Data Science

Lecture 10 - Regression

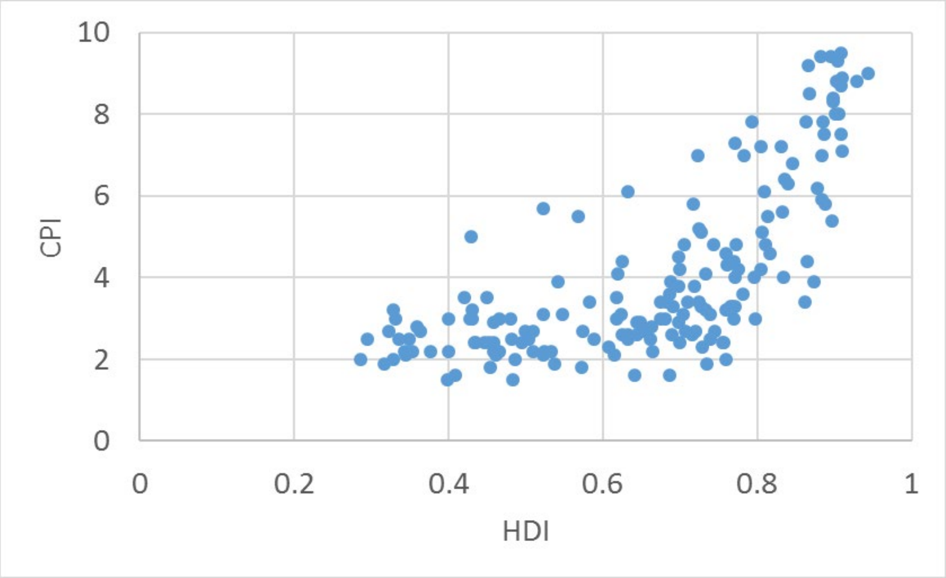
Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

BIG DATA
EXPERIENCE
CENTER

Motivation: why regression

- Consider the relationship between the human development index (HDI) and the corruption perception index (CPI) in Economist data.



- CPI seems to correlate with HDI
- Question: Given a desirable HDI, can we estimate the country CPI which reflects the cost?

Simple linear regression

- Consider modeling the CPI y_i as an “approximate” linear function of their HDI x_i.

$$y = mx + b + error$$

- There is an error term because this linear relationship is not perfect.
- y depends on other things besides x that we do not observe in our samples

Simple linear regression

- Why are we approaching the problem in this way? Here are three reasons.
  1. Sometimes you know x and just need to predict y (prediction)
  2. The conditional distribution is an excellent way to think about the relationship between two variables
  3. Linear relationships are easy to work with and are a good approximation in lots of real world problems

Simple linear regression

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2) i.i.d.$$
$$\varepsilon_i \text{ is independent of } X_i$$

- The intercept is α
- The slope is β
- We use the normal distribution to describe the "error"

Simple linear regression: remarks

- The parameters of our models are α, β and σ.
- The slope of β measures the change in y when x increases by 1 unit.
- The intercept α is the value y takes when $x = 0$
- The linear relationship holds for each pair (X_i, Y_i). Consequently, it is common to drop the subscripts and write $Y = \alpha + \beta X + \varepsilon$.
- The assumption that X is independent of ε is important. It implies that they are uncorrelated.

Interpretation of the regression parameters α, β and σ

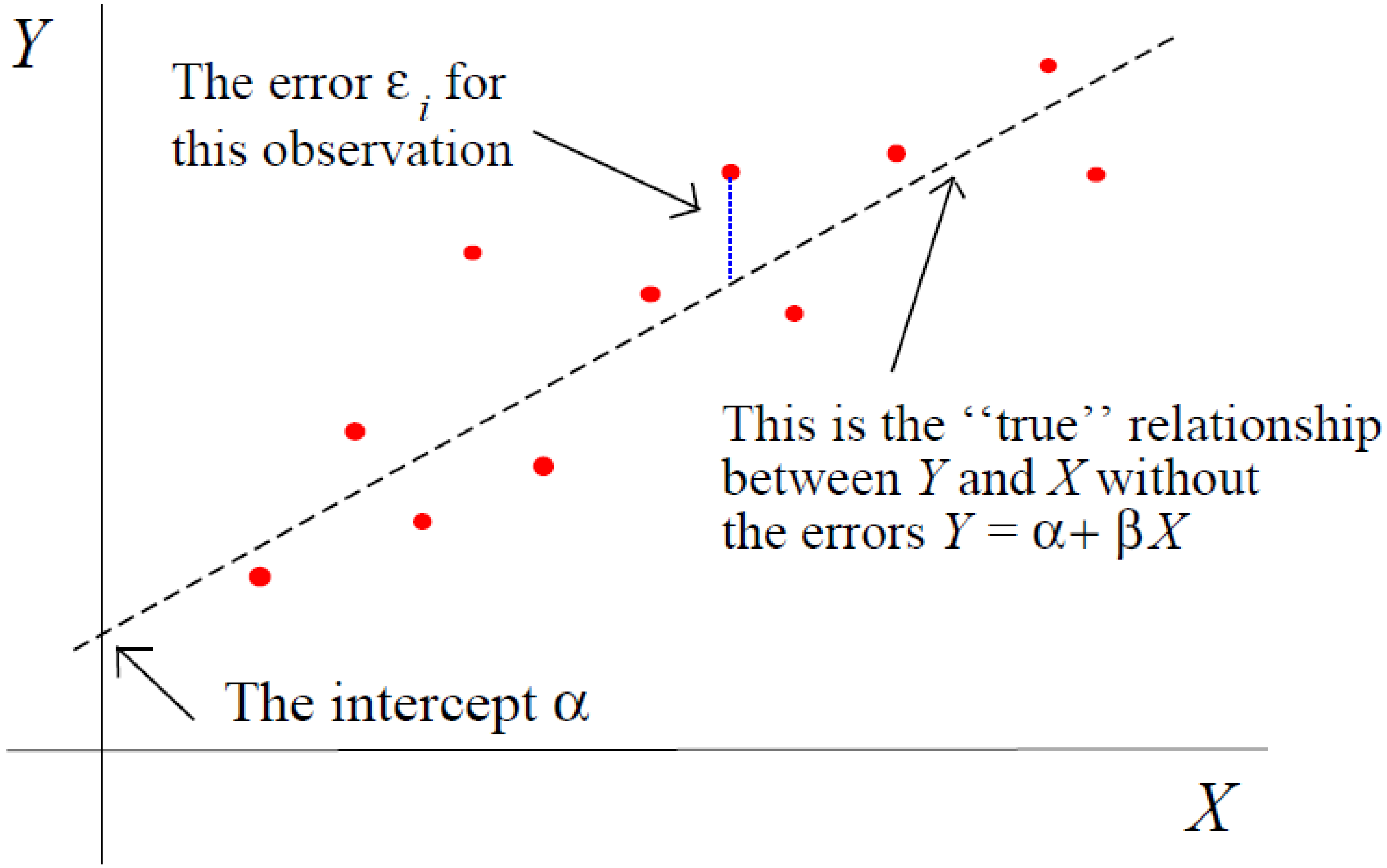
- Given a specific value X = x, how do we interpret α, β and σ.

