# Fourth Industrial Summer School

## Module 4: ML

## Supervised Learning: Regression

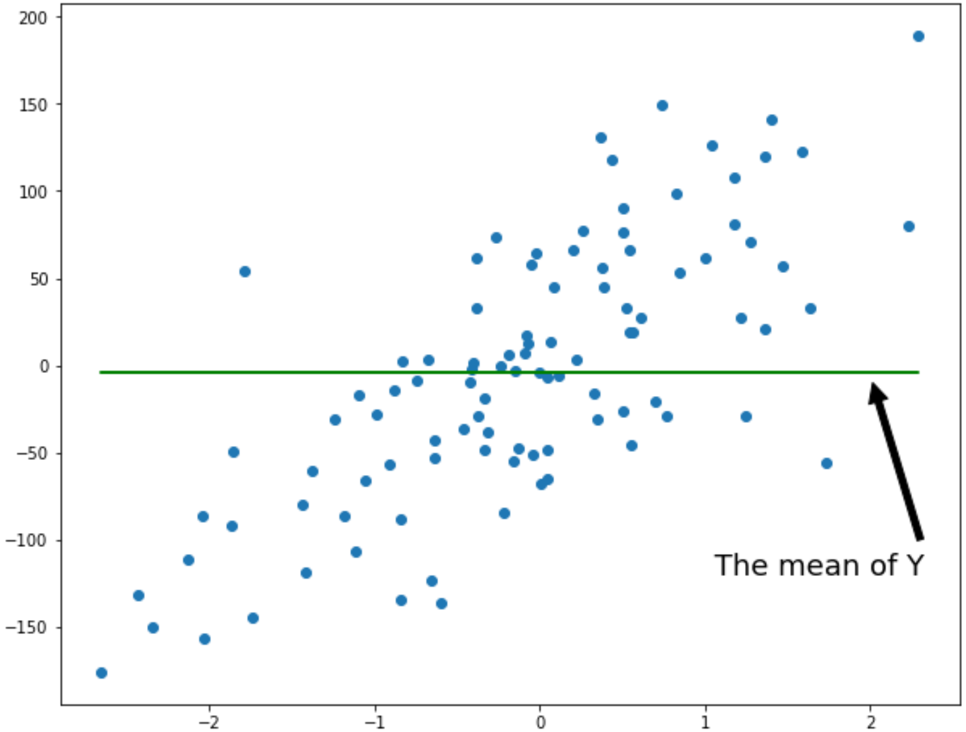
Friday, July 5, 2024

# Outlines

- ✓ How good our model is?
- ✓ Evaluation Metrics

# Coefficient of Determination

- We need to know the total variability present in the data!
- The measure of total variation in the Y is (SST)

