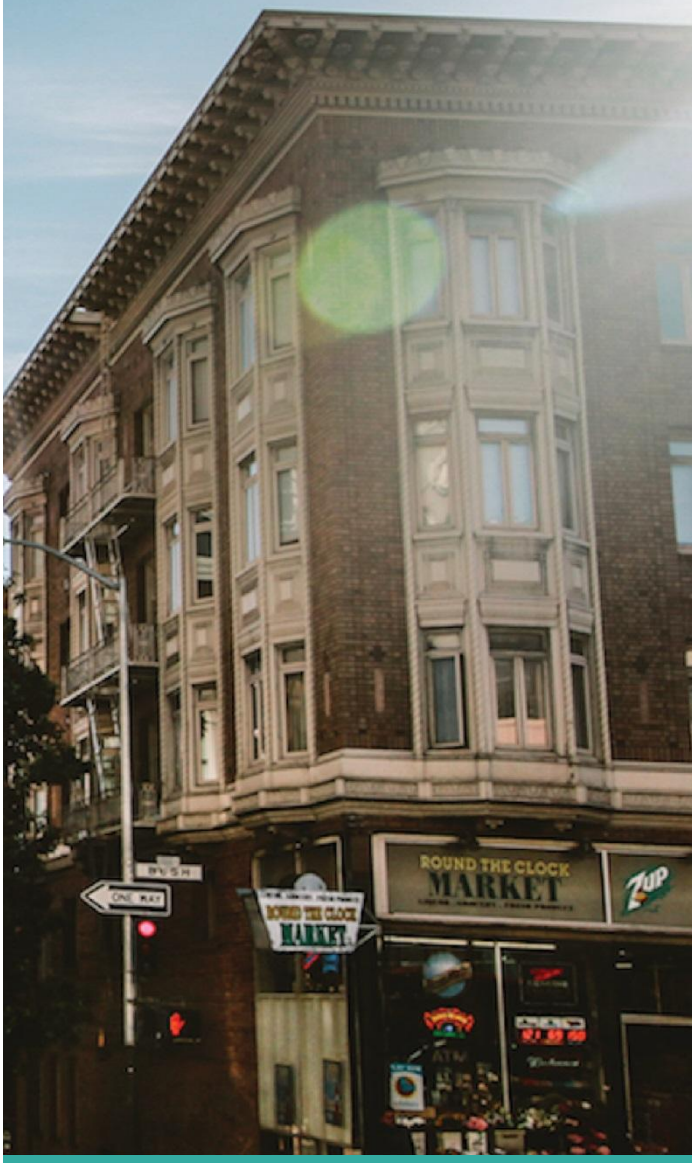
LAB 2: ADVANCED REGRESSION ANALYSIS



OCTOBER 22

BSAN6070 INTRODUCTION TO MACHINE LEARNING

Authored by: Andres Sebastian Gaibor Heredia

1. **Based on your analysis from Lab 1, modify your model accordingly and apply it to Lab 2. Examine all diagnostic tests, regression results, and graphs provided in Lab 2.**