β tells us: if the value we saw for X was one unit bigger, how much would our prediction for Y changes?

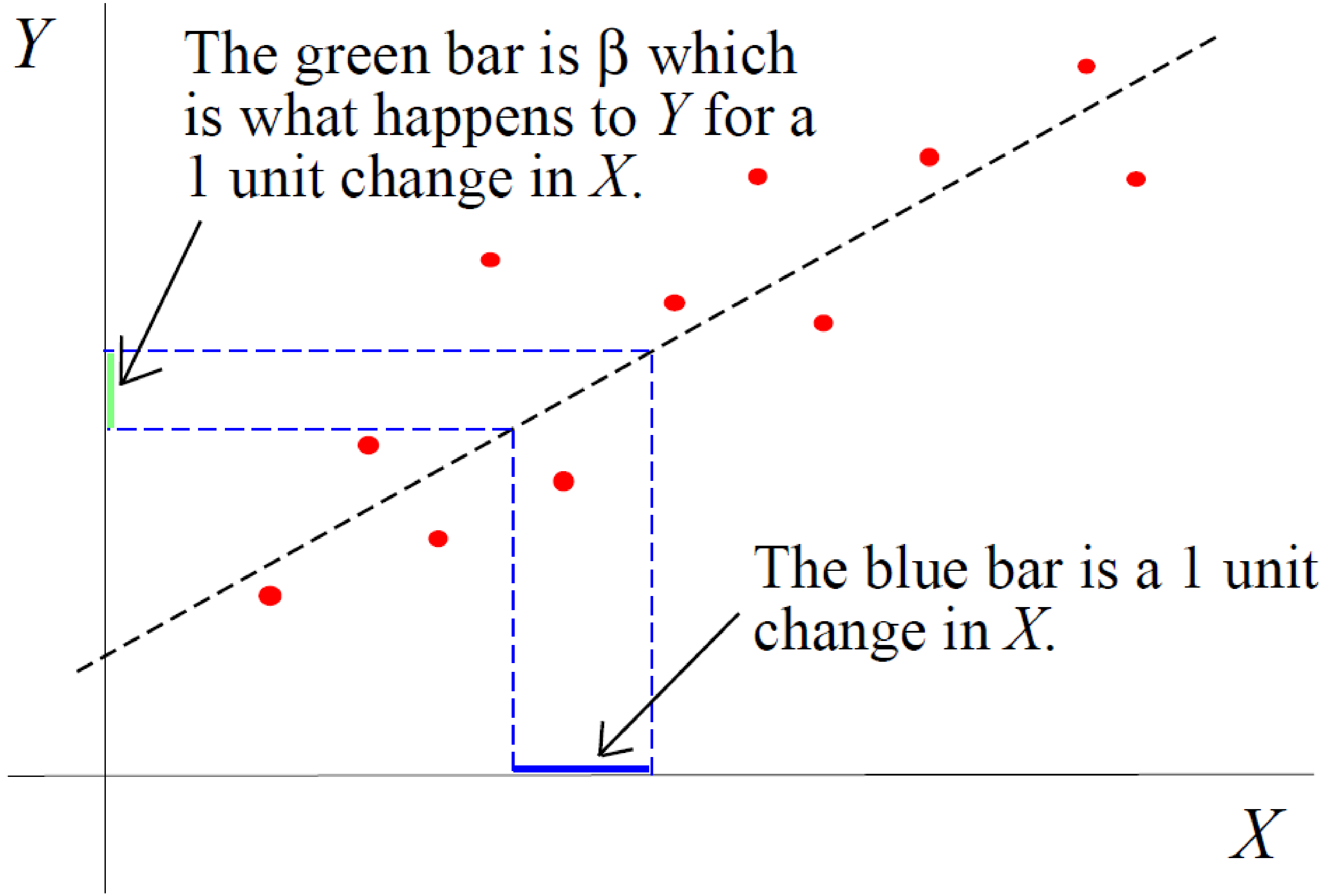
α tells us: what would we predict for Y if x = 0?

σ tells us: if $\alpha + \beta x$ is our prediction for Y given x, how big is the error associated with this prediction?

