Module 4 — Model Development

Learning Objectives

In this module you will learn about:

1. Simple and Multiple Regression.
2. Model evaluation using visualization.
3. Polynomial Regression and Pipelines
4. R-squared and MSE for In-Sample Evaluation
5. Prediction and Decision Making

Question

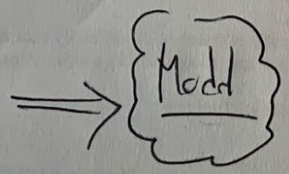
- How can you determine a fair value for a used car ???

Model Development

- A model can be thought of as mathematical equation used to predict a value given one or more other values.
- Relating one or more independent variables to dependent variables.

Independent variables or features

"highway-mpg"
55 mpg



Dependent variables

"predicted price"
$5000

Model Development

- Usually the more relevant data you have the more accurate your model is

"highway-mpg"
"curb-weight"
"engine-size"
$\Rightarrow$ Model $\Rightarrow$ 'price' $\$5400$

To understand why more data is important consider the following situation:

1. You have two almost identical cars
2. Pink cars sell for significantly less

Pink car: "highway-mpg" "curb-weight" "engine-size" $\Rightarrow$ Model $\Rightarrow$ $Y = \$5400$

Red car: "highway-mpg" "curb-weight" "engine-size" $\Rightarrow$ Model $\Rightarrow$ $Y = \$5400$

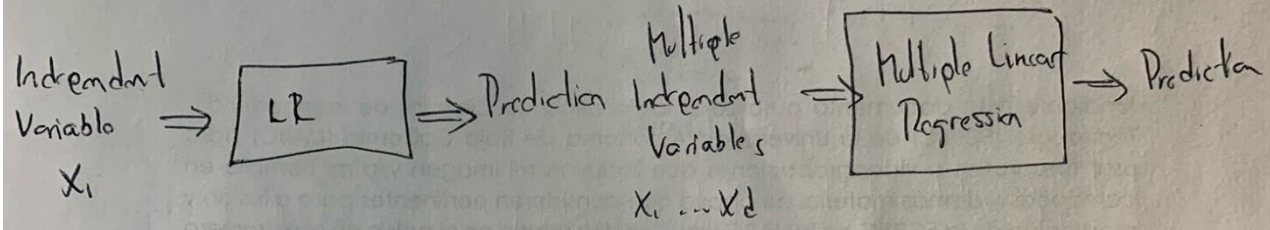
In addition to getting more data you can try different types of models

1. Simple Linear Regression
2. Multiple Linear Regression
3. Polynomial Regression

Simple Linear Regression and Multiple Linear Regression

One independent variable to make a prediction

Multiple independent variable to make a prediction



Simple Linear Regression (SLR)