1.1 Descriptive Analytics

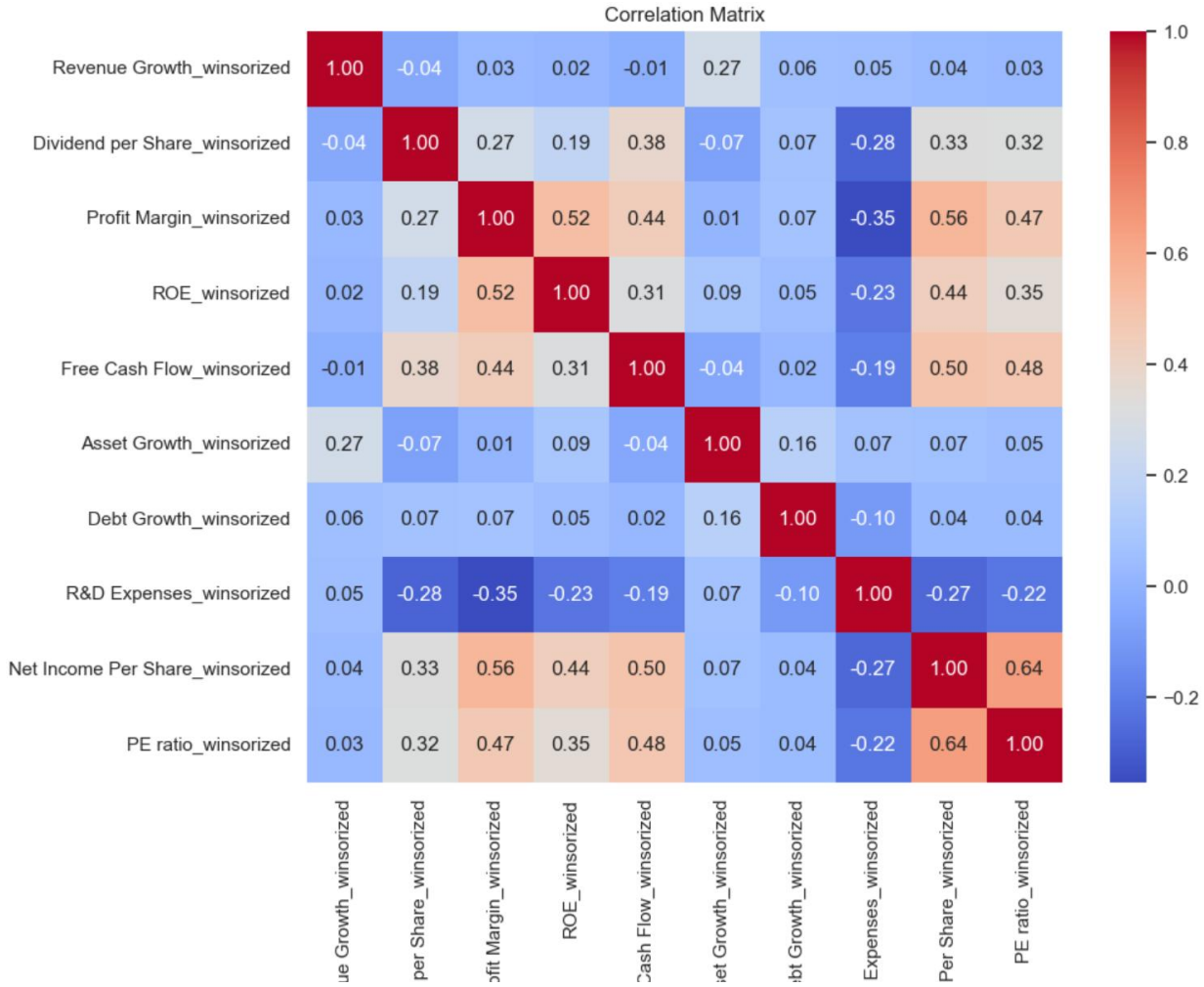
First, we will import the data from 2018 from 4,391 different companies among 225 different financial metrics (fields). We created the same subset of fields as the previous lab with 9 different independent variables (X): "revenue growth", "dividend per share", "profit margin", "return on equity", "free cash flow", "asset growth", "debt growth", "r&d expenses", "net income per share" and the dependent variable (y) "pe ratio". There were missing values on each of the fields as shown below:

```
Missing values in each column:
revenue growth          139
dividend per share      250
profit margin           306
roe                     256
free cash flow          167
asset growth            214
debt growth             264
r&d expenses            237
net income per share    257
pe ratio                252
dtype: int64
```

Hence, we had to clean the data to continue with the modelling of future steps. After addressing the missing values we know that there will be a large amount of outliers so we winsorize the data using percentile 1st and 99th. Afterwards, we use the log transformation lambda x: np.log(x+1) to make data easier to analyze, and to avoid errors resulting from log 0.and obtain the statistical summary of the dataset shown below:

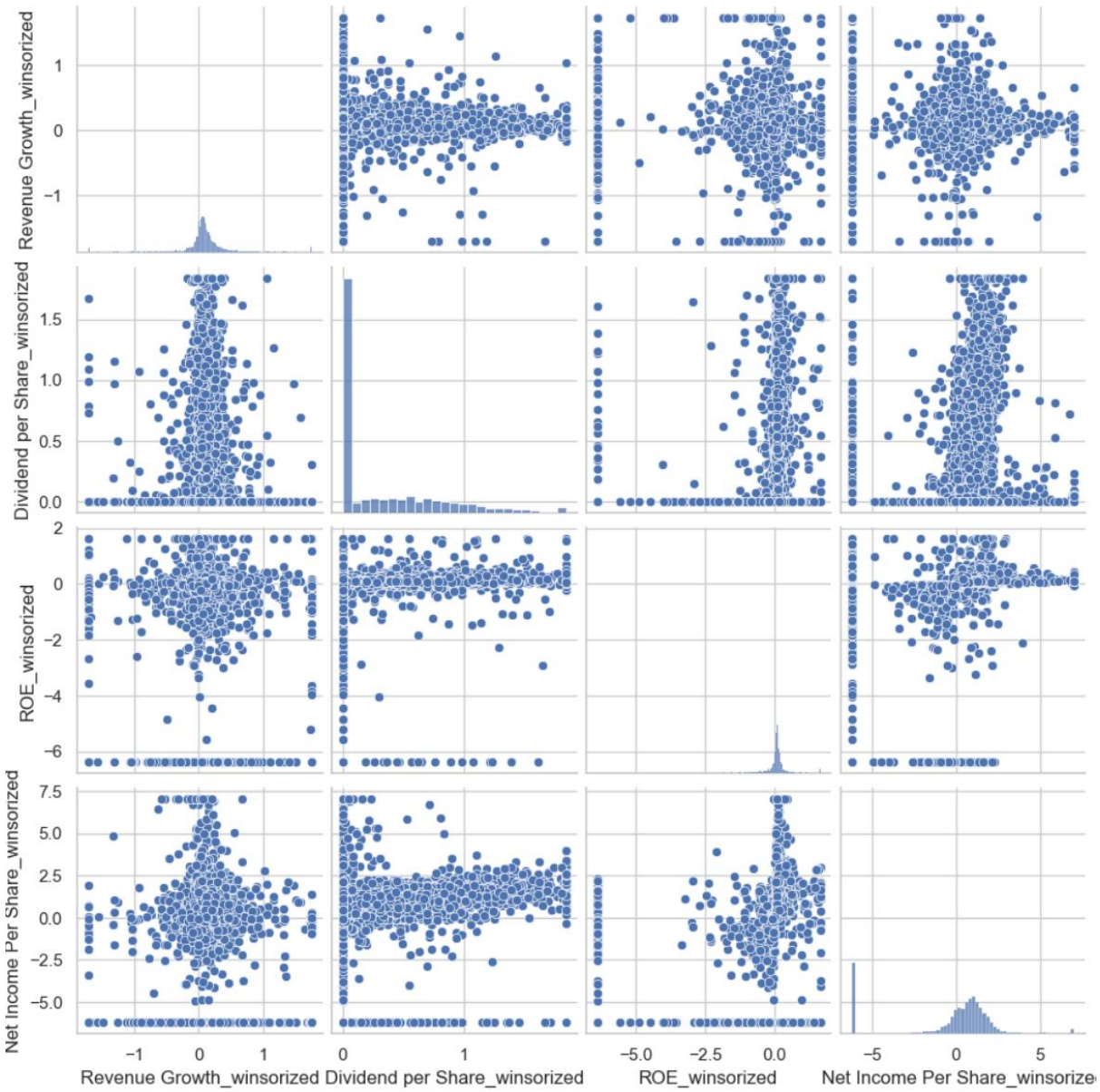
```
Statistical summary of the dataset:
       Revenue Growth_winsorized  Dividend per Share_winsorized  \
count                3998.000000                    3998.000000
mean                    0.097000                       0.332664
std                     0.365072                       0.458263
min                    -1.710364                       0.000000
25%                     0.000000                       0.000000
50%                     0.075849                       0.000000
75%                     0.175402                       0.601580
max                     1.743532                       1.840550

       Profit Margin_winsorized  ROE_winsorized  Free Cash Flow_winsorized  \
count                3625.000000     3770.000000                2677.000000
mean                        -inf        0.007992                  18.507182
std                          NaN        0.523021                   2.139105
min                         -inf       -6.377127                  10.307855
25%                   -0.001601       -0.013060                  17.181461
50%                    0.055435        0.081257                  18.547313
75%                    0.147558        0.151562                  19.918534
max                    0.583053        1.627494                  23.222195

       Asset Growth_winsorized  Debt Growth_winsorized  \
count              3998.000000             3998.000000
mean                  0.073369                    -inf
std                   0.270152                     NaN
min                  -0.765503                    -inf
25%                  -0.034721               -0.089979
50%                   0.035029                0.000000
75%                   0.146046                0.115981
max                   1.124572                2.607124

       R&D Expenses_winsorized  Net Income Per Share_winsorized  \
count              3998.000000                      3352.000000
mean                  6.140817                         0.884565
std                   8.391752                         1.331258
min                   0.000000                        -6.214608
25%                   0.000000                         0.188241
50%                   0.000000                         0.856838
75%                  16.374620                         1.448165
max                  21.541430                         7.042592

       PE ratio_winsorized
count          3998.000000
mean              1.994224
std               1.553912
min               0.000000
25%               0.000000
```

Then we run the correlation matrix between the fields to check if any correlation is present among the explanatory variables. As we see below in the matrix, most correlation values are close to 0, indicating little to no linear relationship between predictors.