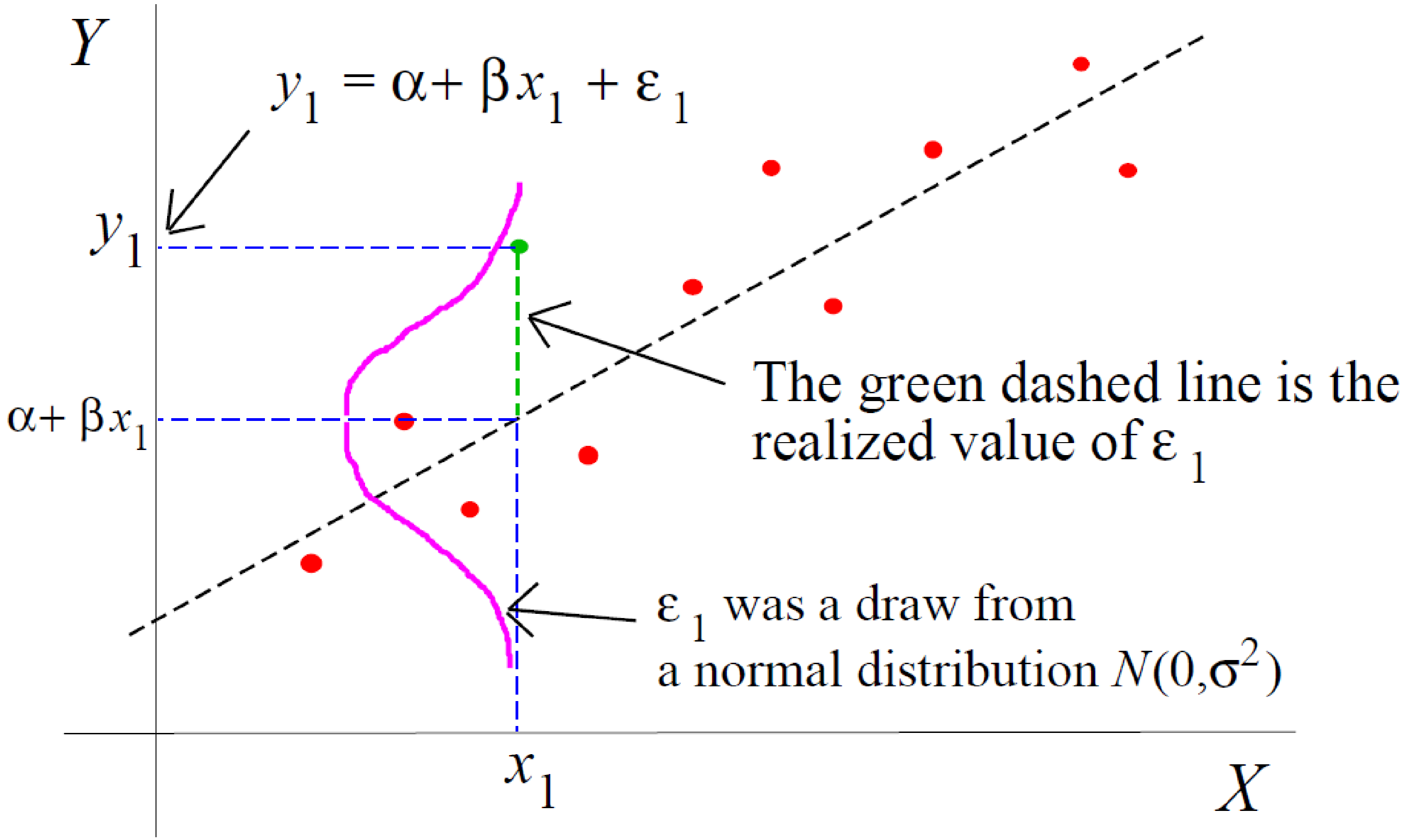
Interpretation of the regression parameters α, β and σ