1.- The predictor (independent) variable - x

2.- The target (dependent) variable - y

$$y = b_0 + b_1 x$$

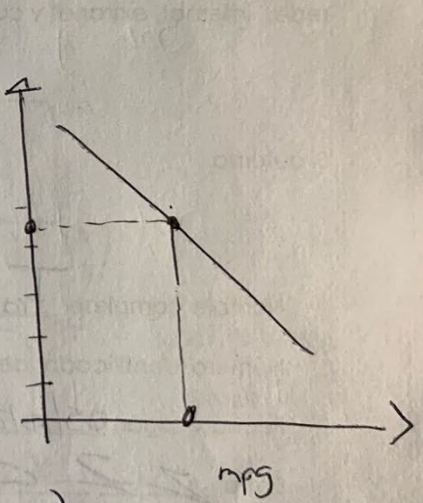
b_0 = the intercept

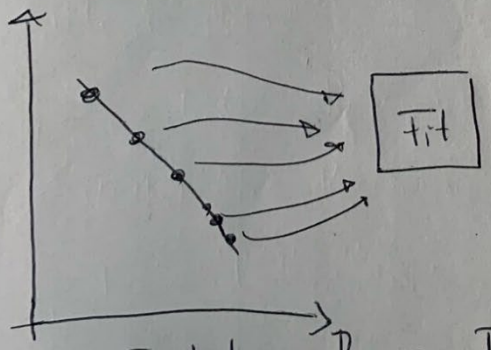
b_1 = the slope

$y = 38423 - 821x$

$= 38423 - 821(20)$

$= 22003$



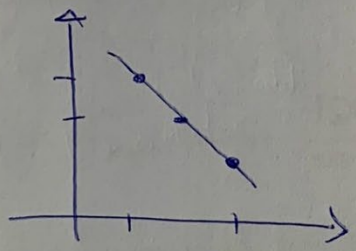


$\Rightarrow (b_0, b_1)$

Result the parameters of the model

Simple Linear Regression: Fit

Simple Linear Regression: Fit



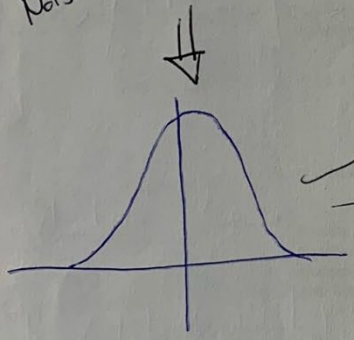
$$X = \begin{bmatrix} 6 \\ 20 \\ 40 \end{bmatrix}$$

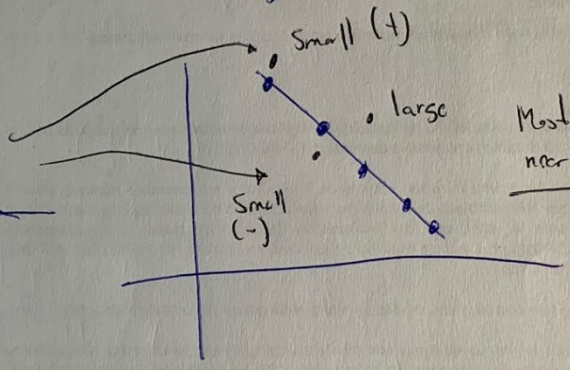
Data frame (Array)

$$Y = \begin{bmatrix} 38423 \\ 22003 \\ 5583 \end{bmatrix}$$

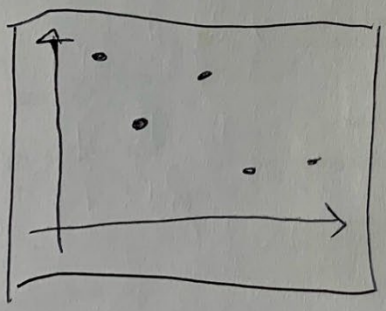
Target (Array)

Noise





Most cases near zero



Fit

(b_0, b_1)

Predict

$\hat{y} = b_0 + b_1 X$

Not always correct

Fitting a simple linear model estimator

X: Predictor Variable

Y: Target Variable

1. Import linear_model from scikit-learn

```
from sklearn.linear_model import LinearRegression
```

2. Create a Linear Regression Object using the constructor:

```
lm = LinearRegression()
```

Fitting a Simple Linear Model

- We define the predictor variable and target variable

```
X = df[['highway-mpg']]
Y = df['price']
```

- Then use lm.fit(X,Y) to fit the model, i.e the parameters b0 and b1

```
lm.fit(X,Y)
```

- We obtain a prediction

```
Yhat = lm.predict(X)
```

SLR - Estimated Linear Model

- We can view the intercept (b0): lm.intercept_ 38423.3058
- We can view the slope (b1): lm.coef_ -821.7333

The Relationship between Price and Highway MPG is given by:

Price = 38423.31 - 821.73 * highway-mpg

$Y = b_0 + b_1 X$

Multiple Linear Regression (MLR)

This method is used to explain the relationship between:

* One continuous target (Y) variable
* Two or more predictor (X) variable

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$$

b_0 = intercept ($X=0$)

