Linear Regression

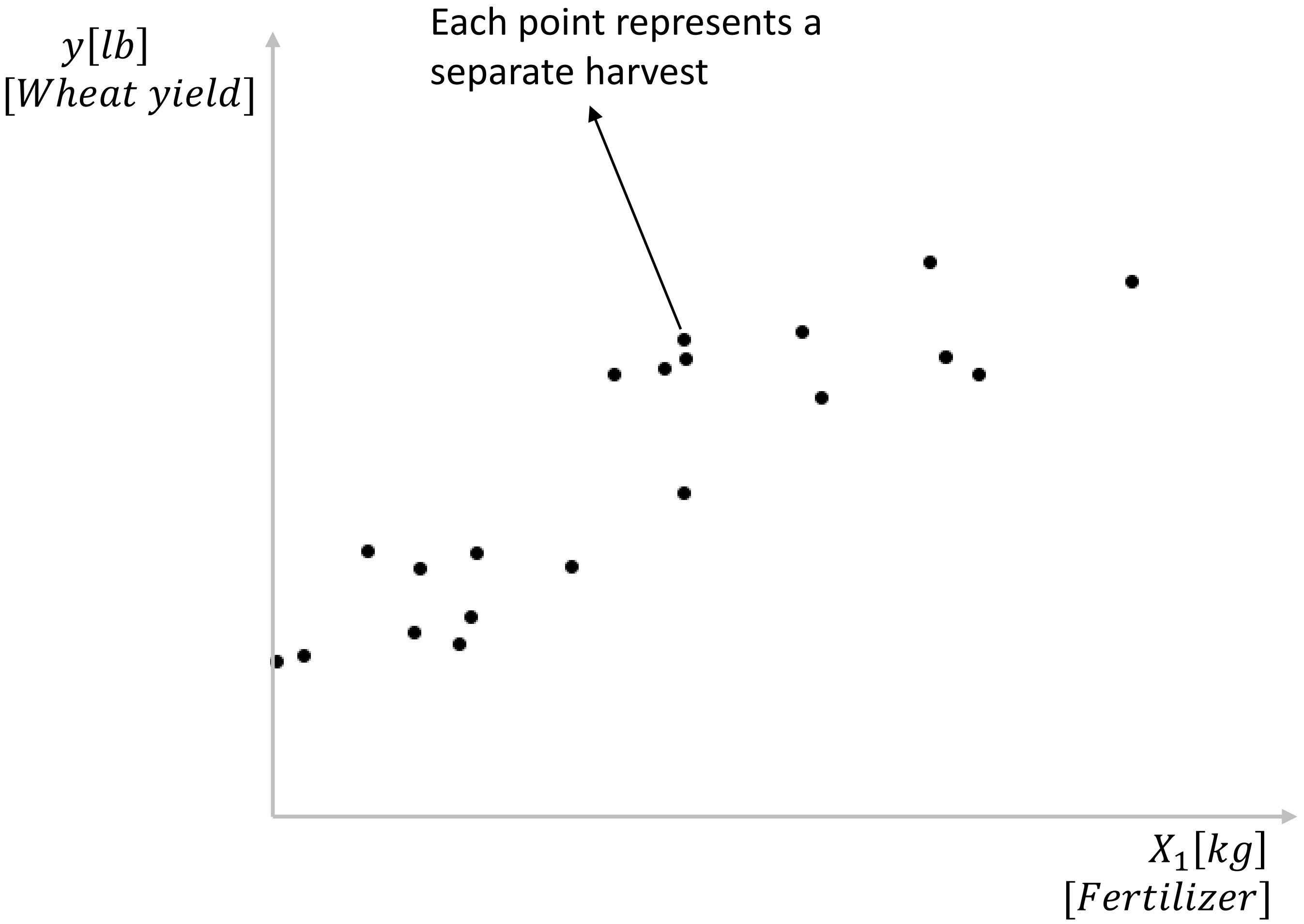
Supervised Learning

Simple Linear Regression

 $\sim$ 

Simple Linear Regression

$\sim$

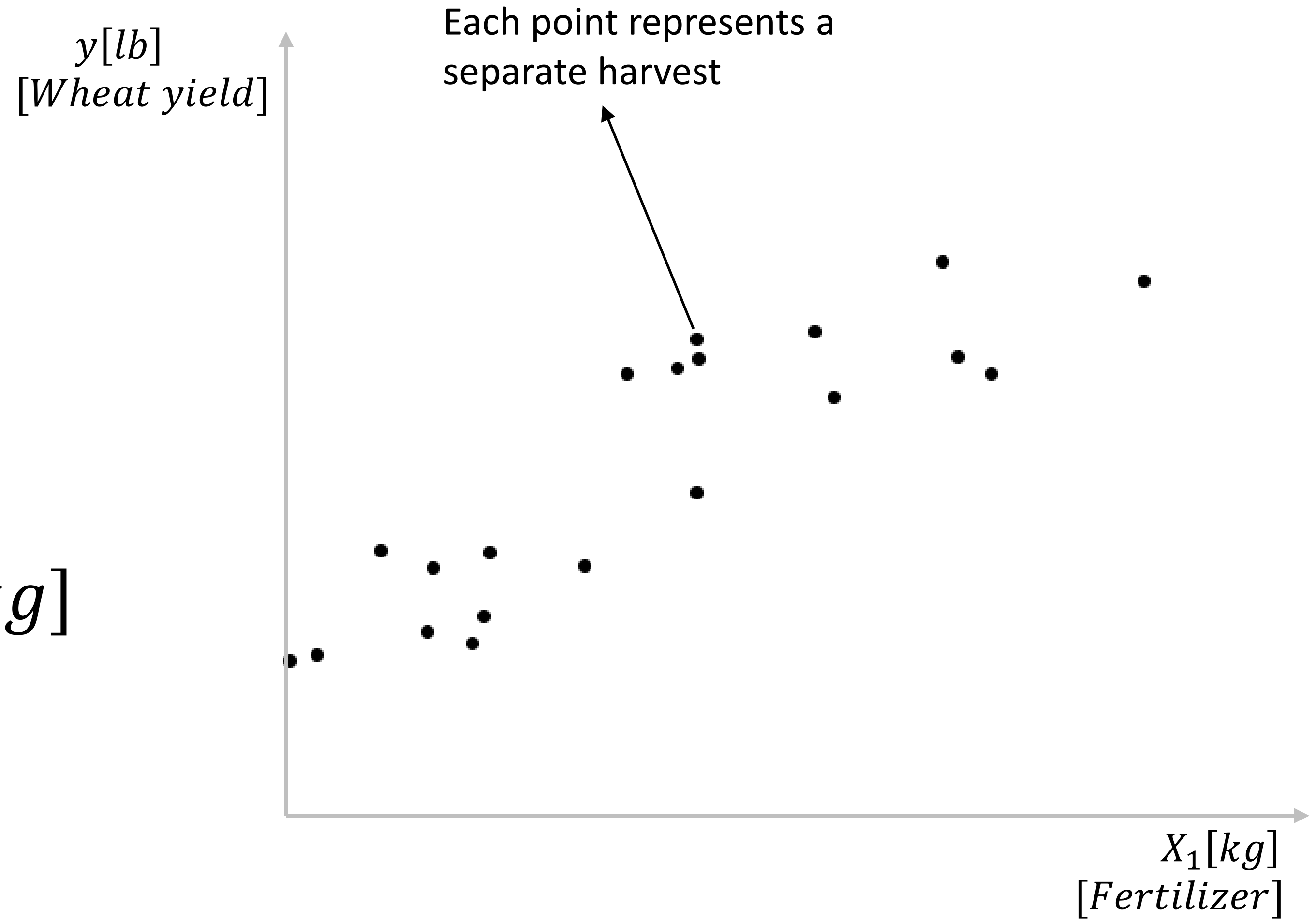


Simple Linear Regression

$$\hat{y} = b_0 + b_1 X_1$$

$\sim$

$$wheat[lb] = b_0 + b_1 * Fertilizer[kg]$$



Simple Linear Regression

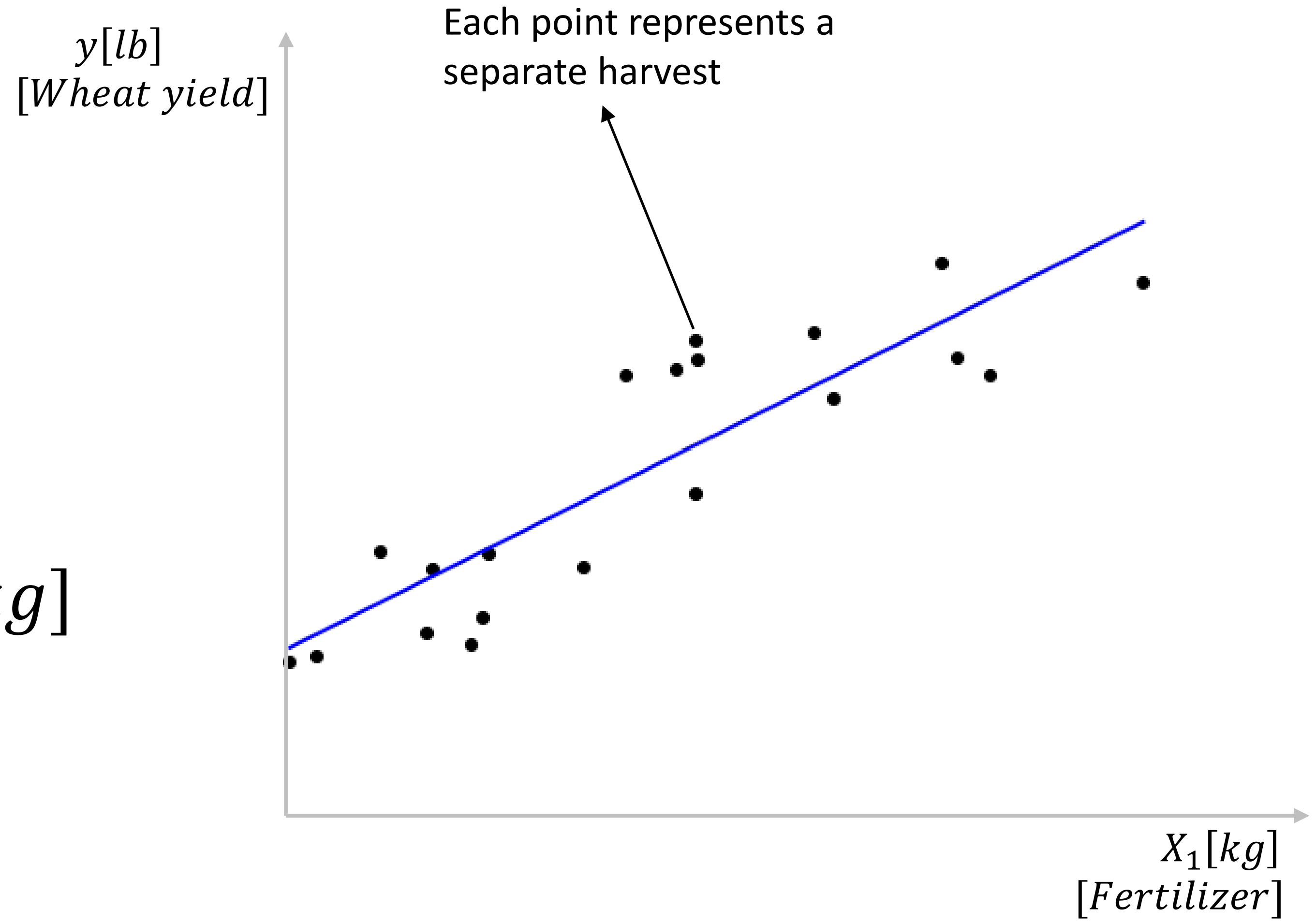
$$\hat{y} = b_0 + b_1 X_1$$

 $\sim$ 

$$wheat[lb] = b_0 + b_1 * Fertilizer[kg]$$

$$b_0 = 10[lb]$$

$$b_1 = 4[\frac{lb}{kg}]$$



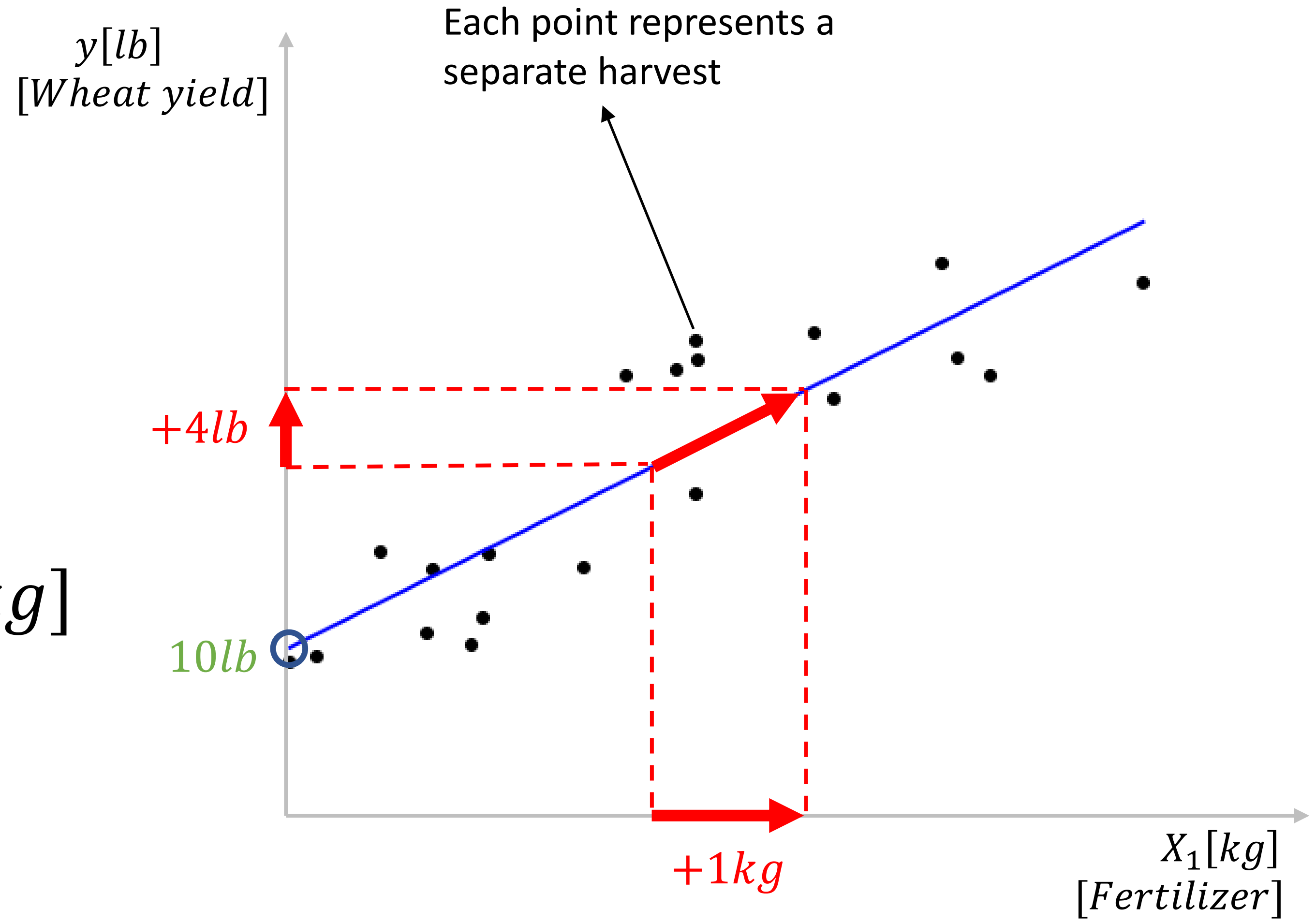
Simple Linear Regression