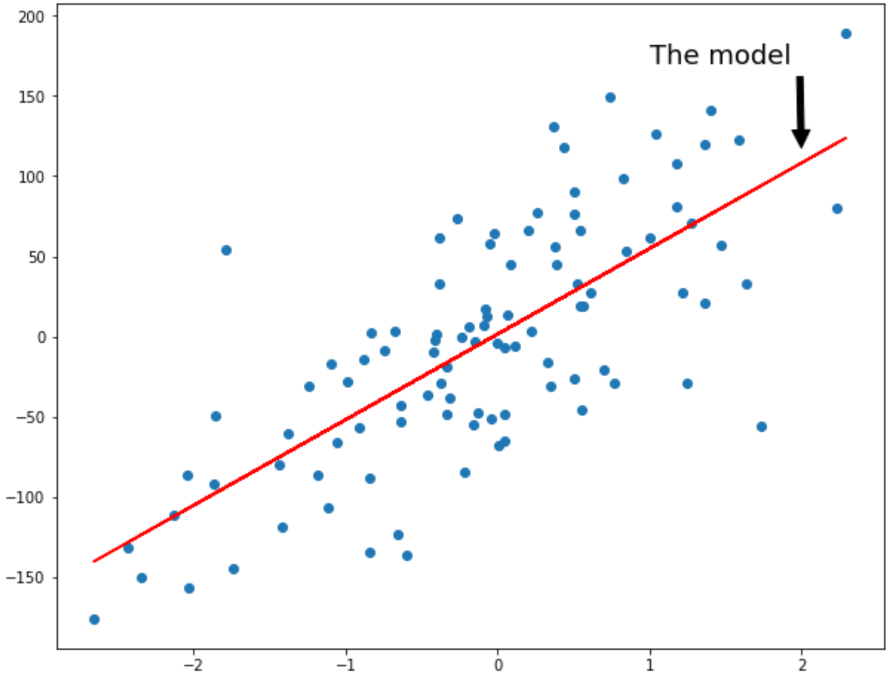



$$SST = \sum(y_i - \bar{y})^2$$

where $\bar{y}$ is the sample mean of Y variable

# Coefficient of Determination

- Since SSE is the optimal sum of squared error of any linear model! Hence SSE is always smaller than SST



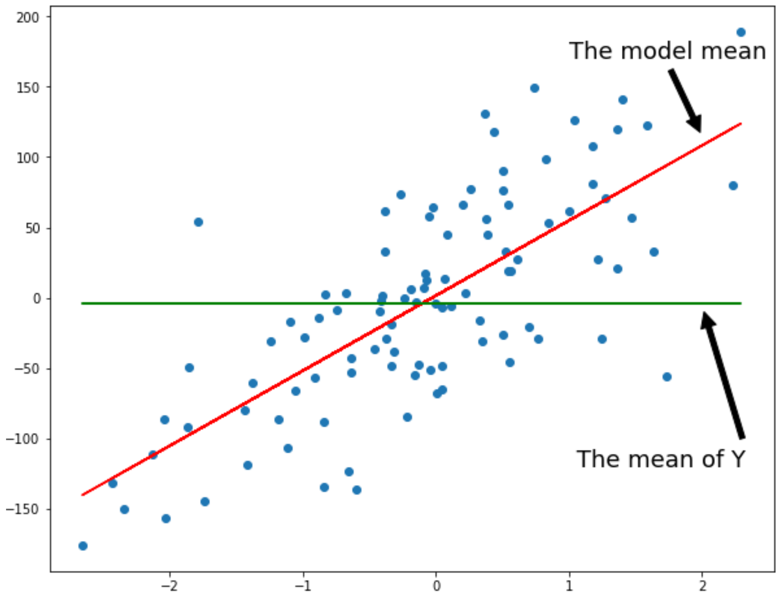

$$SSE = \sum(y_i - \hat{y}_i)^2$$

*where* $\hat{y} = (\beta_0 + \beta_1 x_1)$

- ***SSE*** *is the total variation that is* ***not*** *described by the model line!*
- ***How much of the variation in y described or explained by the variation by the model line?***
- *The answer to this question is the called the* ***r-squared score*** *or* ***coefficient of determination***

# Evaluation: How well our model is?

- **Does the red line fit the data better than the mean (green line)?** If so, how much better
- R-squared is the **proportion** of **variation explained** by the model.
- R-squared can be used to understand the power of the predictions




$$R^2 = 1 - \frac{\mathbf{SSE}}{\mathbf{SST}}$$

where $\frac{\mathbf{SSE}}{\mathbf{SST}}$ is the proportion of the variation that is not described by the model!

# Evaluation: How well our model is?

- Let's suppose that SST = 50, and SSE = 30

$$R^2 = 1 - \frac{30}{50} = 0.4 = 40\%$$

- The model explained 40% variation of the total variation in y

✓ In what case $R^2$ will approach 1.0 ( or 100%) ?

✓ In what case $R^2$ will approach 0.0 ( 0%) ?

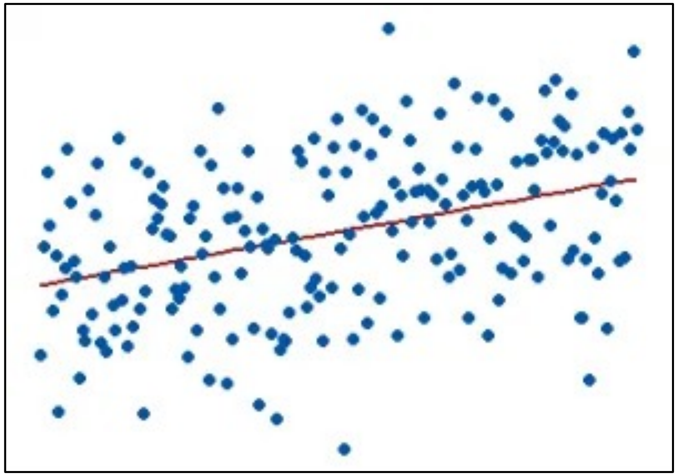
- What is a good value for $R^2$?