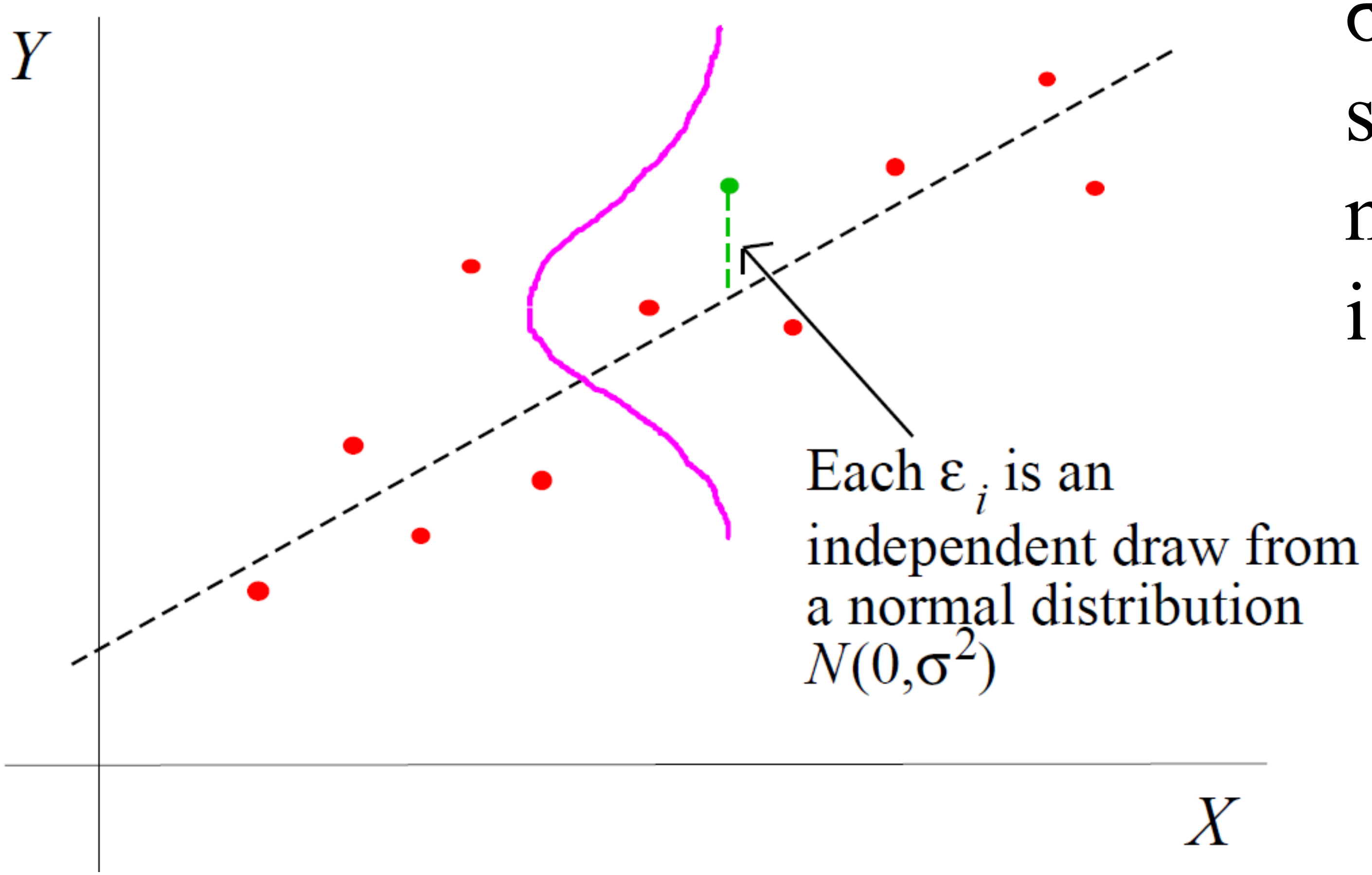
β: measures the slope of the line



α: How do we get y_1 given a specific value of $X_1 = x_1$

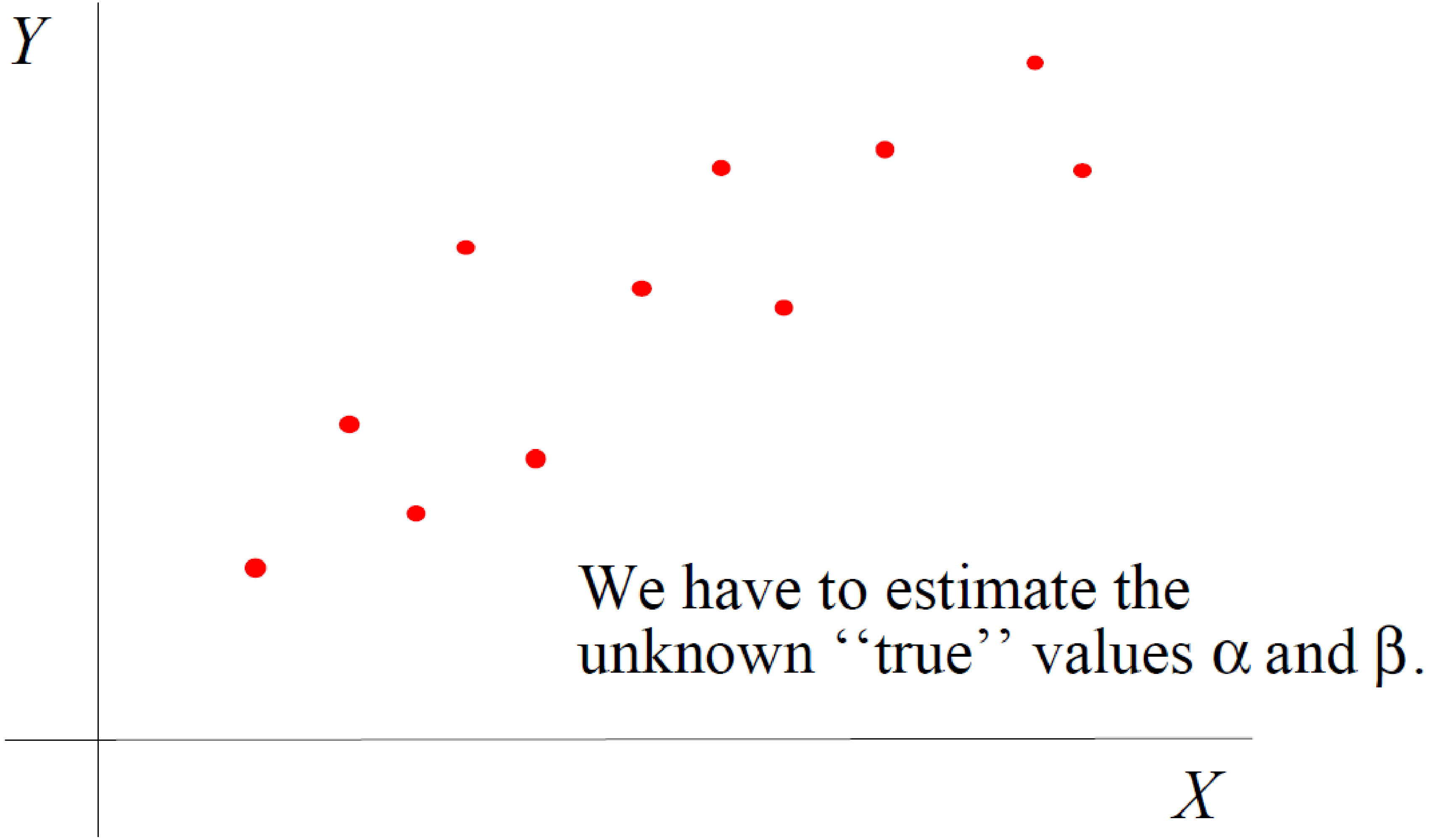


Each ε_i is i.i.d. $N(0,\sigma^2)$

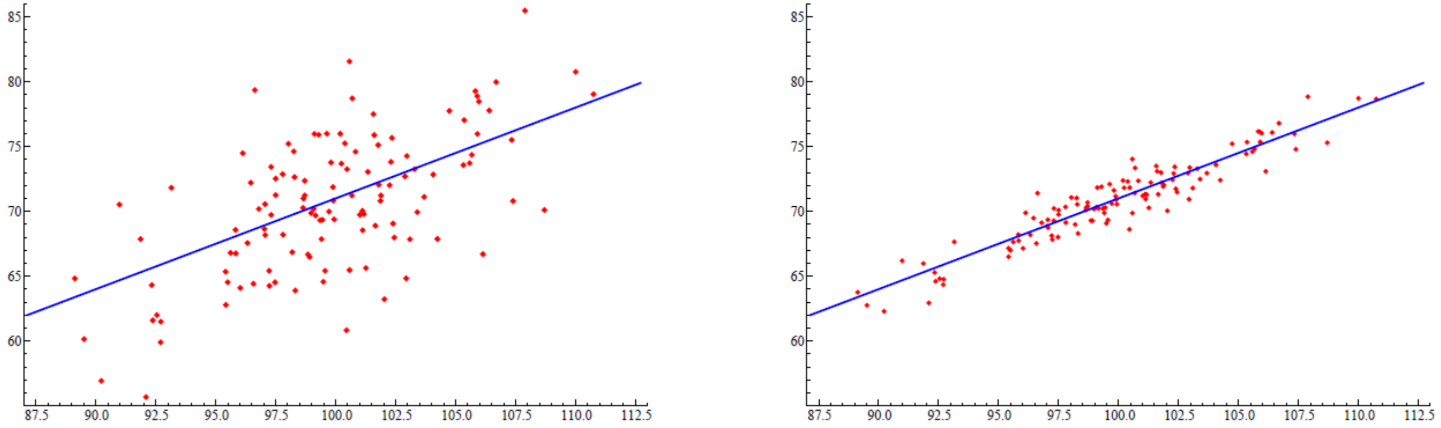


σ^2 measures the spread of the normal distribution, i.e. the error

In practice, we only observe the data!