$$\hat{y} = b_0 + b_1 X_1$$

$$wheat[lb] = b_0 + b_1 * Fertilizer[kg]$$

$$b_0 = 10[lb]$$

$$b_1 = 4[\frac{lb}{kg}]$$



Simple Linear Regression

$$\hat{y} = b_0 + b_1 X_1$$

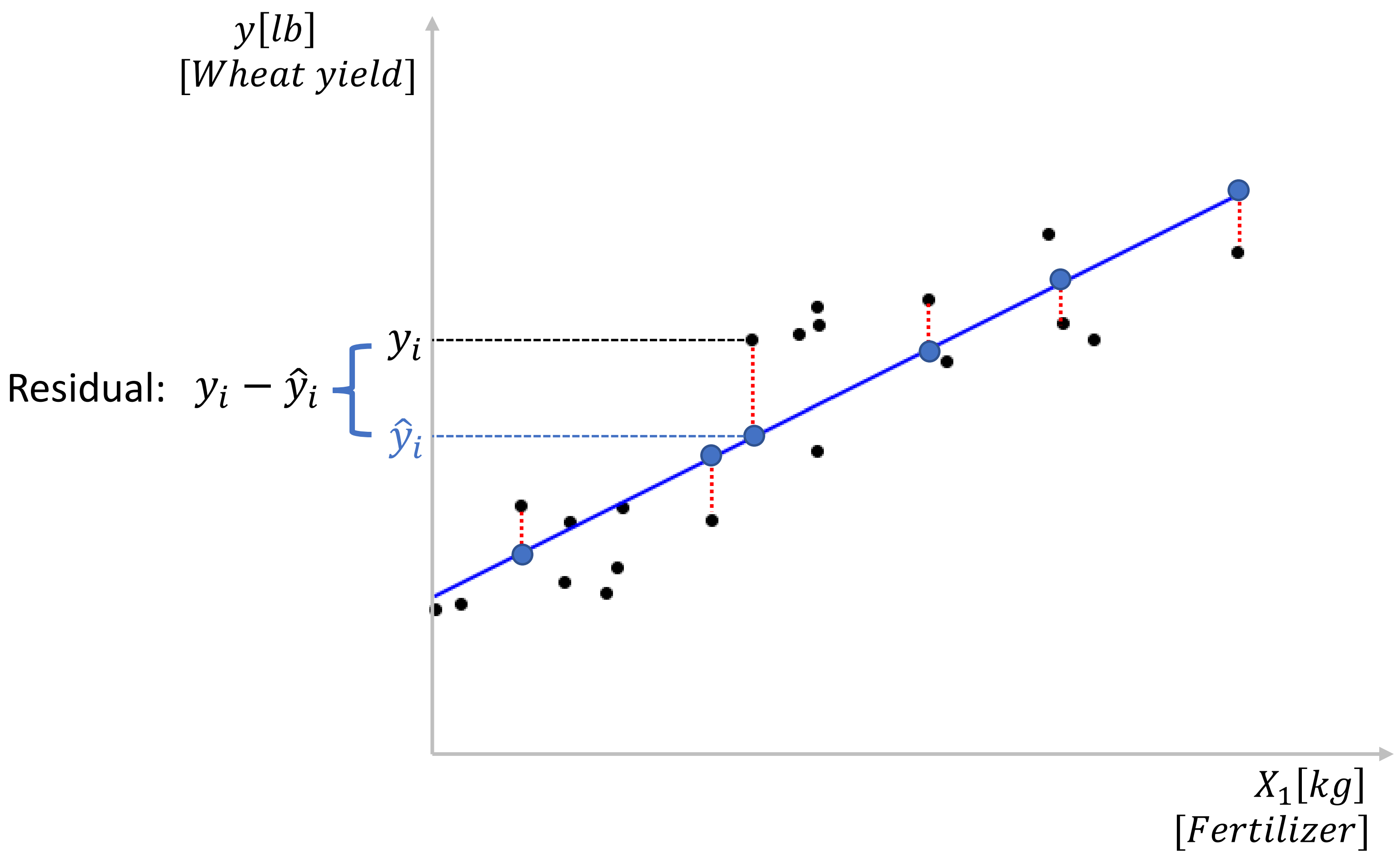
Ordinary Least Squares:

Estimates the values of

b_0, b_1 such that:

$$\sum_i (y_i - \hat{y}_i)^2$$

is minimized.



Simple Linear Regression

$$\underbrace{\hat{y}}_{\text{Dependent variable / Response / Target}} = \underbrace{b_0}_{\text{y-intercept}} + \underbrace{b_1}_{\text{Slope coefficient}}\underbrace{X_1}_{\text{Independent variable / Explanatory variable / Predictor / Feature}}$$