b_1 : the coefficient or parameter of X_1

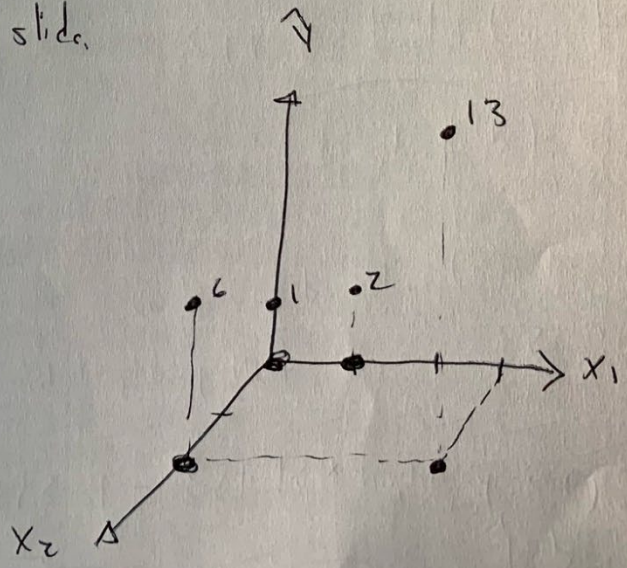
b_2 : the coefficient or parameter of X_2 and so on..

$$\hat{Y} = 1 + 2X_1 + 3X_2$$

The parameters X_1 and X_2 can be visualized on a 2D plane, lets do an example on the next slide.

$$\hat{Y} = 1 + 2X_1 + 3X_2$$

n	X_1	X_2	$\hat{Y}$
1	0	0	1
2	0	2	6
3	1	0	2
4	3	2	13



Fitting a Multiple Linear Model Estimator

1._ We can ~~store~~ extract 4 predictor variables and store them in the variable Z

```
Z = df[['horsepower', 'curb-weight', 'engine-size', 'highway-mpg']]
```

2._ Then train the model as before

```
lm.fit(Z, df['price'])
```

3._ We can also obtain a prediction

```
Yhat = lm.predict(X)
```

MLR – Estimated Linear Model

1._ Find the intercept (b_0)

```
lm.intercept_
```

2._ Find the coefficients (b_1, b_2, b_3, b_4)

```
lm.coef_
```

3._ The Estimated Linear Model:

$$Price = -15678.74 + (52.66) * horsepower + (4.70) * curb\text{-}weight + (81.96) * engine\text{-}size + (33.58) * highway\text{-}mpg$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$$

Model

Model Evaluation Using Visualization

Regression Plot

Why use regression plot?

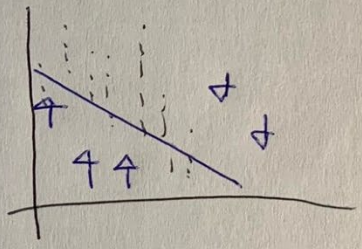
It gives us a good to estimate of:

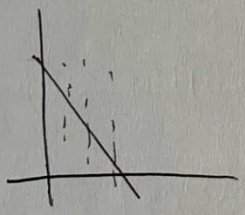
1. The relationship between two variables.
2. The strength of the correlation
3. The direction of the relationship (positive or negative)

Regression Plot Shows us a combination of:

- The scatterplot: where each point represent a different y
- The fitted linear regression line ($\hat{Y}$)

```
import seaborn as sns
sns.regplot(x="highway-mpg", y="price", data=df)
plt.ylim(0,)
```