# Is that a "good" R-squared value?

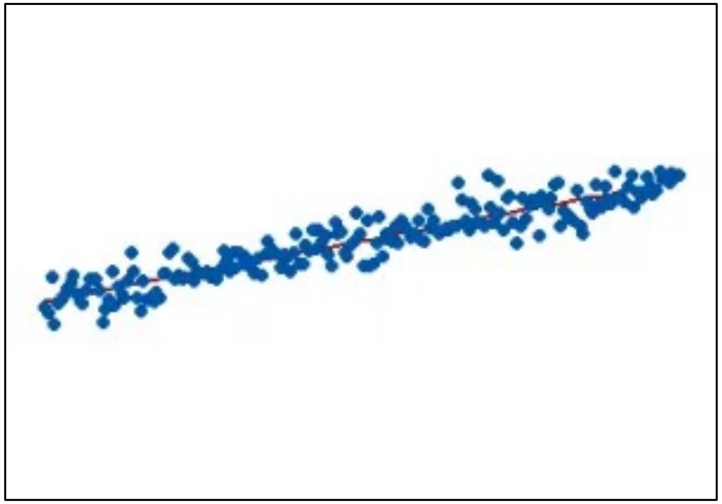
- It's tough to answer this question
- A good R-squared value depends widely on the problem domain
- It is useful as a tool for comparing different models

# Different datasets r-squared

$R^2 = 15\%$



$R^2 = 85\%$



❑ Usually, the larger the better, but not always it depends on the problem data

# What is a good value for $R^2$

- The **higher R-squared, the better the model**
- The threshold for a good R-squared value depends on the domain!
- In other words, the nature of the spread in the dependent variable may vary according to the data.
- R-squared is a fraction by which the variation of the errors is less than the variation of the dependent variable.
- But it is useful as a tool for **comparing different models**

# Sklearn: r2_score

- We can access the r-squared metric library in Sklearn as follows:

```
from sklearn.metrics import r2_score
```

- It ranges (**negative , 1]**
  - R-squared with a value of 1 means the model explains all the variation of the dependent variable.
  - A value of 0 means a bad model.
  - negative values means the model is arbitrary worse than the **simple mean model!**

# R-bar-squared (adjusted)

- As we said, the greater the value of R-squared the btter!
- R-squared could be misleading! Why?
- Because it tends to increase by adding more independent variables (features)
- For multi-feature regression, adding more features and noticing an increase in R-squared is not always a better model than the fewer feature one.
- This take us to compute and look at the adjusted **R-squared**

# R-bar-squared (adjusted)

- Adjusted R-Squared has the form of:

$$\bar{R}^2 = 1 - \frac{(1 - R^2)\times(n - 1)}{n - K - 1}$$

  - where K is number of independent variables, n is the number of samples
- $\bar{R}^2$ deals with additional indpendent variables
- It penalize the value of the R-squared if we add junk new independent variables!

Note: Adjusted R-squared is not implemented in sklearn. But you can develop a function to compute it ( if you need to )

# Performance Evaluation metrics

- There are well known metrics to evaluate the modeling predictions.
  - Mean Square Error(MSE)/Root Mean Square Error(RMSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