The highest correlation is between Net Income Per Share_winsorized and PE ratio winsorized (0.64), which is still moderate.
Conclusion: There is no multicollinearity concern. No need to remove or combine variables.

We then plot the pairplot of 5 selected variables and taking a sample of 1,000 records to visualize the relationship between them. We select Revenue Growth Winsorized, Dividend Per Share Winsorized, ROE Winsorized and Net Income Per Share Winsorized. The resulting plot can be seen below:



1.2 Feature scaling

We use StandardScaler() and fit_transform(X) to transform all features into a standard normal distribution (mean = 0, standard deviation = 1). This change ensures all features have the same scale (important for models that are sensitive to feature magnitudes, e.g., linear regression, k-means, PCA). It also prevents features with larger scales from dominating the model.

Handling multicollinearity

After running the VIF on all 9 financial features we note that all VIF values are below 5, meaning none of the independent variables are highly correlated. The highest VIF (1.846) is for "Profit Margin Winsorized", which is still very low and does not indicate any collinearity concerns. The DataFrame for VIF > 5 is empty, confirming no multicollinearity exists.

Conclusion: The predictors are independent of each other and do not introduce redundancy into the model. The regression coefficients will be stable, meaning the model will not suffer from

inflated standard errors due to multicollinearity. Since multicollinearity is not an issue, the interpretation of the regression coefficients remains valid.

1.3 Splitting the Data

We divide the data into training data and testing data, this ensures the model is trained on one portion of the data and tested on unseen data. It also prevents overfitting, where a model memorizes the training data but fails on new data. We used a 80-20 split for training and testing data respectively. We include a random_state to ensure results are reproducible.
The training set size is 3,198 records, while the testing set size is 800 records.

1.4 OLS results

After running OLS with the standardized features, we can note that the results seem to have improved significantly.

OLS Regression Model Summary:

OLS Regression Results

Dep. Variable:	PE ratio_winsorized	R-squared:	0.459
Model:	OLS	Adj. R-squared:	0.457
Method:	Least Squares	F-statistic:	300.0
Date:	Fri, 07 Mar 2025	Prob (F-statistic):	0.00
Time:	09:26:53	Log-Likelihood:	-4978.0
No. Observations:	3198	AIC:	9976.
Df Residuals:	3188	BIC:	1.004e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.9923	0.020	97.988	0.000	1.952	2.032
Revenue Growth_winsorized	-0.0040	0.021	-0.192	0.848	-0.045	0.037
Dividend per Share_winsorized	0.1155	0.023	5.054	0.000	0.071	0.160
Profit Margin_winsorized	0.1641	0.027	5.976	0.000	0.110	0.218
ROE_winsorized	0.0413	0.024	1.718	0.086	-0.006	0.088
Free Cash Flow_winsorized	0.2680	0.025	10.607	0.000	0.218	0.317
Asset Growth_winsorized	0.0413	0.021	1.950	0.051	-0.000	0.083
Debt Growth_winsorized	-0.0136	0.021	-0.651	0.515	-0.055	0.027
R&D Expenses_winsorized	-0.0010	0.022	-0.043	0.966	-0.045	0.043
Net Income Per Share_winsorized	0.7105	0.027	26.189	0.000	0.657	0.764

Omnibus:	85.142	Durbin-Watson:	2.038
Prob(Omnibus):	0.000	Jarque-Bera (JB):	103.929
Skew:	0.336	Prob(JB):	2.70e-23
Kurtosis:	3.574	Cond. No.	2.68

We note that R-squared suggests the Winsorized-log model explains still ~46% of the variance in the P/E ratio. The F-statistic increased to 0.843 and still confirms the model is not statistically significant.
From all features, we see that there is many insignificant variables as their p-value > 0.05. We then have to drop Revenue Growth_winsorized, ROE_winsorized, Asset Growth_winsorized, Debt Growth_winsorized and R&D Expenses_winsorized. This is different from our previous model where we did not apply the log transformation and winsorization where R&D Expenses was the only significant variable.
After adjusting for insignificant variables, we conduct again the OLS regression where we have the following summary of results

OLS Regression Model Summary:

OLS Regression Results

Dep. Variable:	PE ratio_winsorized	R-squared:	0.457
Model:	OLS	Adj. R-squared:	0.457
Method:	Least Squares	F-statistic:	672.4
Date:	Fri, 07 Mar 2025	Prob (F-statistic):	0.00
Time:	09:56:43	Log-Likelihood:	-4981.9
No. Observations:	3198	AIC:	9974.
Df Residuals:	3193	BIC:	1.000e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.9921	0.020	97.942	0.000	1.952	2.032
Dividend per Share_winsorized	0.1116	0.022	4.995	0.000	0.068	0.155
Profit Margin_winsorized	0.1797	0.025	7.249	0.000	0.131	0.228
Free Cash Flow_winsorized	0.2662	0.025	10.592	0.000	0.217	0.316
Net Income Per Share_winsorized	0.7240	0.027	27.222	0.000	0.672	0.776

