our case, the DW statistic is approximately 1.920896. Since this value is close to 2, it indicates that there is no strong evidence of autocorrelation in the residuals of our regression model.

2. **Breusch-Godfrey Test**: The Breusch-Godfrey test is a more general test for higher-order autocorrelation in the residuals. It assesses whether there is any correlation between the residuals and their lagged values up to a specified order.

```
. estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
```

lags(p)	chi2	df	Prob > chi2
1	0.001	1	0.9715

H0: no serial correlation

Since the p-value (0.9715) is greater than the typical significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no significant evidence of autocorrelation in the residuals at the 5% significance level.

Assumption 10: There should not be Multicollinearity.

```
. corr SalaryofCEOInThousandUSD ReturnonEquityIn ReturnonStockIn NumberofEmployees MarketingBudgetInMillionUS
(obs=209)
```

	Salary~D	Retu~yIn	Retu~kIn	Number~s	Market~S
SalaryofCE~D	1.0000				
ReturnonEq~n	0.1148	1.0000			
ReturnonSt~n	-0.0337	0.2749	1.0000		
NumberofEm~s	-0.0344	0.1136	0.0407	1.0000	
MarketingB~S	0.1185	-0.0555	-0.1383	-0.0569	1.0000

The correlations between the independent variables are relatively small, with the highest being 0.2749 between ReturnonEquityIn and ReturnonStockIn. Generally, correlations below 0.3 or 0.4 are considered weak.

Low correlations between pairs of variables can suggest the absence of multicollinearity, it doesn't guarantee its absence. Multicollinearity can still exist even if correlations are not very high. Additionally, multicollinearity can also occur between three or more variables even if pairwise correlations are low.

To thoroughly assess multicollinearity, further diagnostics such as variance inflation factor (VIF) should be conducted.

VIF Test: The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. It quantifies how much the variance of an estimated regression coefficient is increased due to multicollinearity in the model.

Generally, a VIF greater than 10 is considered high, indicating potential multicollinearity issues.

```
. vif
```

Variable	VIF	1/VIF
SalaryofCE~D	1.00	1.000000
Mean VIF	1.00	

```
. vif
```

Variable	VIF	1/VIF
ReturnonEq~n	1.00	1.000000
Mean VIF	1.00	

```
. vif

    Variable |      VIF       1/VIF
-------------+----------------------
ReturnonSt~n |     1.00    1.000000
-------------+----------------------
    Mean VIF |     1.00
```

```
. vif

    Variable |      VIF       1/VIF
-------------+----------------------
NumberofEm~s |     1.00    1.000000
-------------+----------------------
    Mean VIF |     1.00
```

```
. vif

    Variable |      VIF       1/VIF
-------------+----------------------
MarketingB~S |     1.00    1.000000
-------------+----------------------
    Mean VIF |     1.00
```

Based on the above results, we can conclude that multicollinearity is not present in the model overall.

Conclusion

1. The assumption of normality for the stochastic (disturbance) term Ui is violated.
2. Unable to substantiate the assumption of fixed independent variables, suggesting potential variability in the explanatory variables over repeated sampling.
3. The assumption of homoscedasticity is violated, indicating unequal variances in the error terms, which may compromise the efficiency and reliability of parameter estimates.
4. Our analysis provides no definitive proof that the model is correctly specified, highlighting the challenges inherent in capturing the true underlying relationships between variables.