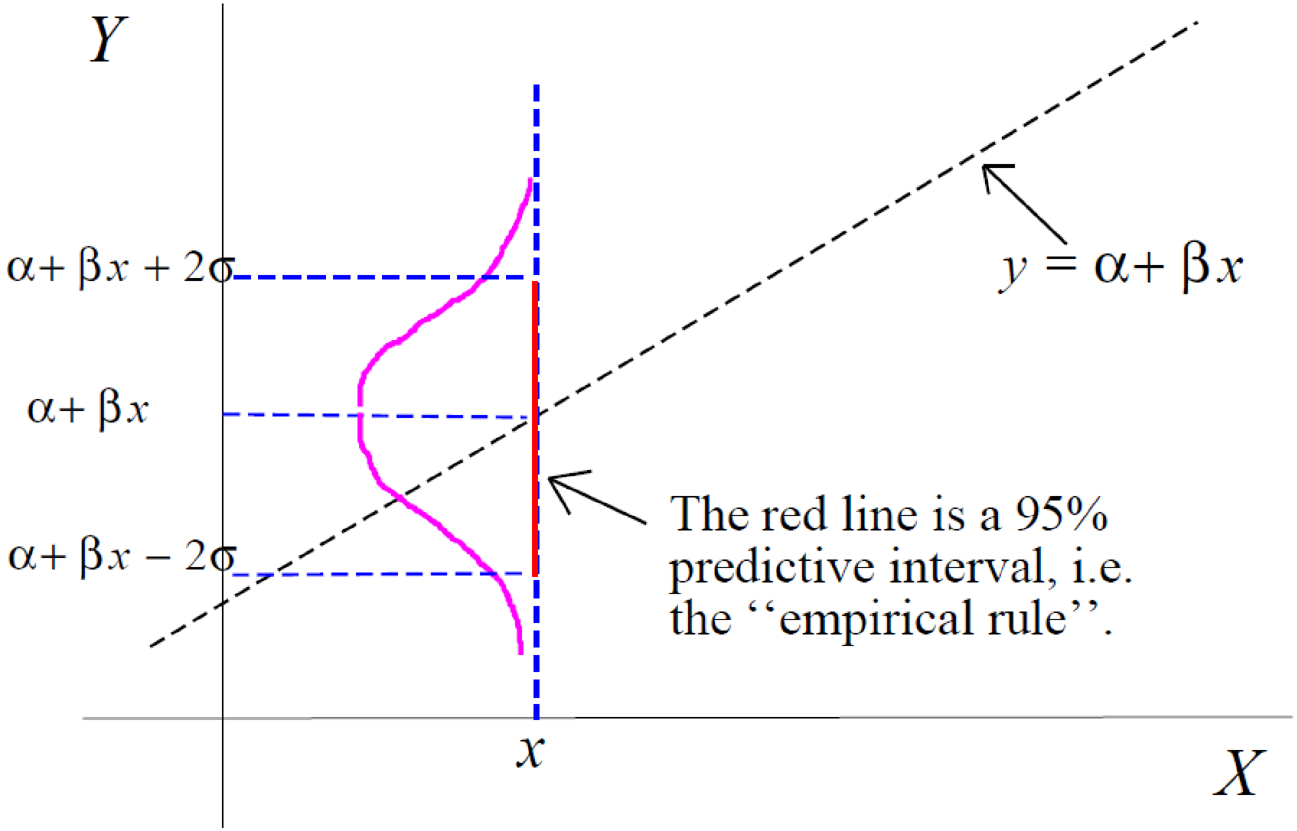
What role does the variance σ^2 play?



- The variance σ^2 of the error describes how big the error on average
- When σ^2 is smaller (right) the data are closer to the “true” regression line.
- The variance will determine how “wide” (or narrow) our predictive intervals are

Prediction using regression

Given a specific value for X = x, we can predict



Simple linear regression

- Simple linear regression model has one input x and one output y.
- The relationship can be explain as the following equation

$$f(\boldsymbol{x}) = w_0 + w_1 x$$

- Mathematically, parameters are obtained by least square method

Multiple regression

- Multiple linear regression model structure is exactly the same as the linear regression

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$

- Mathematically, parameters are obtained by least square method

Multiple regression with interaction

- Adding interaction terms to a regression model can greatly expand understanding of the relationships among the variables in the model
- This occurs when two or more variables depend on one another for the outcome

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + \cdots$$

Method of least squares

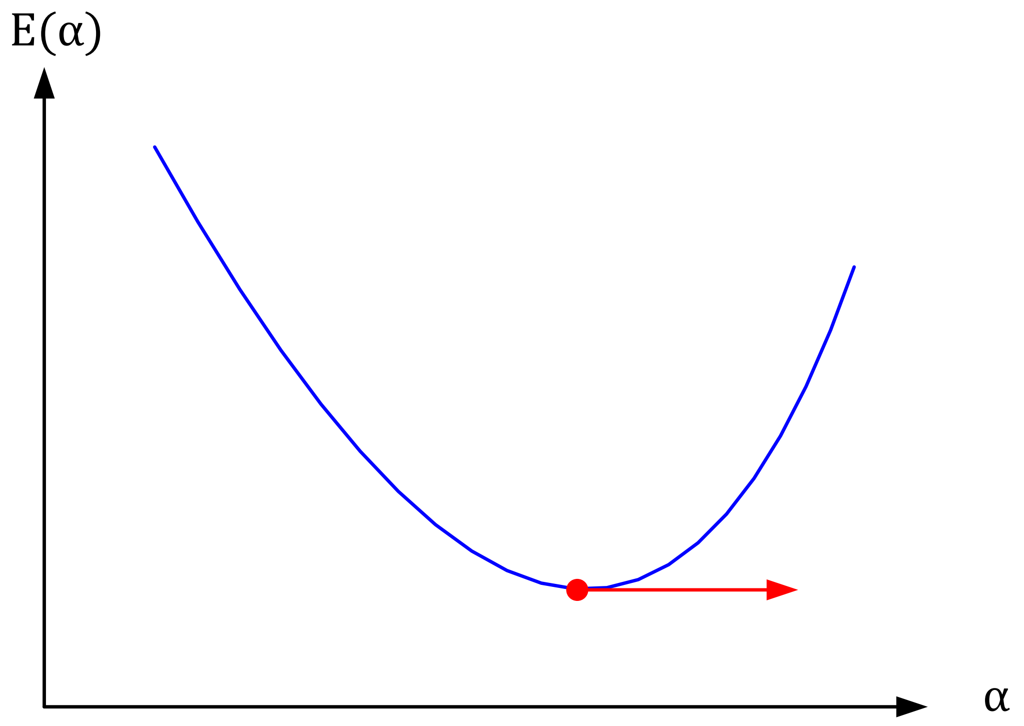
- Choose the β's so that the sum of the squares of the errors, ε_i, are minimized
- The least squares function is

$$S = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right)^2$$

OLS solution

Minimum of a function is the point where the slope is zero



Learning model parameters

$$S = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i \right)^2$$

$$\frac{\partial S}{\partial \alpha} \qquad \frac{\partial S}{\partial \beta}$$

unknown: α, β

Solving system of equations

calculated: α, β

Derivative of the error functions

The function S is to be minimized with respect to β_0, β_1

$$\frac{\partial S}{\partial \alpha} = -2\sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right) = 0$$

and

$$\frac{\partial S}{\partial \beta} = -2\sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right)x_i = 0$$