Omnibus:	85.476	Durbin-Watson:	2.041
Prob(Omnibus):	0.000	Jarque-Bera (JB):	103.908
Skew:	0.338	Prob(JB):	2.73e-23
Kurtosis:	3.568	Cond. No.	2.31

Model Fit & Performance

R-squared: 0.457 (Adjusted R^2: 0.457): The model explains 45.7% of the variance in PE ratio_winsorized, indicating a moderate fit. Adjusted R^2 is the same as R^2, meaning that the removal of variables did not negatively impact the model's explanatory power. The model is likely simpler and more interpretable without unnecessary predictors. This suggests that the retained predictors (Dividend per Share, Profit Margin, Free Cash Flow, and Net Income Per Share) are the key drivers of PE ratio_winsorized.

Statistical Significance of Variables

A variable is considered statistically significant if its p-value (P>|t|) is below 0.05. All remaining variables are significant ($p < 0.05$):

Dividend per Share_winsorized ($p = 0.000$)

Profit Margin_winsorized ($p = 0.000$)

Free Cash Flow_winsorized ($p = 0.000$)

Net Income Per Share_winsorized ($p = 0.000$)

The model now only includes significant variables, improving reliability. All four predictors are strongly associated with PE ratio_winsorized.

1.5 Ridge Regression

This time around, we are conducting a Ridge Regression using the Winsorized log-transformed and standardized fields. The Ridge Regression was applied to the dataset using cross-validation (CV=5) and grid search to find the best regularization parameter (alpha). The best alpha value is 10.0

Ridge regression introduces penalty terms to shrink coefficients and prevent overfitting. The model found that alpha=10.0 is optimal, balancing bias and variance trade-off.

R^2 = 0.4732 means that 47.32% of the variation in the target variable is explained by the predictors. This is slightly better than the OLS model (which had $R^2 \approx 0.457$), meaning Ridge improved generalization.

Key Takeaway: Ridge Regression slightly improved model performance compared to OLS. This suggests that some overfitting was present in the OLS model, and Ridge helped correct it.

1.6 Lasso Regression

Lasso applies L1 regularization, which shrinks some coefficients to exactly 0, effectively performing feature selection.

A small alpha=0.01 means minimal regularization was applied, allowing most important variables to retain significant values.

MSE 1.2336 Very similar to Ridge (1.2335), meaning no major performance loss despite feature selection.

R^2 0.4732 Same as Ridge, explaining 47.32% of variance in PE ratio_winsorized.

1.7 Elastic Net Regression

alpha: 0.1: Controls the overall regularization strength.

A low l1_ratio (0.1): means this model leans more toward Ridge regression but still includes some feature selection.

This model prefers keeping most features but applies slight sparsity, meaning it retains relevant predictors while reducing overfitting.

MSE 1.2436 Slightly higher than Lasso/Ridge, but within an acceptable range.

R^2 0.4689 Explains 46.89% of variance in PE ratio_winsorized, slightly lower than Ridge and Lasso (which had ~0.4732).

Elastic Net has slightly lower R^2 than Lasso/Ridge but still provides robust generalization.

The small trade-off in performance may be acceptable if reducing overfitting is a priority.

2. Model Comparison

By making a comparison of the four models OLS, Ridge regression, Lasso and Elastic Net we see the following output:

Model Comparison:

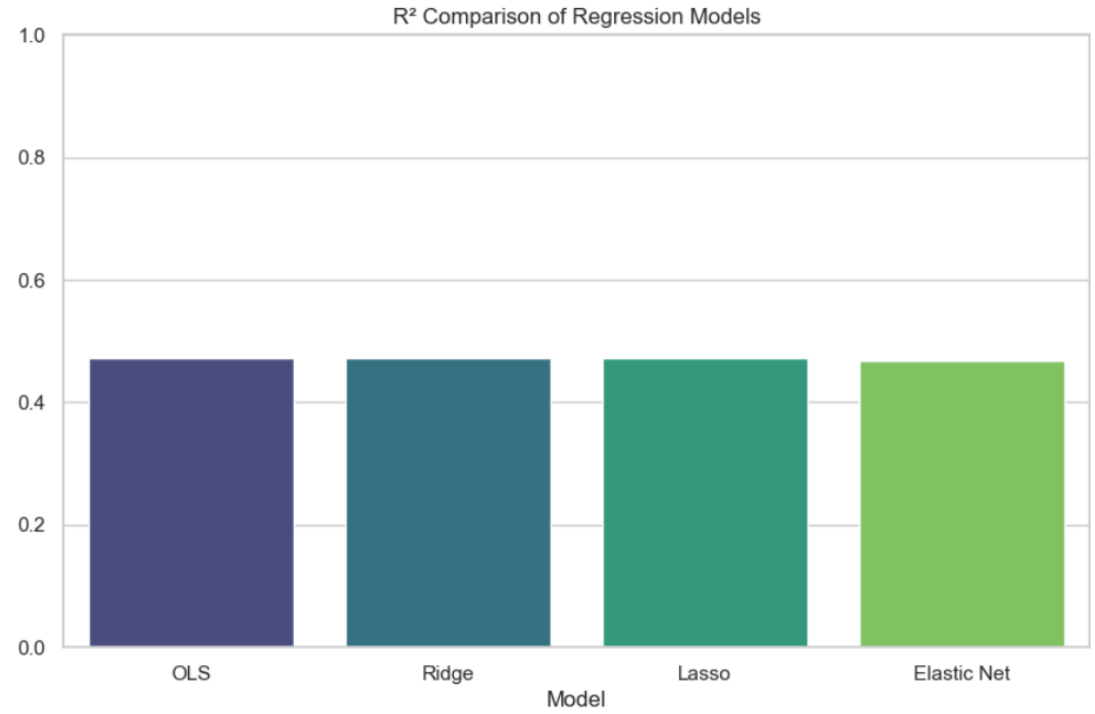
	Model	MSE	R^2
0	OLS	1.233321	0.473303
1	Ridge	1.233522	0.473218
2	Lasso	1.233577	0.473194
3	Elastic Net	1.243567	0.468928

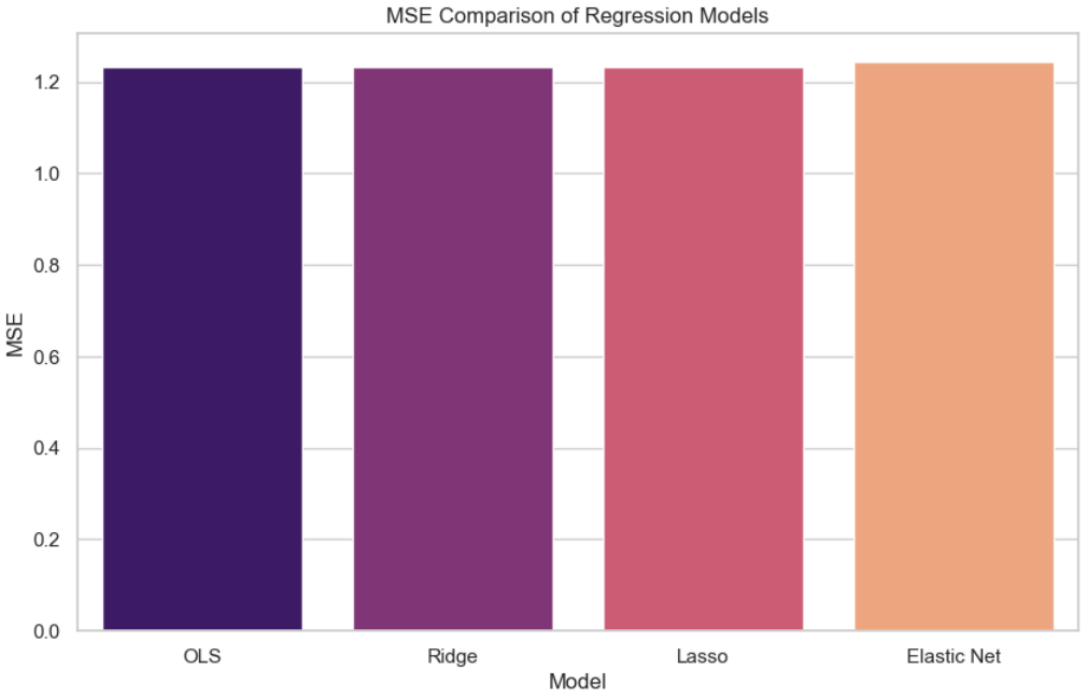
OLS has the best performance (Lowest MSE, Highest R^2)

OLS (Ordinary Least Squares) regression achieved the lowest MSE (1.2333) and the highest R^2 (0.4733). Indicates that no significant overfitting was present, so regularization did not provide a major benefit. If interpretability is the goal, OLS is the best choice.