$\hat{Y} - Y_0 = 84$

$\hat{Y}_1 - \hat{Y} = 50$

$Y_0 = b_0 + b_1 X_0 + e_0$

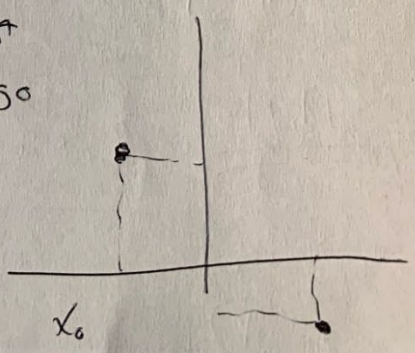
$\hat{Y}_1 = b_0 + b_1 X_1$

$Y_1 = b_0 + b_1 X_1 + e_1$

$\hat{Y} = b_0 + b_1 X$

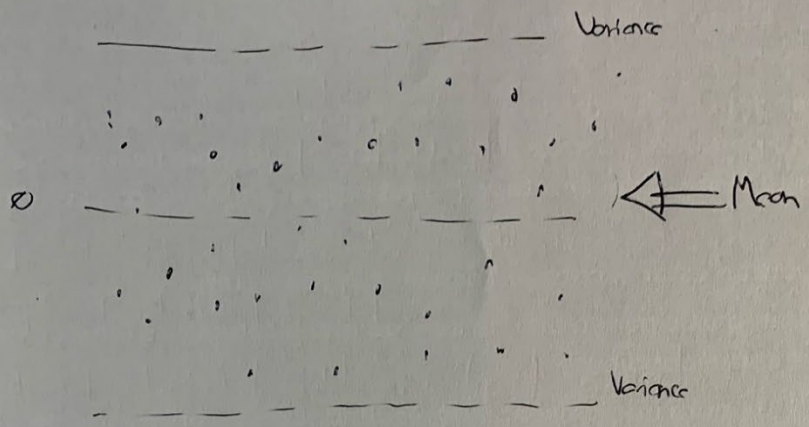
84

X_0 X_1

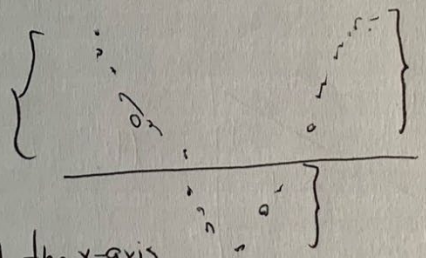


X_0

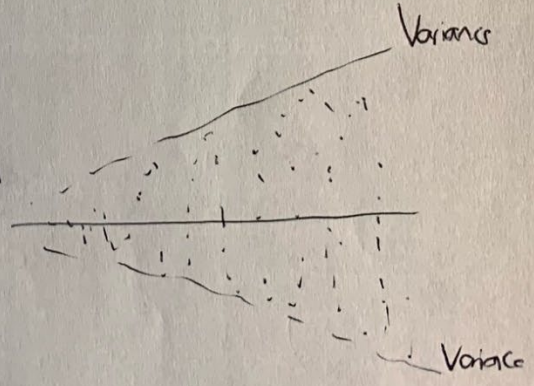
Residual Plot



- Look at the spread of the residuals:
  - → Randomly spread out around x axis then a linear model is appropiate.



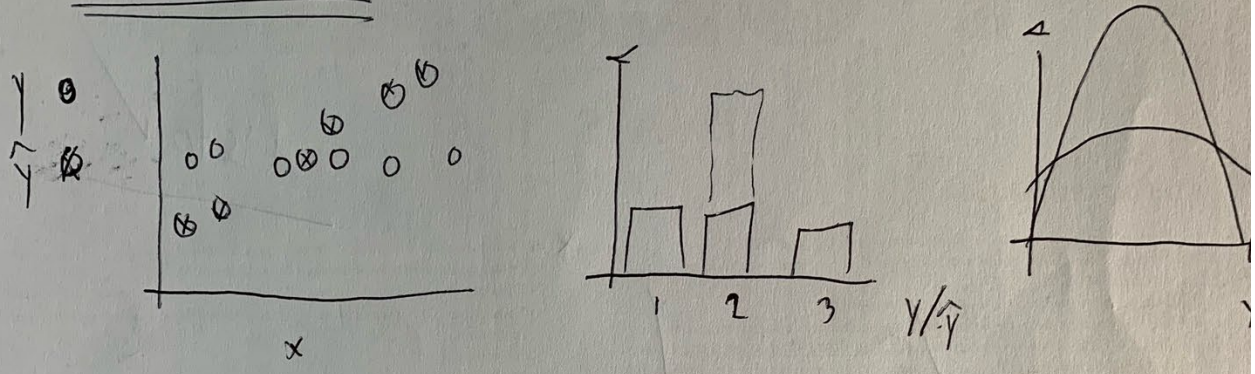
- Not randomly spread out around the x-axis.
- Nonlinear model may be more appropiate.
- Variance appears to change with x axis.