Multiple Linear Regression

$$\hat{y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

Dependent variables
Responses
Targets

y-intercept

Coefficients

Independent variables
Explanatory variables
Predictors
Features

Multiple Linear Regression

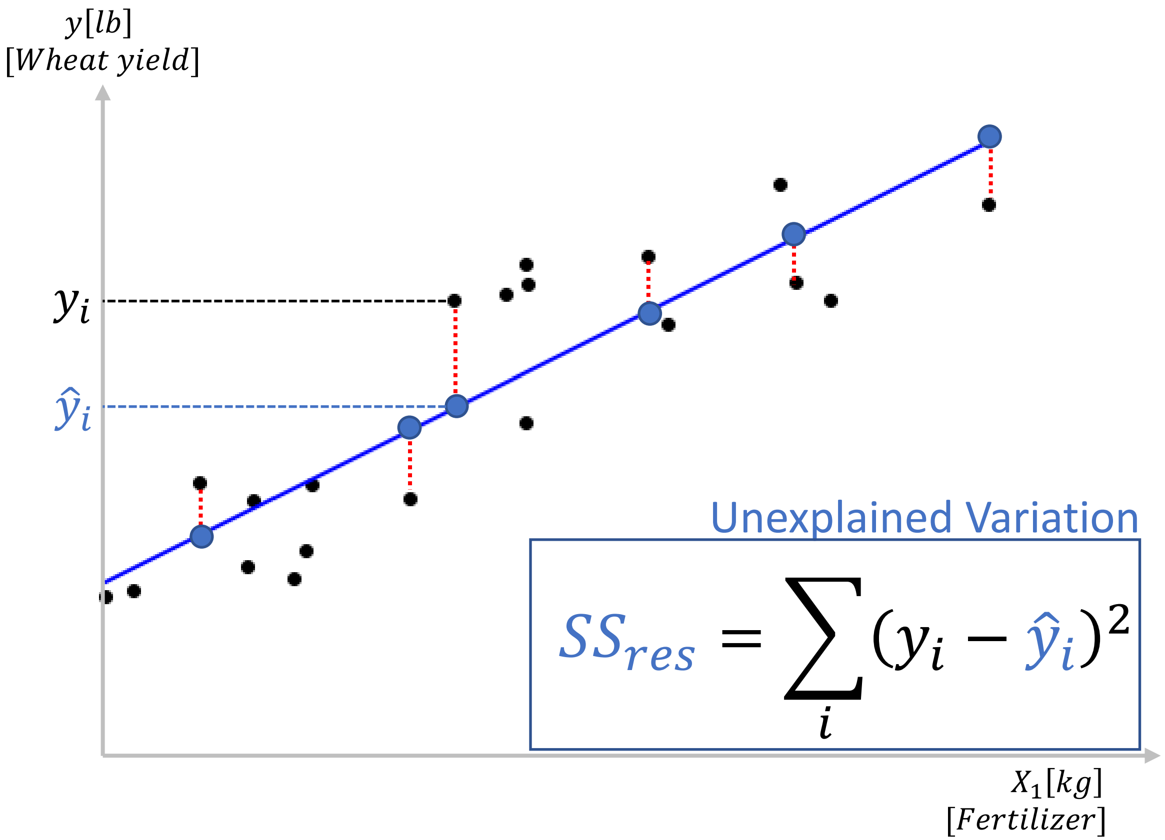
$$\hat{y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

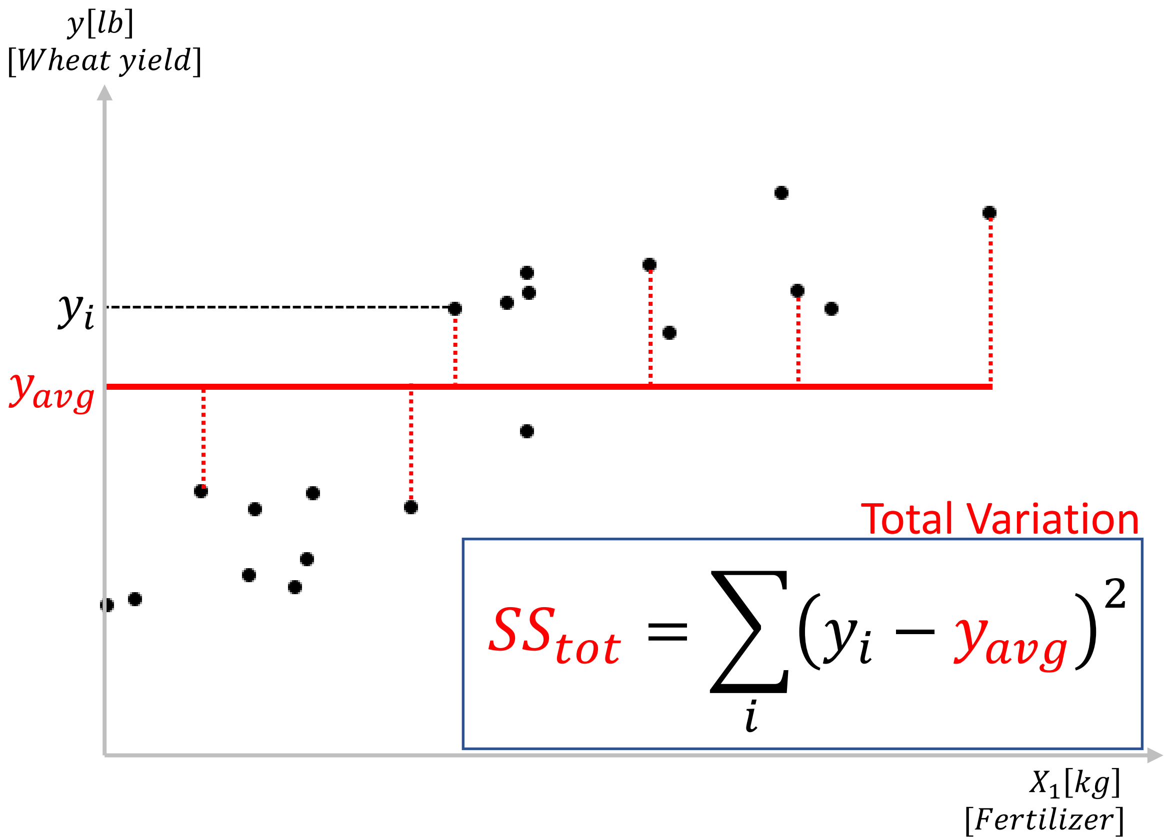
 $\sim$ 

$$wheat[lb] = 7lb + 3\frac{lb}{kg} * Fertilizer[kg] - 0.5\frac{lb}{F} * Temp[F] + 1.2\frac{lb}{mm} * Rain[mm]$$