Ridge and Lasso Perform Similarly to OLS: Ridge (MSE: 1.2335, R^2: 0.4732) and Lasso (MSE: 1.2336, R^2: 0.4732) perform almost identically to OLS. Ridge keeps all features, while Lasso performs feature selection, but this did not significantly impact accuracy.

Elastic Net Performed Slightly Worse: Elastic Net (MSE: 1.2436, R^2: 0.4689) performed slightly worse than other models. This suggests that the mix of L1 and L2 regularization was not optimal for this dataset. While it still provides feature selection benefits, it led to slight underperformance.



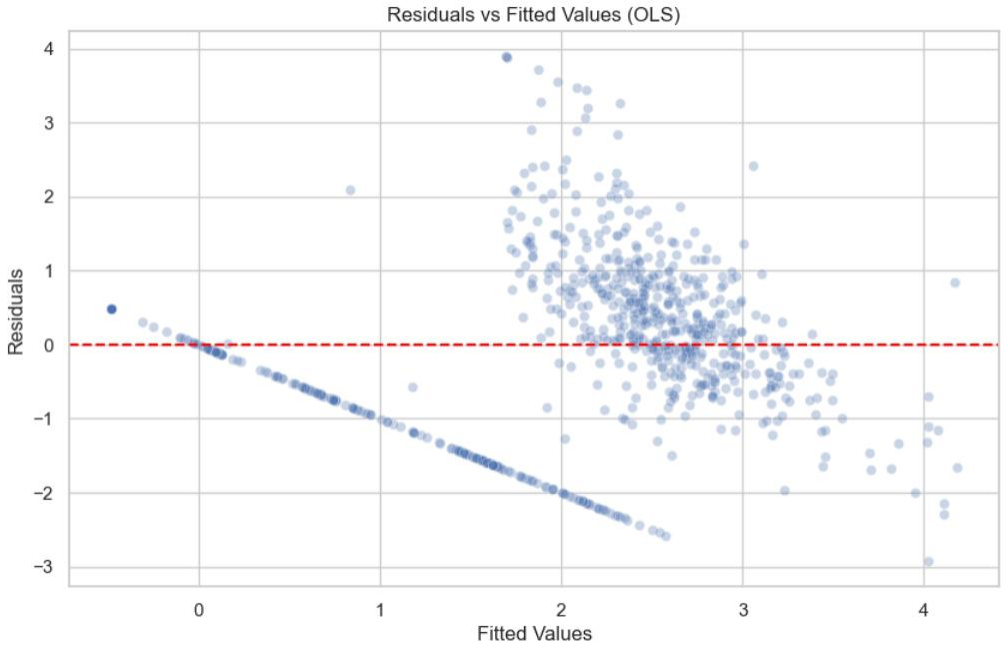


2.1 Analysis using the "best" model

By comparing MSE between the four models we can determine that the smallest (1.233321) corresponds to OLS regression, also the R-squared is slightly the highest (0.473303), therefore we choose it as the best model among the four.

Residual vs Fitted Values plot

This Residuals vs. Fitted Values plot is used to assess the assumptions of Ordinary Least Squares (OLS) regression and the overall fit of the model. The goal is to check if residuals (errors) are randomly distributed, which would indicate that OLS is a good model for the data.



Key Observations

The Y-axis (Residuals) represents the difference between actual values and predicted values.
The X-axis (Fitted Values) represents the predicted values from the OLS model.
The red dashed line represents the zero residual baseline—ideally, residuals should be randomly scattered around this line.
Pattern in Residuals (Signs of Model Issues): A downward trend in residuals at lower fitted values (left side), meaning the model systematically underpredicts or overpredicts for small-fitted values. Increasing variance of residuals at higher fitted values (right side), indicating potential heteroscedasticity (non-constant variance in errors).

Key Takeaway

The model may be missing key nonlinear relationships or not fully capturing variance in the data. The residual pattern suggests OLS may not be the best fit, as it assumes a linear relationship between predictors and the target variable.

Durbin-Watson Statistic: The DW statistic ranges from 0 to 4, where:

2	indicates no autocorrelation.
0 to <2	suggests positive autocorrelation.
>2 to 4	suggests negative autocorrelation.

In this context, a DW statistic of 2.0992 is very close to 2, indicating that there is no significant autocorrelation in the residuals. This suggests that the residuals are independent, which is an assumption of the linear regression model.