  - Mean Absolute Error(MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

  - Root Mean Squared Error

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

# Robustness of MAE, MSE and RMSE

| No Noise | | | |
|---|---|---|---|
| ID | Error | ABS ERROR | SQ Error |
| 1 | 2 | 2 | 4 |
| 2 | 2 | 2 | 4 |
| 3 | 2 | 2 | 4 |
| 4 | 2 | 2 | 4 |
| 5 | 2 | 2 | 4 |
| 6 | 2 | 2 | 4 |
| 7 | 2 | 2 | 4 |
| 8 | 2 | 2 | 4 |
| 9 | 2 | 2 | 4 |
| 10 | 2 | 2 | 4 |

| MAE | RMSE | MSE |
|---|---|---|
| 2.00 | 2.00 | 4.00 |

| 1/2 samples with Noise | | | |
|---|---|---|---|
| ID | Error | ABS ERI | SQ Erro |
| 1 | 3 | 3 | 9 |
| 2 | 3 | 3 | 9 |
| 3 | 3 | 3 | 9 |
| 4 | 3 | 3 | 9 |
| 5 | 3 | 3 | 9 |
| 6 | 1 | 1 | 1 |
| 7 | 1 | 1 | 1 |
| 8 | 1 | 1 | 1 |
| 9 | 1 | 1 | 1 |
| 10 | 1 | 1 | 1 |

| MAE | RMSE | MSE |
|---|---|---|
| 2.00 | 2.24 | 5.00 |

| skewed data | | | |
|---|---|---|---|
| ID | Error | ABS ERI | SQ Erro |
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 |
| 10 | 20 | 20 | 400 |

| MAE | RMSE | MSE |
|---|---|---|
| 2.00 | 6.32 | 40.00 |

Same error magnitude, MAE is stable

RMSE increases by increasing error magnitude

MSE is affected much with outliers, still MAE stable, while RMSE shows an increase

# Sklearn: Error score

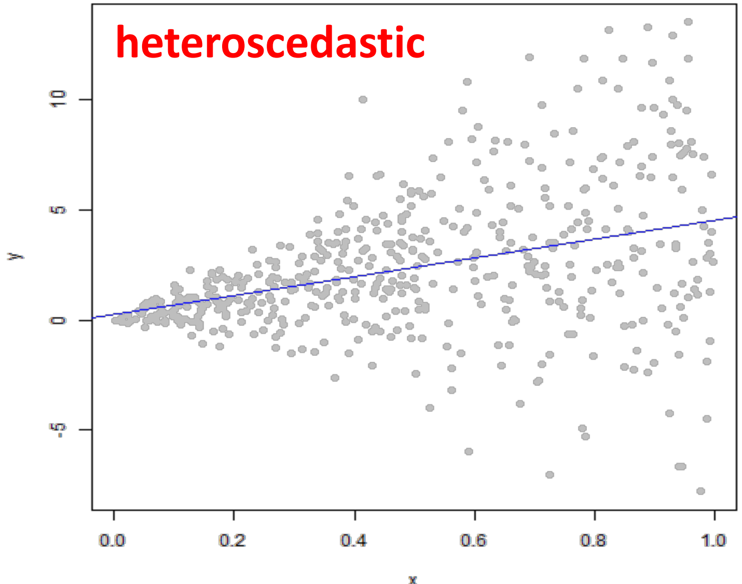
- We can compute the **MSE, MAE** using sklearn library as:

```
# other evaluation metrics
from sklearn.metrics import mean_squared_error, mean_absolute_error
print(mean_absolute_error(Y_test, Y_predicted))
print(mean_squared_error(Y_test, Y_predicted))
```

- The **RMSE** can be simply computed by taking the sqrt of the MSE!