Least square normal equation

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Find alpha (intercept)

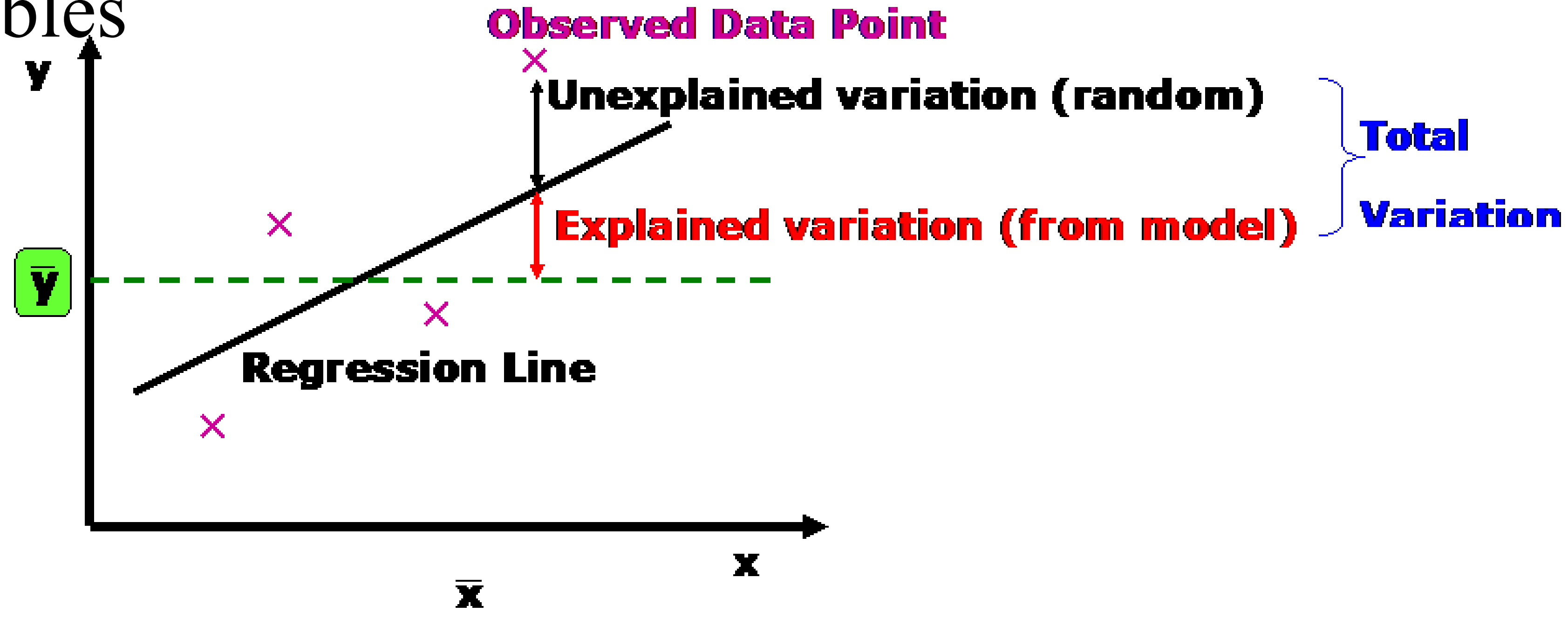
$$\alpha = \frac{\begin{vmatrix} \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Find beta (slope)

$$\beta = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Coefficient of determination (r^2)

The coefficient of determination is a number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables



Coefficient of determination

- The coefficient of determination R^2 (or sometimes r^2) is another measure of how well the least squares equation

$$Y = \alpha + \beta X$$

perform as a predictor of y
- R^2 is computed as:

$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SS_{yy}}{SS_{yy}} - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

- R^2 measures the relative sizes of SS_{yy} and SSE.
- The smaller SSE, the more reliable the predictions obtained from the model.

Coefficient of determination

- SS_{yy} measures the deviation of the observations from their mean:

$$SS_{yy} = \sum_i (y_i - \bar{y})^2$$

- SSE measures the deviation of observations from their predicted values

$$SSE = \sum_i (y_i - Y_i)^2$$

Coefficient of determination

- The higher the R^2, the more useful the model
- R^2 takes on values between 0 and 1
- Essentially, R^2 tells us how much better we can do in predicting y by using the model and computing Y than by just using the mean of y as a predictor.
- Note that when we use the model and compute Y the prediction depends on X because $Y = \alpha + \beta X$.
- Thus, we act as if x contains information about y.
- If we just use the mean of y to predict y, then we are saying that x does not contribute information about y and thus our predictions of y do not depend on x.

Evaluation: MAE, MAPE

$$MAE = \frac{1}{N}\sum \left|y_{true} - y_{pred}\right|$$

$$MAPE = \frac{1}{N}\sum \frac{\left|y_{true} - y_{pred}\right|}{\left|y_{true}\right|} \times 100$$

Lab: Load data

https://colab.research.google.com/drive/1gOQexVOk3mNwzlYsgS3jrMx1QjTnZdpq?usp=sharing

```
from sklearn.datasets import fetch_california_housing
import pandas as pd
import numpy as np

housing = fetch_california_housing()
housing
```

```
{'data': array([[   8.3252    ,   41.        ,    6.98412698, ...,    2.55555556,
          37.88      , -122.23      ],
        [   8.3014    ,   21.        ,    6.23813708, ...,    2.10984183,
          37.86      , -122.22      ],
        [   7.2574    ,   52.        ,    8.28813559, ...,    2.80225989,
          37.85      , -122.24      ],
        ...,
        [   1.7       ,   17.        ,    5.20554273, ...,    2.3256351 ,
          39.43      , -121.22      ],
        [   1.8672    ,   18.        ,    5.32951289, ...,    2.12320917,
          39.43      , -121.32      ],
        [   2.3886    ,   16.        ,    5.25471698, ...,    2.61698113,
          39.37      , -121.24      ]]),
 'target': array([4.526, 3.585, 3.521, ..., 0.923, 0.847, 0.894]),
```

Repackage the data

```python
X = pd.DataFrame(housing['data'], columns=housing['feature_names'])
X.head()
```

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
0	8.3252	41.0	6.984127	1.023810	32			
1	8.3014	21.0	6.238137	0.971880	240			
2	7.2574	52.0	8.288136	1.073446	49			
3	5.6431	52.0	5.817352	1.073059	55			
4	3.8462	52.0	6.281853	1.081081	56			

```python
X.shape
```