Coefficient of Determination (R^2)





$$R^2 = 1 - \frac{Unexplained\ Variation}{Total\ Variation} = 1 - \frac{SS_{res}}{SS_{tot}}$$

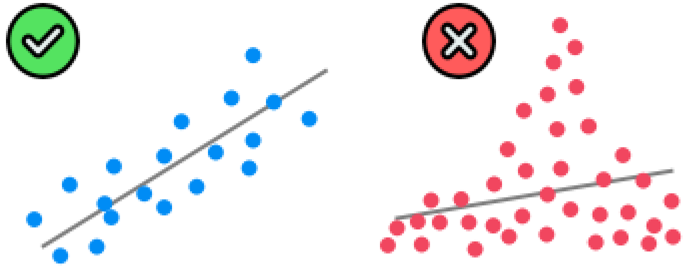
- Measures the % of variance in the target variable explained by the model
- Values between [0, 1]. Higher the better in general.

Evaluation metrics

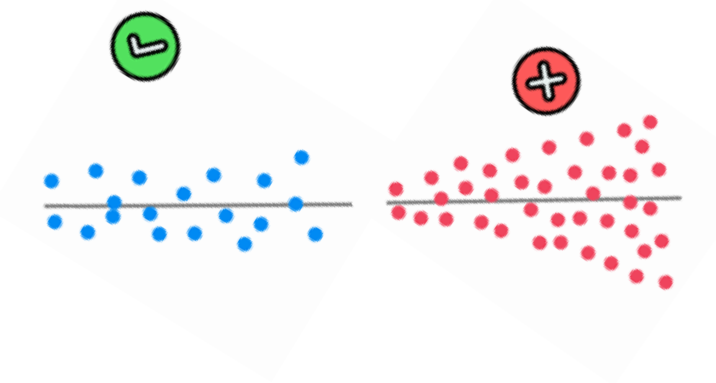
R^2	Adjusted R^2	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$	$Adj.R^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
• Measures the % of variance in the target variable explained by the model • Values between [0, 1] • Higher the better in general	• Similar to R^2, but penalizes for the addition of too many variables • Adding more variables always increase R^2, but not Adj. R^2 • Higher the better in general	• Easy to understand and interpret • Same unit as target variable • Not sensitive to outliers • Lower the better	• Same unit as target variable • Sensitive to outliers – errors will be magnified due to the squared term • Lower the better

Assumptions of Linear Regression

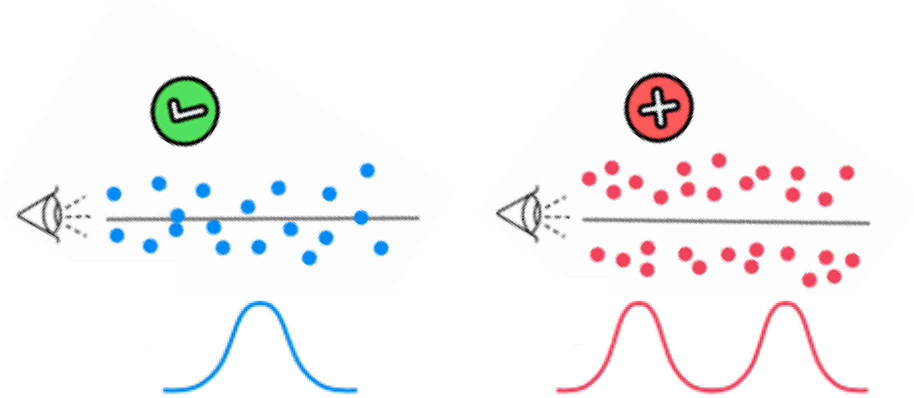
1. Linearity
(Linear relationship between Y and each X)



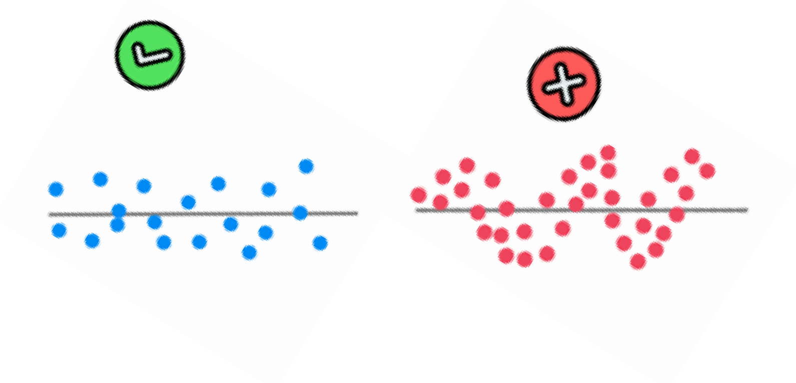
2. Homoscedasticity
(Equal variance)



3. Multivariate Normality
(Normality of error distribution)



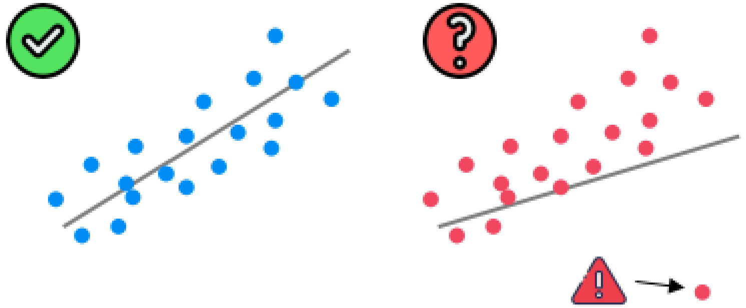
4. Independence
(of observations. Includes “no autocorrelation”)



5. Lack of Multicollinearity
(Predictors are not correlated with each other)

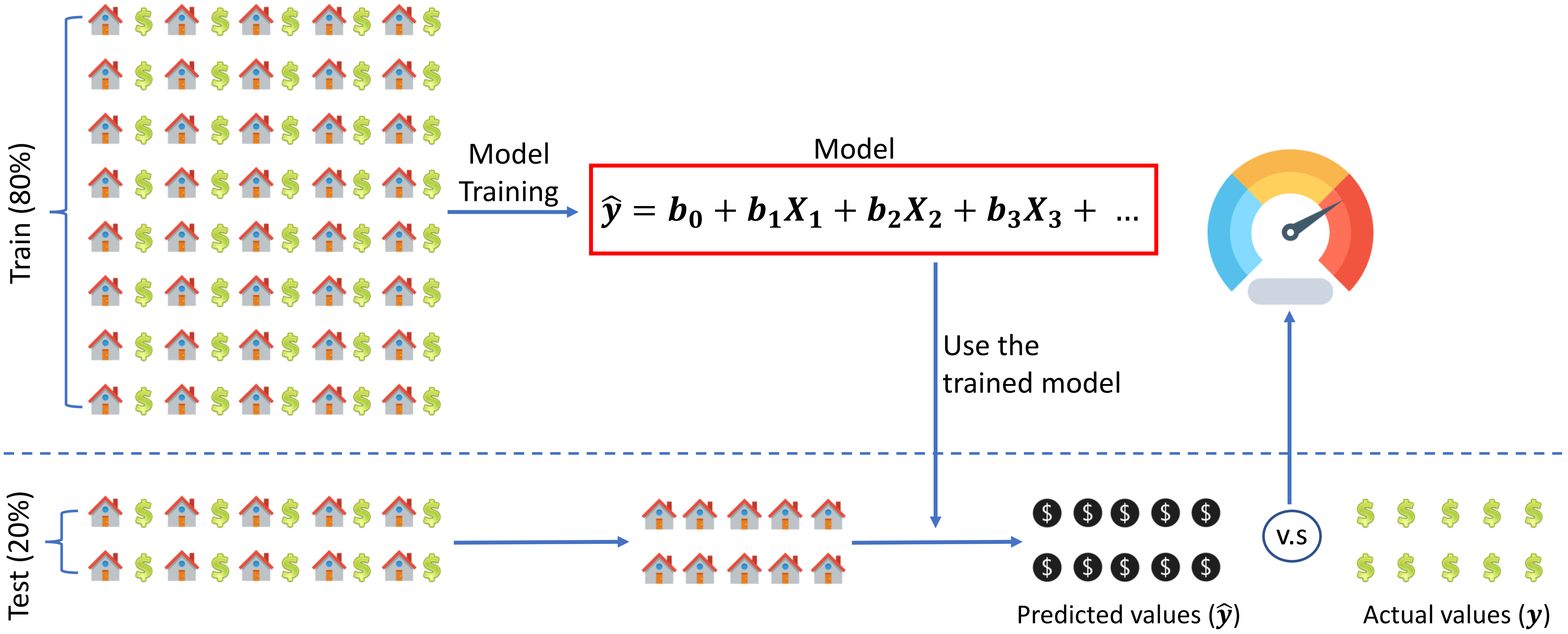
$X_1 \nsim X_2$ $X_1 \sim X_2$

6. The Outlier Check
(This is not an assumption, but an “extra”)



Source: https://www.superdatascience.com/blogs/assumptions-of-linear-regression

Linear Regression in action



Pros and cons of Linear Regression

Pros

- Easy to implement and understand.
- Computationally efficient.