Breusch-Pagan Test: The Breusch-Pagan (BP) test is used to detect heteroscedasticity in a regression model—whether the variance of residuals changes across different values of the independent variables.

Lagrange Multiplier Statistic	62.4454	A high test statistic suggests heteroscedasticity (unequal variance in residuals).
p-value	0.0000	Since $p < 0.05$, we reject the null hypothesis of homoscedasticity.

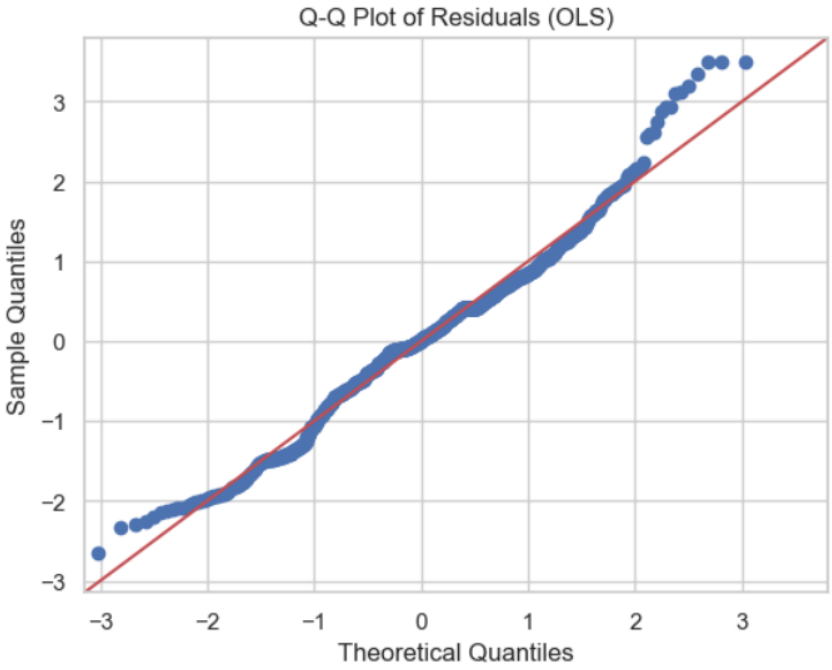
F-statistic	16.8273	The F-test confirms the presence of heteroscedasticity.
f p-value	0.0000	Again, since p < 0.05, heteroscedasticity is statistically significant.

Key Takeaways

Heteroscedasticity is present: The variance of residuals is not constant across fitted values. This violates a key OLS assumption, which assumes homoscedasticity. Standard OLS regression assumes homoscedasticity, when this assumption is violated:

- Coefficient estimates remain unbiased.
- Standard errors become unreliable leading to incorrect p-values and confidence intervals.
- Hypothesis testing (t-tests, F-tests) may be misleading.

Q-Q Plot and Shapiro-Wilk Test: By running the q-q plot on the Lasso regression we can obtain the following plot:




```
Shapiro-Wilk Test Statistic for OLS: 0.9835
p-value: 0.0000
```

This Q-Q (Quantile-Quantile) plot and Shapiro-Wilk test are used to assess whether the residuals from the OLS regression follow a normal distribution, which is an important assumption for OLS regression.

Interpretation of the Q-Q Plot

The X-axis (Theoretical Quantiles) represents the expected quantiles from a normal distribution. The Y-axis (Sample Quantiles) represents the quantiles of the residuals from the OLS model. The red diagonal line represents a perfect normal distribution. If residuals follow a normal distribution, the points should closely follow this line.

Near-Normality in the Middle Range

Most of the residuals align well with the red line in the middle section. This suggests approximate normality for moderate values.

Deviations at the Tails (Outliers): Bottom left (Lower Tail): Some residuals are lower than expected, suggesting slight left-skewness.
Top right (Upper Tail): There are multiple large residuals, indicating the presence of outliers or heavy tails (right-skewness).

Interpretation of the Shapiro-Wilk Test
The Shapiro-Wilk test confirms that residuals are not normally distributed (p = 0.0000), we reject the null hypothesis of normality. Even though the Q-Q plot looks close to normal in the middle, the deviations in the tails are statistically significant. Non-normal residuals suggest that hypothesis testing results (t-tests, F-tests) may be unreliable.

VIF: All VIF values for the independent variables are close to 1, meaning that there is minimal collinearity between the predictors.
This time, the highest VIF value (for Net Income Per Share Winsorized) is 1.55, which is still below the common VIF threshold of 5 for concern.
Implication: The model does not suffer from multicollinearity issues, meaning each predictor contributes unique information. The coefficients in the regression are stable and not distorted by correlation between variables.