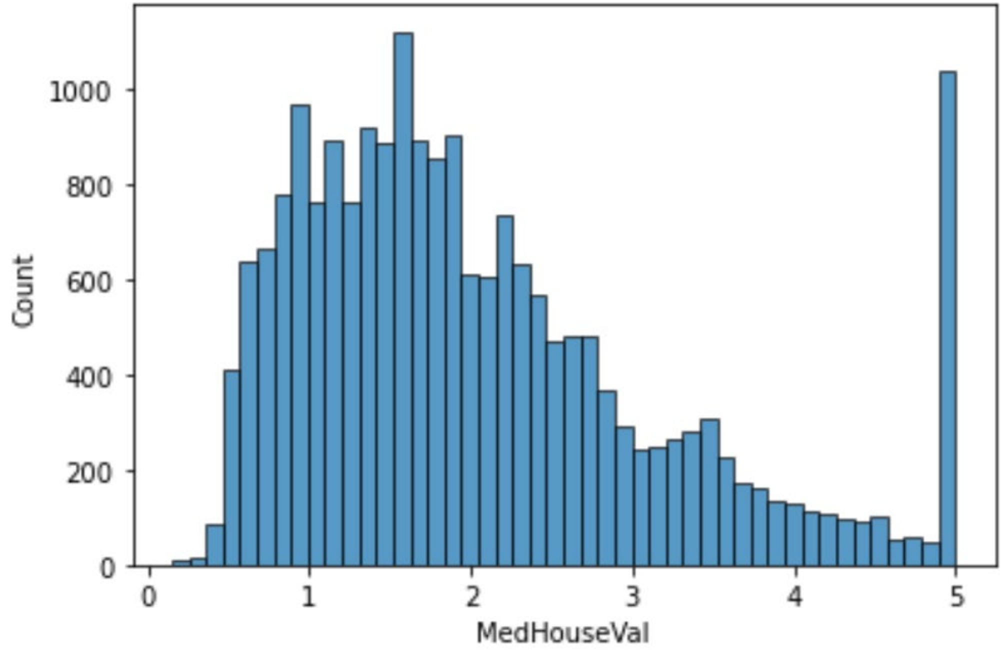
```
(20640, 8)
```

```python
y = pd.Series(housing['target'],
              name=housing['target_names'][0])
y
```

```
0        4.526
1        3.585
2        3.521
3        3.413
4        3.422
         ...
20635    0.781
20636    0.771
20637    0.923
20638    0.847
20639    0.894
Name: MedHouseVal, Length: 20640, dtype: float64
```

EDA
Distribution of y

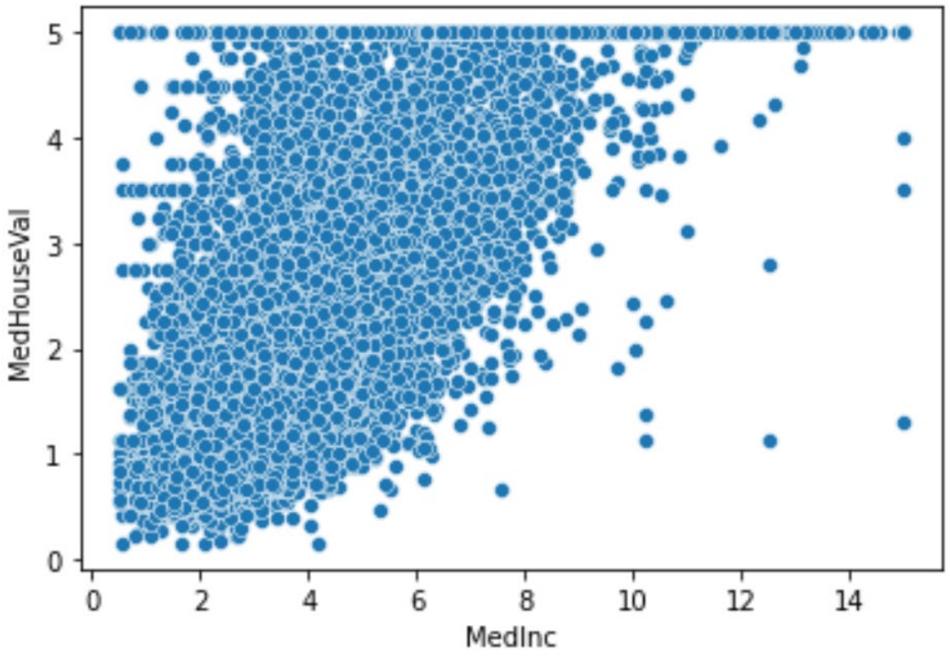
```python
import seaborn as sns
import matplotlib.pyplot as plt
sns.histplot(y)
plt.show()
```



EDA
Association with features

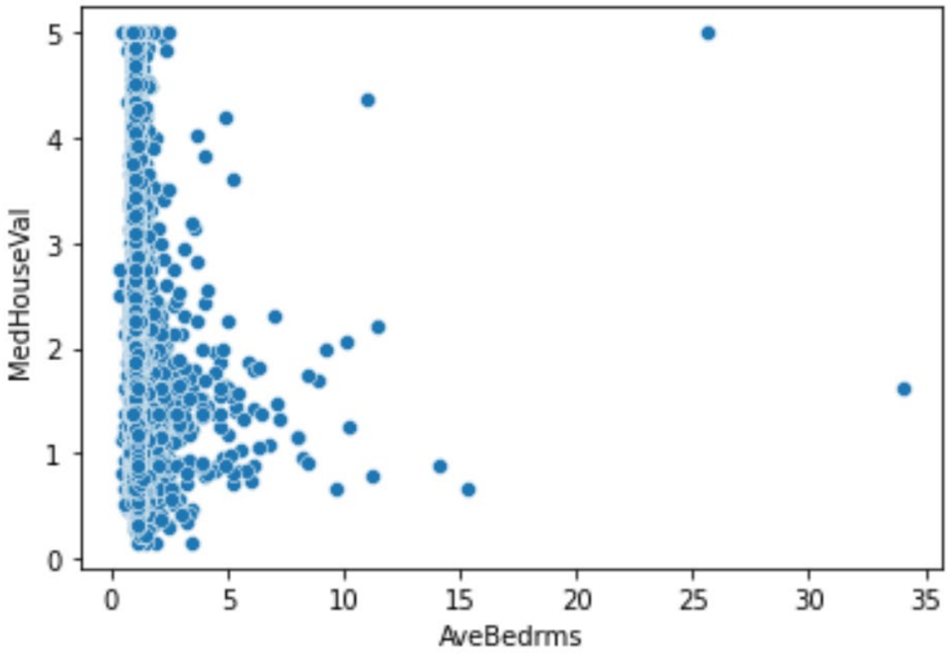
```
sns.scatterplot(x=X['MedInc'], y=y)
```