Cons

- Too simple to capture many of real-world complexities.
- Sensitive to outliers.
- Too many statistical assumptions which are usually hard to satisfy in real world data.

Categorical Encodings

Categorical Encodings

- Real world data is usually a combination of numerical columns (features) and categorical (strings) columns.
- Most implementations of machine learning algorithms require the input data to be numeric.
- Need to convert categorical columns to some numeric counterpart so that the machine learning algorithms can process them.
- Categorical Encoding - Process of transforming a categorical column into one or more numeric column(s).
- E.g. OneHotEncoding, OrdinalEncoding, CountEncoding, HashingEncoding, TargetEncoding, etc.

OneHotEncoding (Dummy variables)

y Life expectancy	X_1 Adult Mortality	X_2 Total expenditure	X_3 Schooling	Country
76.8	114	13.7	12.3	USA
76.9	115	13.73	12.3	USA
77	115	14.55	12.3	USA
71	179	7.19	14.3	Brazil
71.4	176	7.13	14.6	Brazil
71.8	172	6.94	14.8	Brazil
68.6	16	4.18	8	Morocco
69	155	4.44	8.5	Morocco
69.5	15	5.31	8.8	Morocco

$$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + ?$$

OneHotEncoding (Dummy variables)

Categorical variable

y Life expectancy	X_1 Adult Mortality	X_2 Total expenditure	X_3 Schooling	Country
76.8	114	13.7	12.3	USA
76.9	115	13.73	12.3	USA
77	115	14.55	12.3	USA
71	179	7.19	14.3	Brazil
71.4	176	7.13	14.6	Brazil
71.8	172	6.94	14.8	Brazil
68.6	16	4.18	8	Morocco
69	155	4.44	8.5	Morocco
69.5	15	5.31	8.8	Morocco

How should we handle this?

$$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + ?$$

OneHotEncoding (Dummy variables)

Categorical variable

Dummy variables

y	X_1	X_2	X_3		D_1	D_2	D_3
Life expectancy	Adult Mortality	Total expenditure	Schooling	Country	USA	Brazil	Morocco
76.8	114	13.7	12.3	USA	1	0	0
76.9	115	13.73	12.3	USA	1	0	0
77	115	14.55	12.3	USA	1	0	0
71	179	7.19	14.3	Brazil	0	1	0
71.4	176	7.13	14.6	Brazil	0	1	0
71.8	172	6.94	14.8	Brazil	0	1	0
68.6	16	4.18	8	Morocco	0	0	1
69	155	4.44	8.5	Morocco	0	0	1
69.5	15	5.31	8.8	Morocco	0	0	1

$$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 D_1 + b_5 D_2 + b_6 D_3$$

OneHotEncoding (Dummy variables)

Dummy variables

Categorical variable

y Life expectancy	X_1 Adult Mortality	X_2 Total expenditure	X_3 Schooling	Country	D_1 USA	D_2 Brazil	D_3 Morocco
76.8	114	13.7	12.3	USA	1	0	0
76.9	115	13.73	12.3	USA	1	0	0
77	115	14.55	12.3	USA	1	0	0
71	179	7.19	14.3	Brazil	0	1	0
71.4	176	7.13	14.6	Brazil	0	1	0
71.8	172	6.94	14.8	Brazil	0	1	0
68.6	16	4.18	8	Morocco	0	0	1
69	155	4.44	8.5	Morocco	0	0	1
69.5	15	5.31	8.8	Morocco	0	0	1

Dummy variable Trap!!!

$$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 D_1 + b_5 D_2 + b_6 D_3$$

but, $D_3 = 1 - D_1 - D_2$

OneHotEncoding (Dummy variables)

y	X_1	X_2	X_3	Categorical variable	Dummy variables D_1	D_2	D_3
Life expectancy	Adult Mortality	Total expenditure	Schooling	Country	USA	Brazil	~~Morocco~~
76.8	114	13.7	12.3	USA	1	0	~~0~~
76.9	115	13.73	12.3	USA	1	0	~~0~~
77	115	14.55	12.3	USA	1	0	~~0~~
71	179	7.19	14.3	Brazil	0	1	~~0~~
71.4	176	7.13	14.6	Brazil	0	1	~~0~~
71.8	172	6.94	14.8	Brazil	0	1	~~0~~
68.6	16	4.18	8	Morocco	0	0	~~1~~
69	155	4.44	8.5	Morocco	0	0	~~1~~
69.5	15	5.31	8.8	Morocco	0	0	~~1~~

$$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 D_1 + b_5 D_2 + \cancel{b_6 D_3}$$

- Always need to omit one dummy variable (doesn't matter which one is omitted)
- If there are n levels in a categorical variable, we only need $n - 1$ dummy variables

OneHotEncoding (Dummy variables)

Original data

Life expectancy	Adult Mortality	Total expenditure	Schooling	Country
76.8	114	13.7	12.3	USA
76.9	115	13.73	12.3	USA
77	115	14.55	12.3	USA
71	179	7.19	14.3	Brazil
71.4	176	7.13	14.6	Brazil
71.8	172	6.94	14.8	Brazil
68.6	16	4.18	8	Morocco
69	155	4.44	8.5	Morocco
69.5	15	5.31	8.8	Morocco

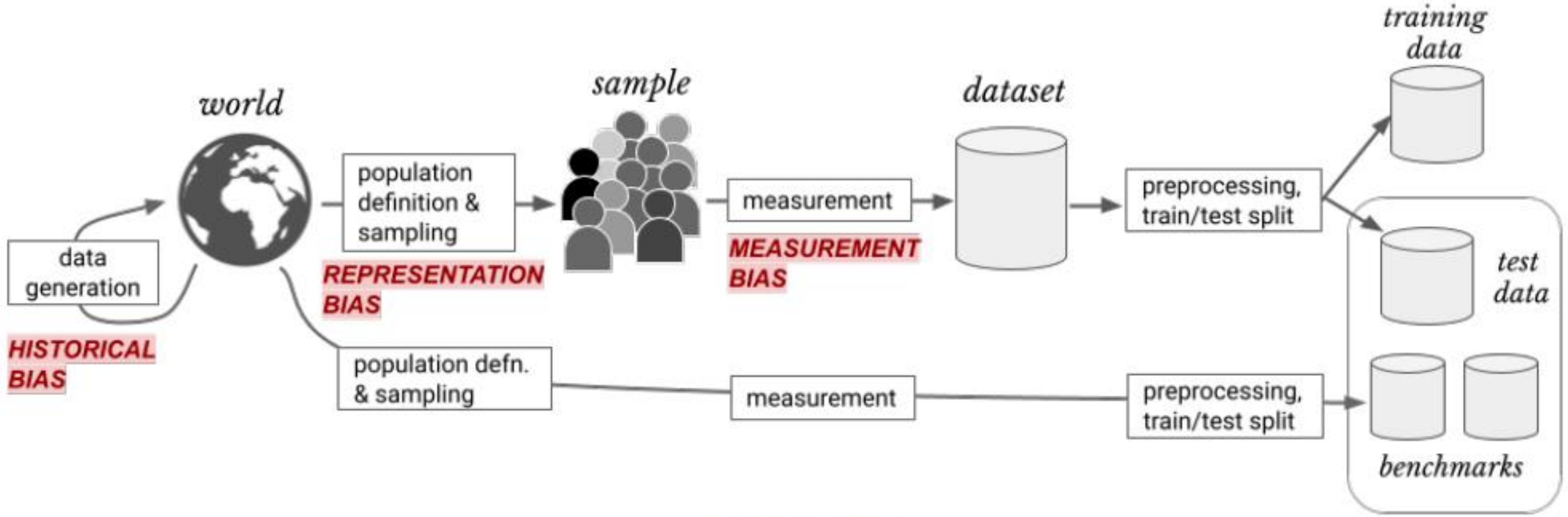
Modified data (Dummified)