Residual Plot

```
import seaborn as sns
sns.residplot(df['highway-mpg'], df['price'])
```

Distribution Plot



Compare the distribution plots

- The fitted values that result from the model
- The actual values

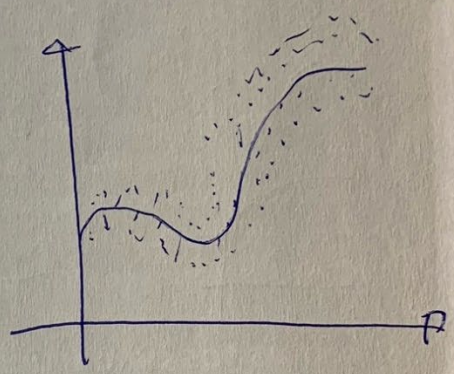
```
import seaborn as sns
ax1 = sns.distplot(df['price'], hist=False, color="r", label="Actual Value")
sns.distplot(Yhat, hist=False, color="b", label="Fitted Values", ax=ax1)
```

Polynomial Regression and Pipelines

- A special case of the general linear regression model
- Useful for describing curvilinear relationships.

Curvilinear relationships:

By squaring or setting higher-order terms of the predictor variables



- Quadratic - 2nd Order

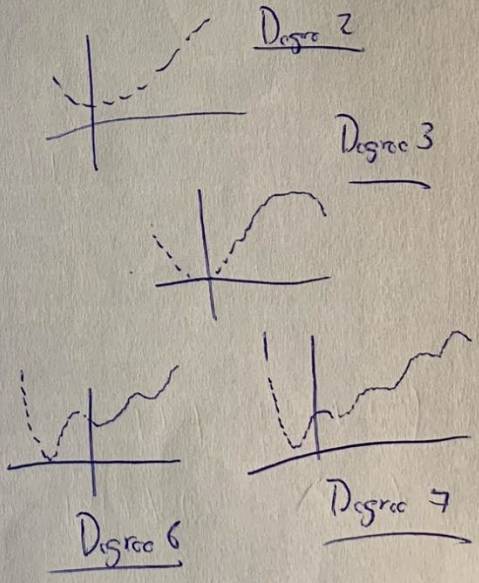
$$\hat{Y} = b_0 + b_1 X_1 + b_2 (X_1)^2$$

- Cubic - 3rd Order

$$\hat{Y} = b_0 + b_1 X_1 + b_2 (X_1)^2 + b_3 (X_1)^3$$

- Higher order

$$\hat{Y} = b_0 + b_1 X_1 + b_2 (X_1)^2 + b_3 (X_1)^3 + \dots$$



1. Calculate Polynomial of 3rd order

```
f = np.polyfit(x,y,3)
p = np.poly1d(f)
```

2. We can print out the model

```
print(p)
```

$$-1.557 (X_1)^3 + 204.8 (X_1)^2 + 8965 X_1 + 1.37 \times 10^5$$

We can also have multi-dimensional polynomial linear regression

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 + b_4 (X_1)^2 + b_5 (X_2)^2 + \dots$$

- The "preprocessing" library in scikit-learn

```
from sklearn.preprocessing import PolynomialFeatures
pr = PolynomialFeatures(degree=2, include_bias=False)
X_poly = pr.fit_transform(X[['horsepower','curb-weight']])
```

```
pr = PolynomialFeatures(degree=2)
pr.fit_transform([1,2], include_bias=False)
```

X_1	X_2
1	2

⇓

X_1	X_2	$X_1 X_2$	X_1^2	X_2^2
1	2	(1)(2)	1	$(2)^2$

1	2	2	1	4

Pre-Processing

- For example we can normalize the each feature simultaneously:

```
from sklearn.preprocessing import StandardScaler
SCALE = StandardScaler()
SCALE.fit(x_data[['horsepower', 'highway-mpg']])
x_scale = SCALE.transform(x_data[['horsepower', 'highway-mpg']])
```

Pipelines

- There are many steps to getting a prediction

Normalization ⇒ Polinomial Transform ⇒ Linear Regression

Normalization + Polinomial Transform: Transformations

Linear Regression: Predictions

Pipelines

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
```

Pipeline Constructor

```
Input = [('scale', StandardScaler()), ('polynomial', PolynomialFeatures(degree=2)),
        ('model', LinearRegression())]
```

- Pipeline Constructor