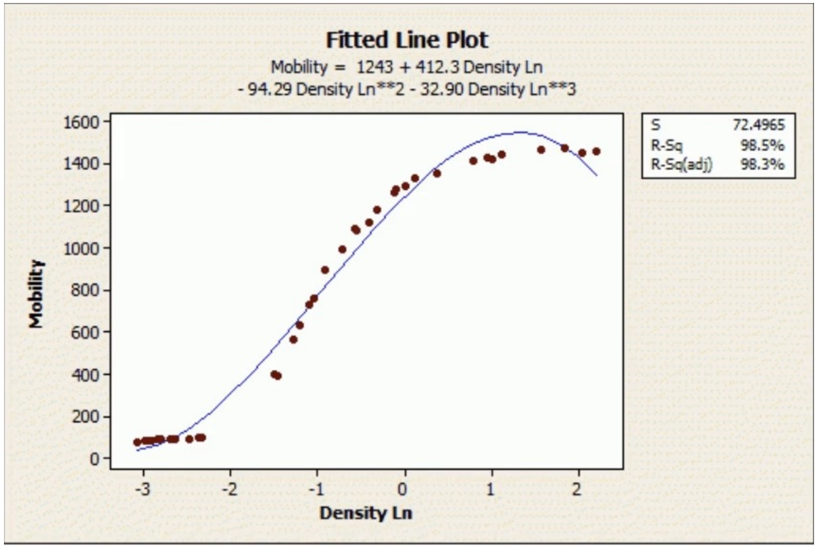
# Evaluation: Residual Plots

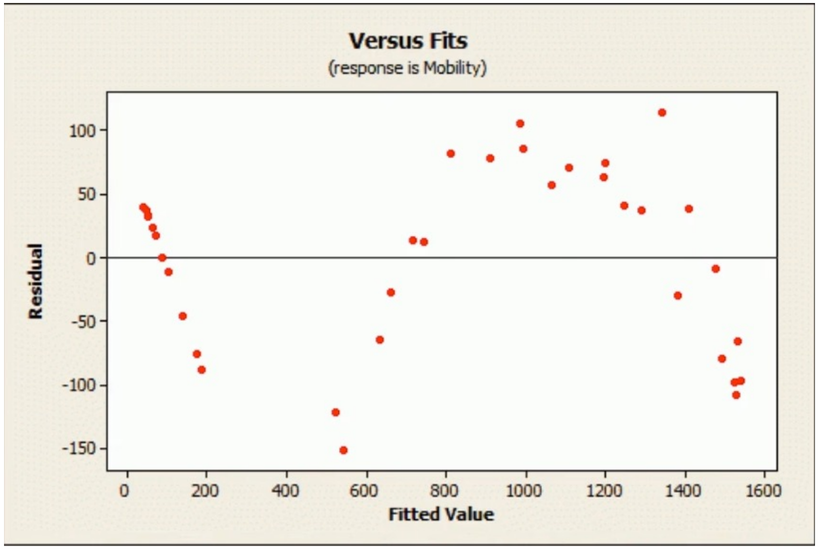
- A residual plot is **typically used to find problems with the regression data**.
- As some datasets could not be good for regression, including Heteroscedastic data (see figure).
- In heteroscedastic data, our error assumption is violated, the error increases by increasing the value of the independent variable



# Residual Plots

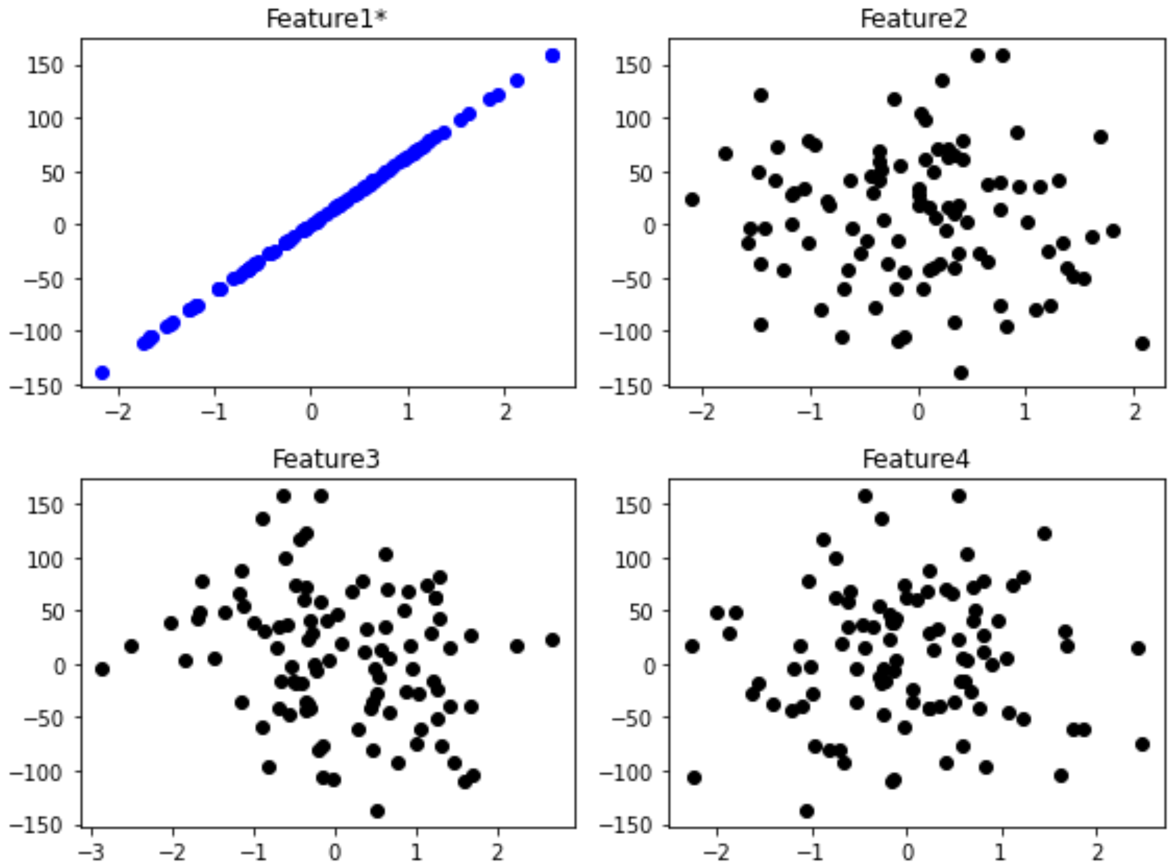
- Even though **R-squared** is high, we may need to plot the residual plot to figure out if we have a problem