Life expectancy	Adult Mortality	Total expenditure	Schooling	USA	Brazil
76.8	114	13.7	12.3	1	0
76.9	115	13.73	12.3	1	0
77	115	14.55	12.3	1	0
71	179	7.19	14.3	0	1
71.4	176	7.13	14.6	0	1
71.8	172	6.94	14.8	0	1
68.6	16	4.18	8	0	0
69	155	4.44	8.5	0	0
69.5	15	5.31	8.8	0	0

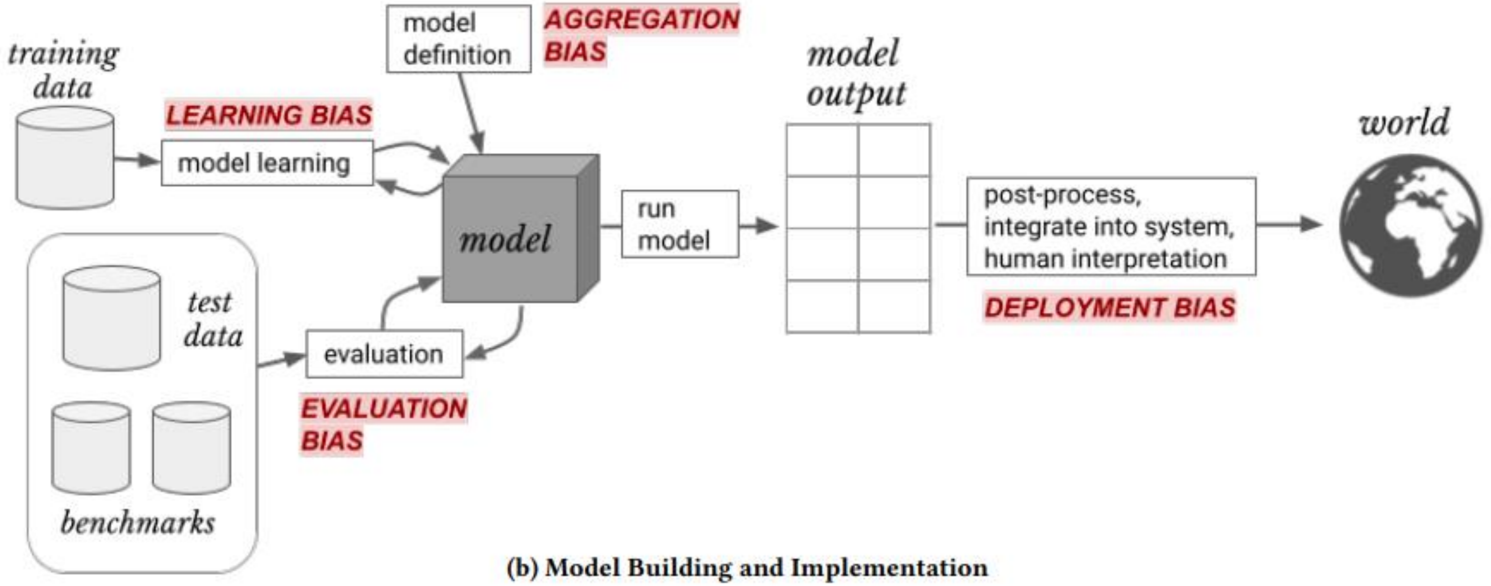
$$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 D_1 + b_5 D_2$$