```
pipe = Pipeline(Input)
```

```
Pipe.train(X['horsepower', 'curb-weight', 'engine-size', 'highway-mpg'], y)
yhat = Pipe.predict(X[['horsepower', 'curb-weight', 'engine-size', 'highway-mpg']])
```

X ⇒ Normalization ⇒ Polinomial Transform ⇒ Linear Regression ⇒ ŷ

Measures for In-Sample Evaluation

- A way to numerically determine how good the model fits on dataset.

- Two important measures to determine the fit of a model:
  - → Mean Squared Error (MSE)
  - → R-Squared (R^2)

MSE

- In python we can measure the MSE as follows

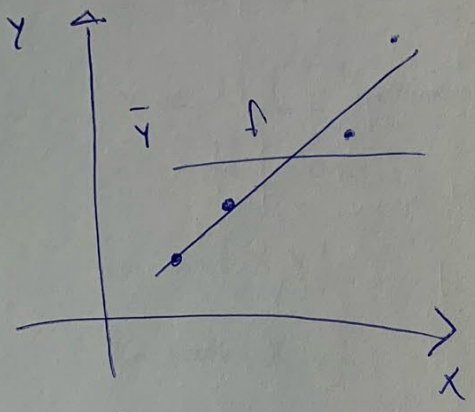
```
from sklearn.metrics import mean_squared_error
mean_squared_error(df['price'], Y_predict_simple_fit)
```

3163502.9446

R-Squared / R^2

- → The coefficient of determination or R squared (R^2)
- → Is a measured to determine how close the data is to the fitted regression line.
- → R^2: the percentage of variation of the target variable (Y) that is explained by the linear model.
- → Think about as comparing a regression model to a simple model i.e the mean of the data points

$$R^2 = \left(1 - \frac{\text{MSE of Regression Line}}{\text{MSE of Average of the data}}\right)$$



- The blue line represents the regression line
- The blue squares represents the MSE of the regression line
- The red line represents the average value of the data points
- The red squares represents the MSE of the red line.
- We see the area of the blue squares is much smaller than the area of the red squares.

, In this case of the areas of MSE is close to zero

$$\frac{\text{MSE of Regression line}}{\text{MSE of } \bar{Y}} = \frac{\square + \square}{\square + \square}$$

$$= 0 \qquad R^2 = \left(1 - \frac{\text{MSE of Regression line}}{\text{MSE of } \bar{Y}}\right)$$

$$= (1 - 0)$$

$$= 1$$

R-squared / R^2

- Generally the values of the MSE are between 0 and 1.
- We can calculate the R^2 as follows

```
X = df[['highway-mpg']]
Y = df['price']
lm.fit(X,Y)
lm.score(X,Y)
0.49659188
```

overfitting
-1

Prediction and Decision Making

To determine the final best fit, we look a combination of

- Do the predicted values make sense.
- Visualization
- Numerical measures for evaluation
- Comparing models

Do the predicted values make sense

- First we train the model

```
lm.fit(df['highway-mpg'], df['prices'])
```

Lets predict the price of a car with 30-mpg

```
lm.predict(30)
```

- Result: $ 13771.36

```
lm.coef
-821.77337
```

Price = 38423.31 - 821.73 highway

- First we import numpy

  import numpy as np

- We use the numpy function arrange to generate a sequence from 1 to 100

new ~~numpy~~

```
new_input=np.arange(1,101,1).reshape(-1,1)
```

1	2	...	99	100

- We can predict new values

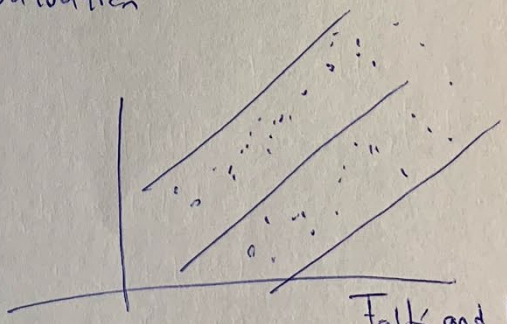
```
yhat=lm.predict(new_input)
```

Visualization

- Simply visualizing your data with regression
- Residual Plot
- Distribution Plot

Numerical Measurements for Evaluation

- Mean Square Error
- R-Squared



Falk and Miller 1992

Equal to or Greater than 0.10

Comparing MLR and SLR

1. Is a lower MSE always implying a better fit?

   No, Not necessarily.

2. MSE for MLR model will be smaller than the MSE for an SLR model, since the errors of the data will decrease when more variables are included in the model.

3. Polynomial Regression will also have a smaller MSE than regular regression.

4. A similar inverse relationship holds for R^2