# Example

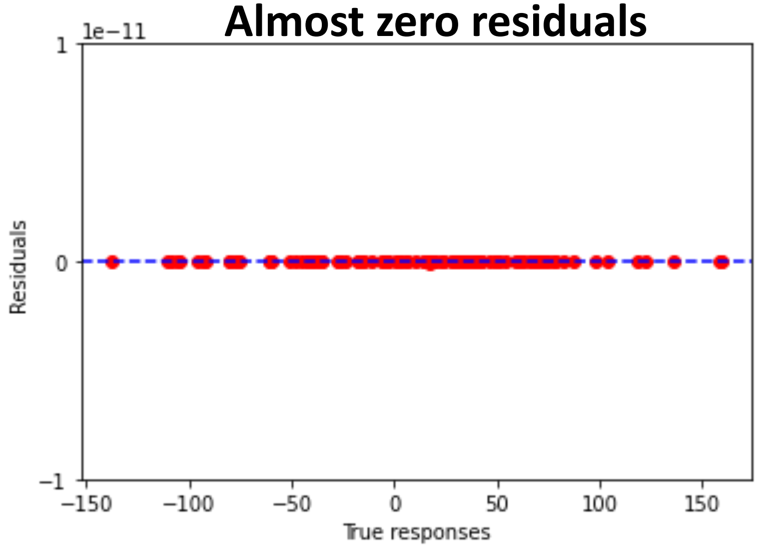
- We generated 4 features with one informative, as shown in the plots below



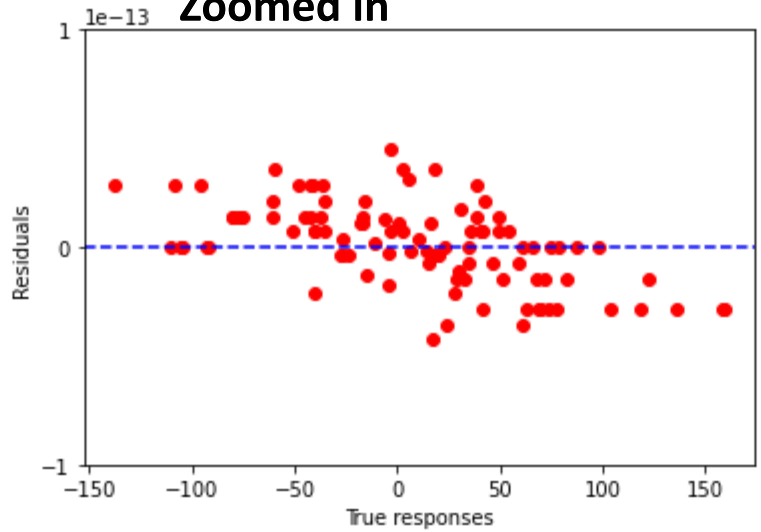
# Results

- trained model achieved $R^2 \approx 1$,
- The model parameters showed 3 features are insignificant
- model parameter: [**6.36487032e+01,** -2.08506557e-15 1.51179609e-14, -5.13041457e-15]
- The residual plot shows good results