Bias and Variance

Bias-Variance



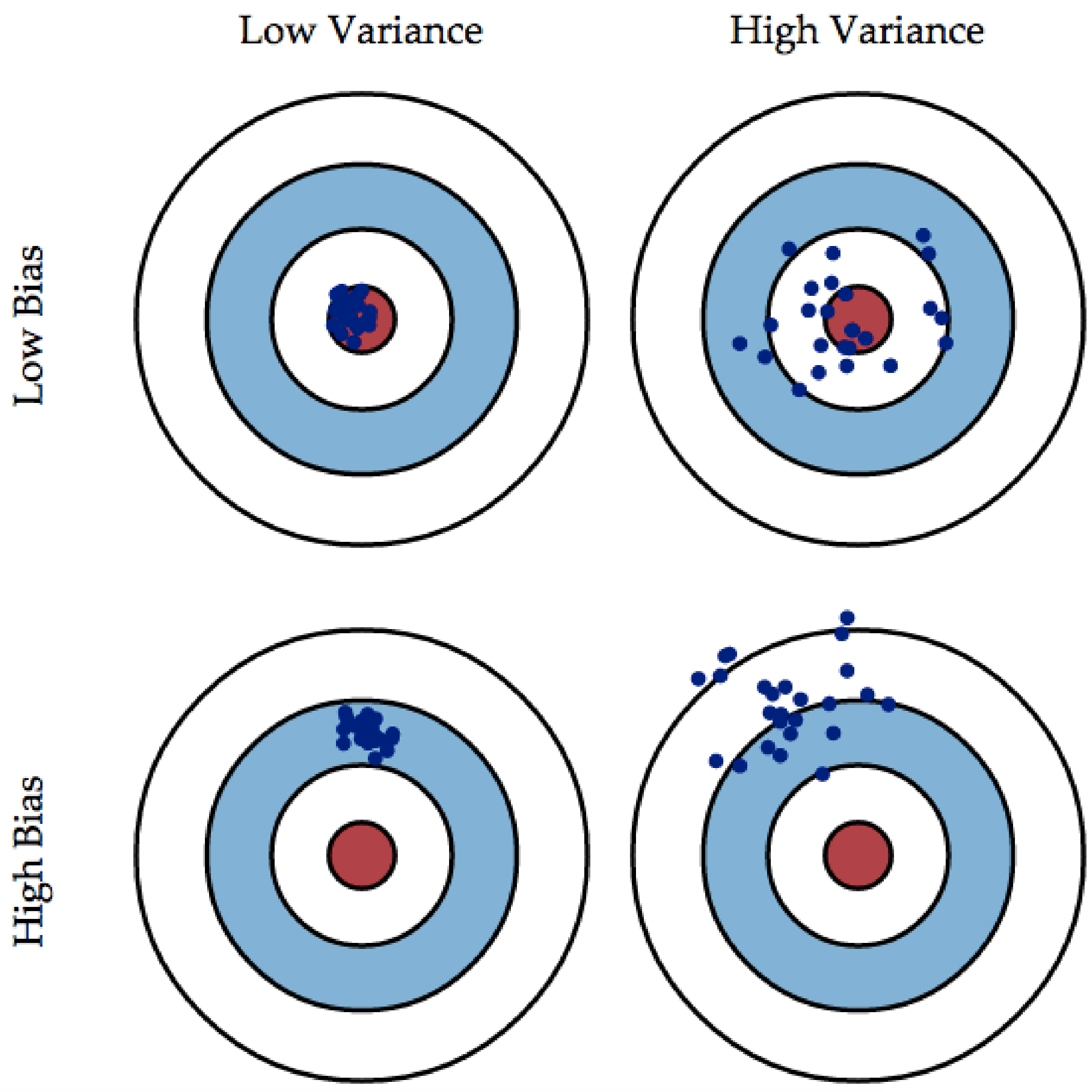
(a) Data Generation



(b) Model Building and Implementation

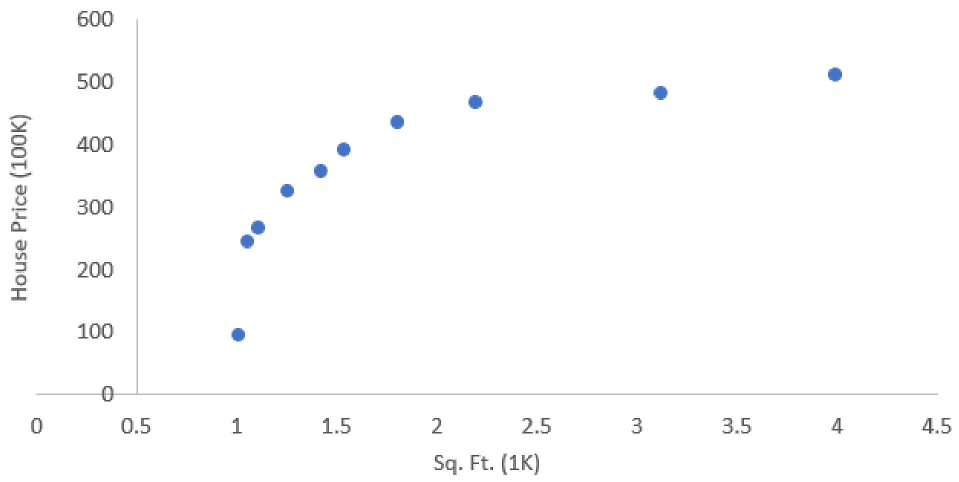
https://arxiv.org/pdf/1901.10002.pdf

Bias-Variance

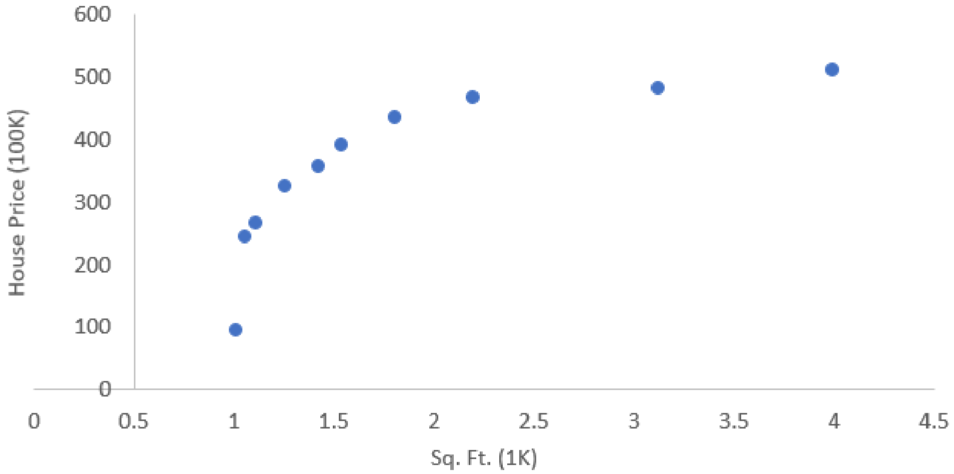


- The ***error due to bias*** is the amount by which the expected model prediction differs from the true value or target, over the training data.
- If these average prediction values are substantially different that the true value, bias will be high.
- The ***error due to variance*** is the amount by which the prediction, over one training set, differs from the expected predicted value, over all the training sets.
- Variance measures how inconsistent are the predictions from one another, over different training sets, not whether they are accurate or not.