Almost zero residuals

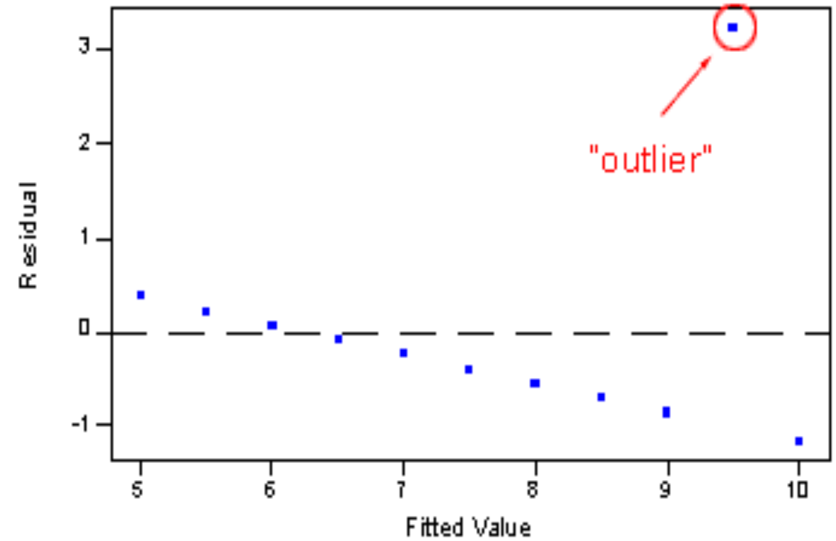



Zoomed in



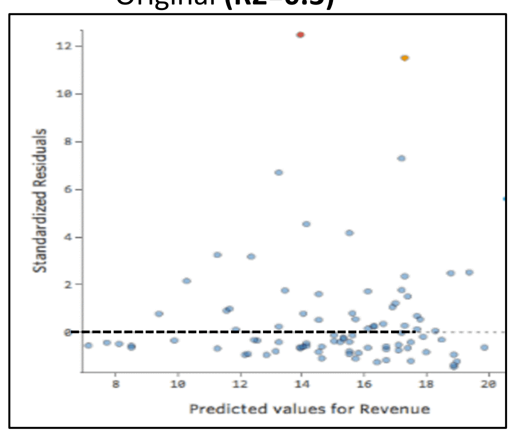
# Residual plot analysis

- The analysis could help us identify outliers, and maybe it is time to remove them to improve the model

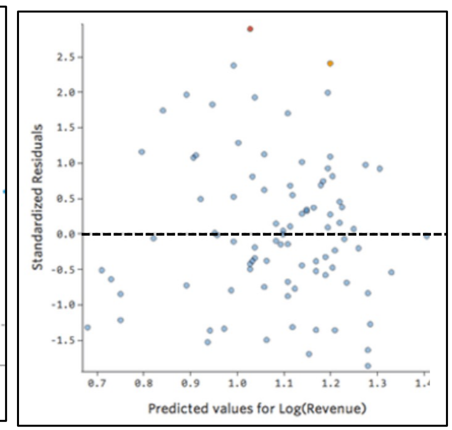



- Also, it could indicate a need to perform transformation to the data to address skewness
  - Several transformations can be tried out such as log-transform, square root, or cube root etc.

Original **(R2=0.3)**




After log transformation**(R2=0.1)**



# Summary

- R- Squared tells how well the developed model explains the variations in the response variable. In other words, whether it captures the trend of the data.

- The error-based performance metrics can tell how well a regression model can predict the value of a response variable. Or, what is the expected cost of our predictions using the developed model.

- The residual plots analysis can help us understand issues within our model and data that are showing if we have non-constant variance, or outliers that might not be reflected in a single number.

# Exercises (6-7)

- How good our regression model is? (Result analysis)
- Build and evaluate regression results using Boston Housing dataset