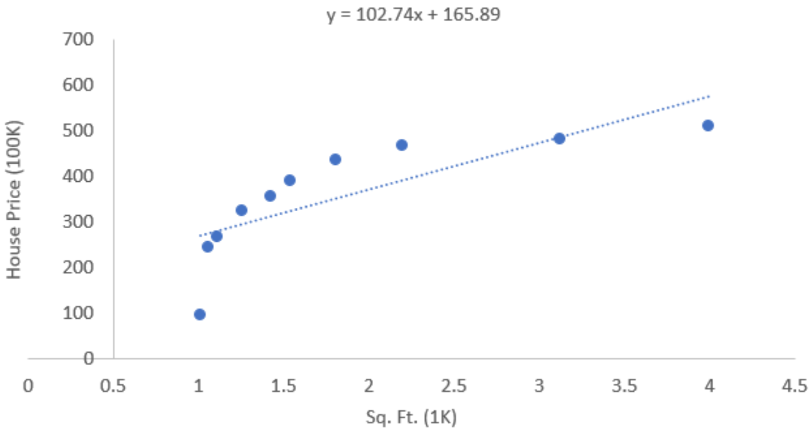
Bias-Variance Tradeoff



Bias-Variance Tradeoff

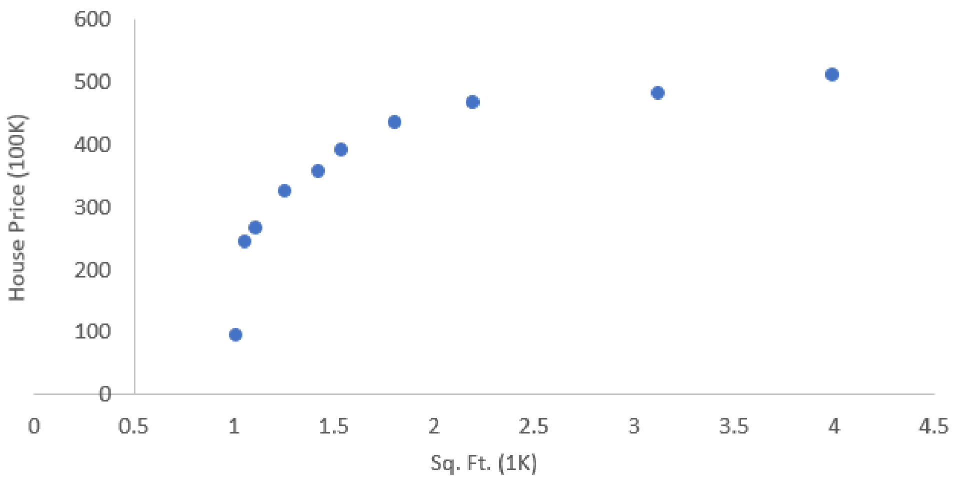


Model 1

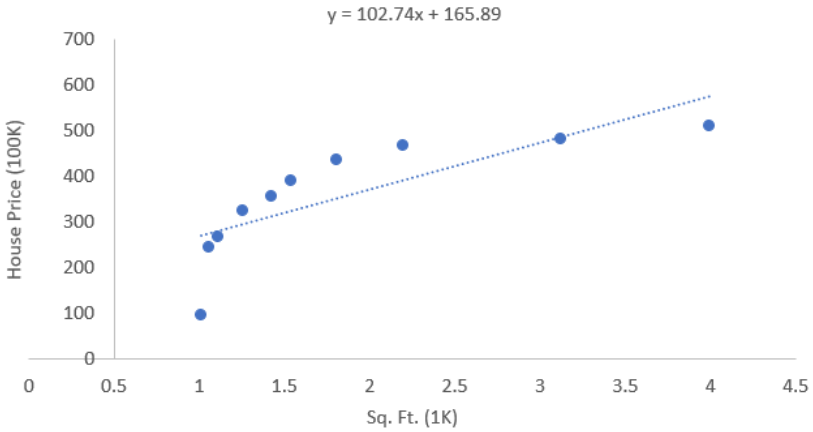


$$y = b_0 + b_1 X_1$$

Bias-Variance Tradeoff

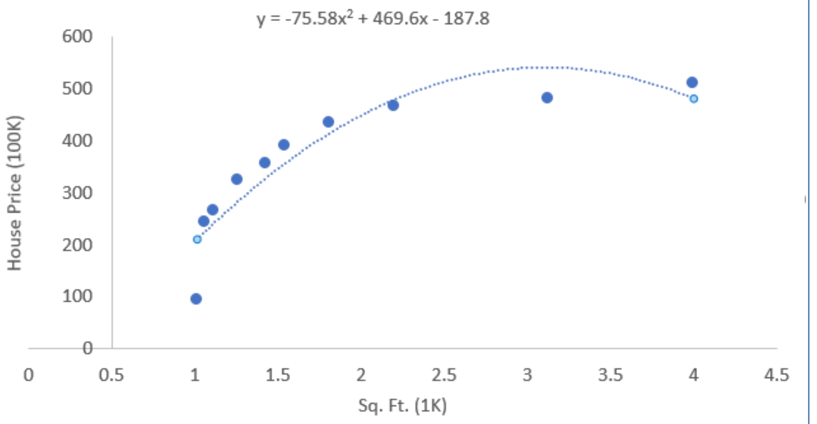


Model 1



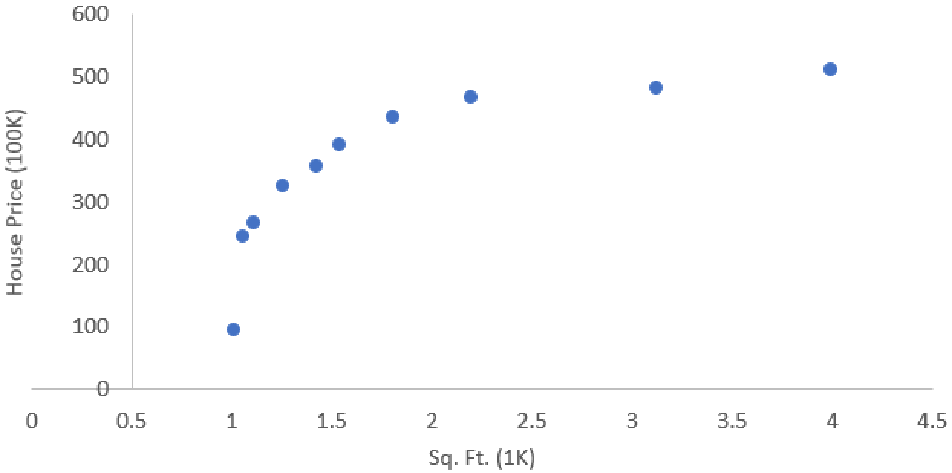
$$y = b_0 + b_1 X_1$$

Model 2



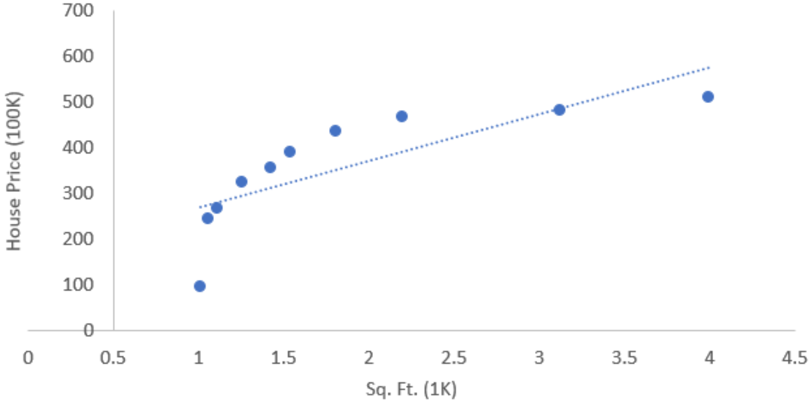
$$y = b_0 + b_1 X_1 + b_2 {X_1}^2$$

Bias-Variance Tradeoff



Model 1

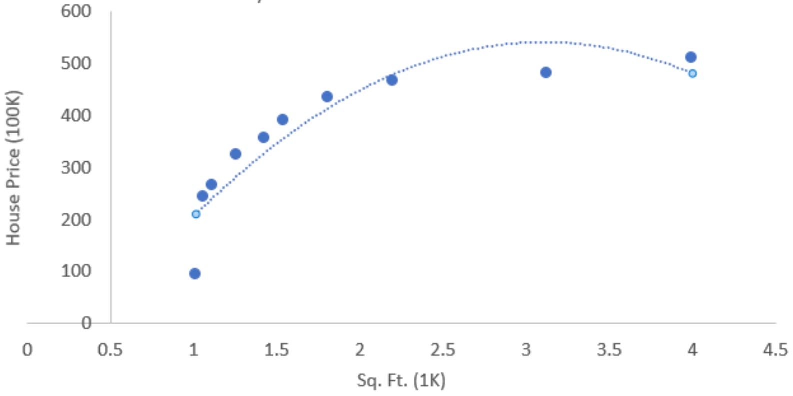
y = 102.74x + 165.89



$$y = b_0 + b_1 X_1$$

Model 2

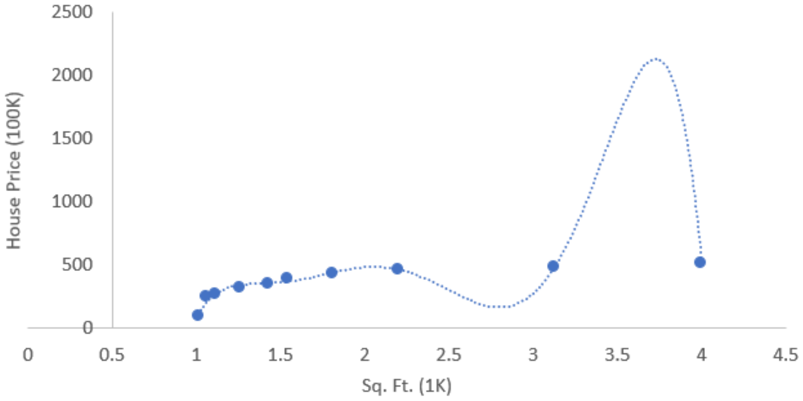
y = -75.58x² + 469.6x - 187.8



$$y = b_0 + b_1 X_1 + b_2 {X_1}^2$$

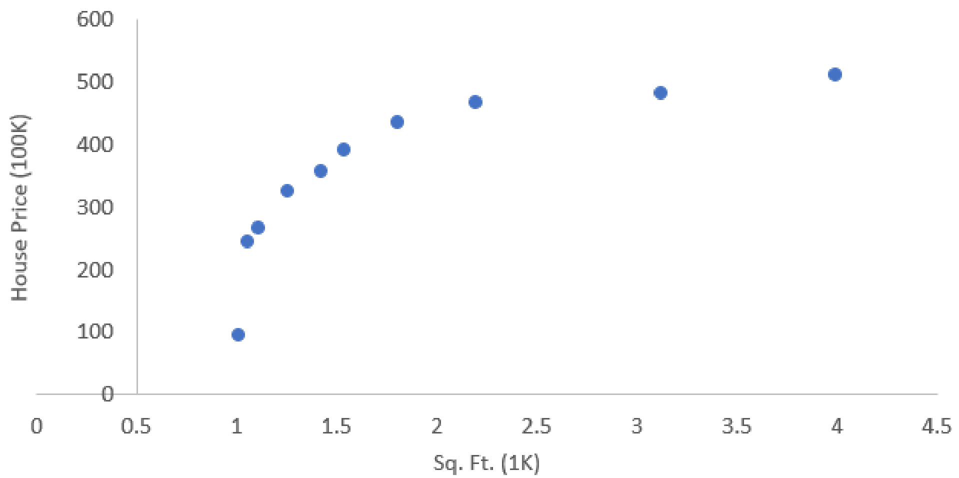
Model 3

y = -541.34x⁶ + 7355.4x⁵ - 39997x⁴ + 111513x³ - 168501x² + 131353x - 41098

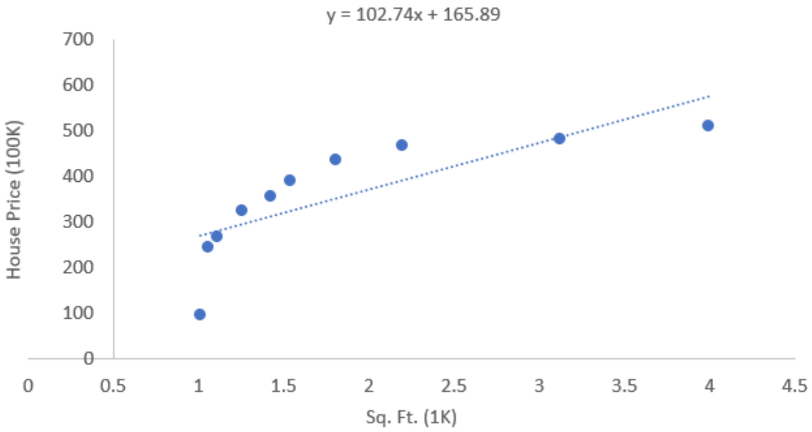


$$y = b_0 + b_1 X_1 + b_2 {X_1}^2 + b_3 {X_1}^3 + b_4 {X_1}^4 + b_5 {X_1}^5 + b_6 {X_1}^6$$

Bias-Variance Tradeoff



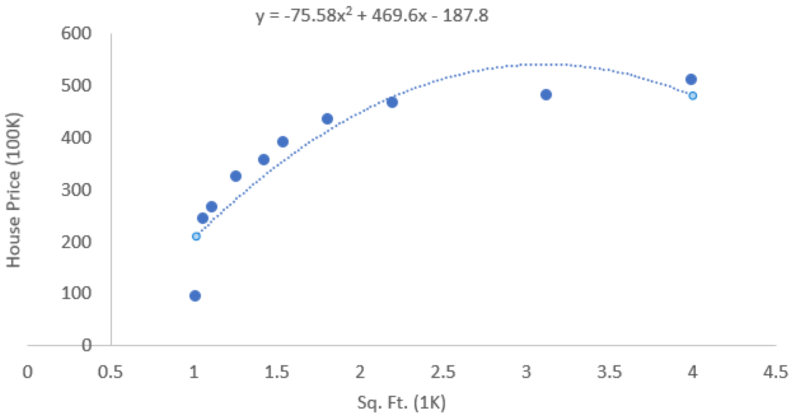
Model 1



$$y = b_0 + b_1 X_1$$

- Underfitting
- Biased

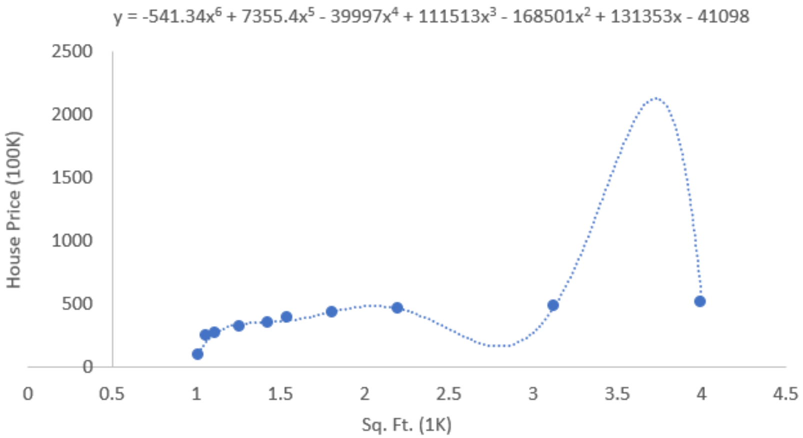
Model 2



$$y = b_0 + b_1 X_1 + b_2 {X_1}^2$$

- Just right

Model 3



$$y = b_0 + b_1 X_1 + b_2 {X_1}^2 + b_3 {X_1}^3 + b_4 {X_1}^4 + b_5 {X_1}^5 + b_6 {X_1}^6$$

- Overfitting
- High variance

Bias-Variance Tradeoff