<AxesSubplot:xlabel='MedInc', ylabel='MedHouseVal'>




```
sns.scatterplot(x=X['AveBedrms'], y=y)
```

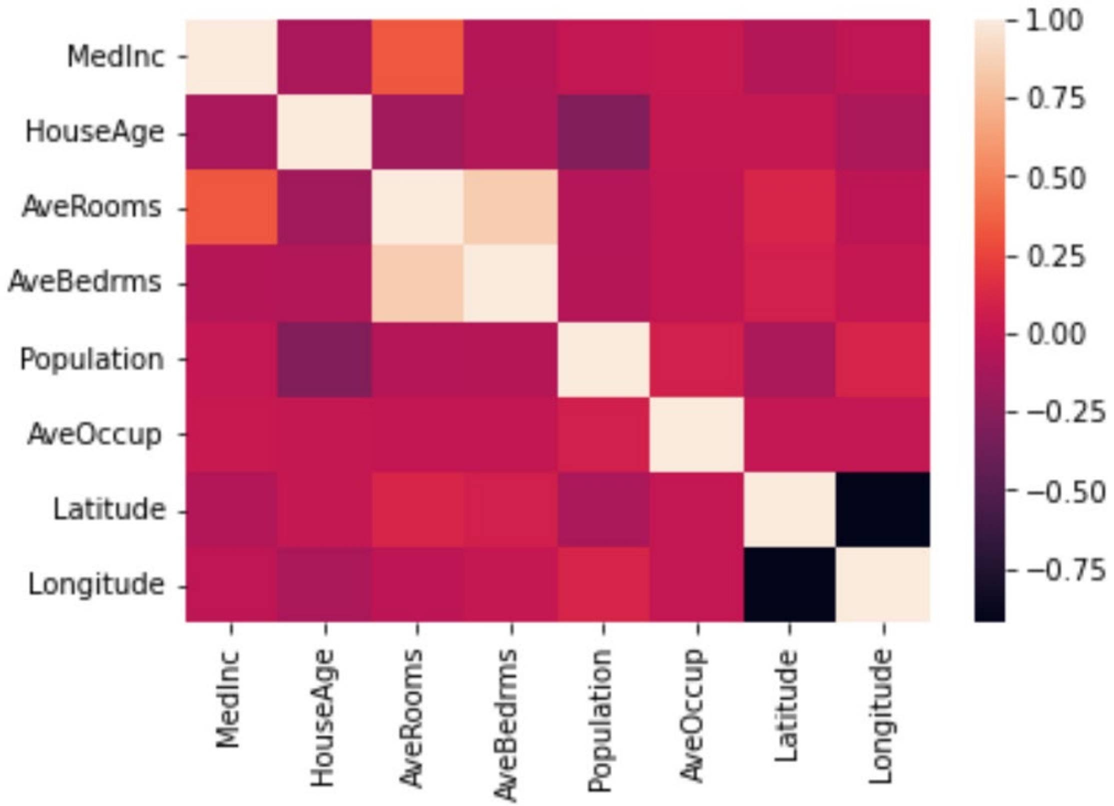
<AxesSubplot:xlabel='AveBedrms', ylabel='MedHouseVal'>



EDA
Feature correlation plot

```
sns.heatmap(X.corr())
```

<AxesSubplot:>



Train/test split

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7)
```

```
X_train.shape, y_train.shape
```

```
((14447, 8), (14447,))
```

```
X_test.shape, y_test.shape
```

```
((6193, 8), (6193,))
```

Modeling statmodels

```python
from statsmodels.api import OLS

lm = OLS(y_train, X_train).fit()
```

```python
lm.summary()
```

Dep. Variable:	MedHouseVal	**R-squared (uncentered):**	0.893
Model:	OLS	**Adj. R-squared (uncentered):**	0.893
Method:	Least Squares	**F-statistic:**	1.502e+04
Date:	Sat, 30 Jul 2022	**Prob (F-statistic):**	0.00
Time:	21:40:59	**Log-Likelihood:**	-16823.
No. Observations:	14447	**AIC:**	3.366e+04
Df Residuals:	14439	**BIC:**	3.372e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
MedInc	0.5182	0.005	100.904	0.000	0.508	0.528
HouseAge	0.0158	0.001	28.437	0.000	0.015	0.017
AveRooms	-0.1927	0.008	-25.579	0.000	-0.208	-0.178
AveBedrms	0.8584	0.035	24.452	0.000	0.790	0.927
Population	6.287e-07	6.03e-06	0.104	0.917	-1.12e-05	1.24e-05
AveOccup	-0.0042	0.001	-7.333	0.000	-0.005	-0.003
Latitude	-0.0643	0.004	-14.973	0.000	-0.073	-0.056
Longitude	-0.0169	0.001	-12.435	0.000	-0.020	-0.014

Evaluation

```
from sklearn.metrics import r2_score, mean_absolute_error, mean_absolute_percentage_error
r2 = r2_score(y_true=y_test, y_pred=y_pred)
mae = mean_absolute_error(y_true=y_test, y_pred=y_pred)
mape = mean_absolute_percentage_error(y_true=y_test, y_pred=y_pred)
print('R2: %0.3f'%r2)
print('MAE: %0.3f'%mae)
print('MAPE: %0.3f'%mape)
```

```
R2: 0.540
MAE: 0.579
MAPE: 0.349
```

Modeling
scikit-learn

```
from sklearn.linear_model import LinearRegression

lr = LinearRegression()
lr.fit(X_train, y_train)
```

```
LinearRegression()
```

```
y_pred_lr = lr.predict(X_test)
y_pred_lr
```

```
array([3.39091271, 2.57418805, 3.33535536, ..., 2.43561109, 1.85466479,
       2.79648288])
```

Evaluation

```
r2 = r2_score(y_true=y_test, y_pred=y_pred_lr)
mae = mean_absolute_error(y_true=y_test, y_pred=y_pred_lr)
mape = mean_absolute_percentage_error(y_true=y_test, y_pred=y_pred_lr)
print('R2: %0.3f'%r2)
print('MAE: %0.3f'%mae)
print('MAPE: %0.3f'%mape)
```

```
R2: 0.599
MAE: 0.540
MAPE: 0.324
```

Summary

- Simple linear regression
- Multiple regression
- Error measurements
- Implementation of regression in statmodels
- Implementation of regression in scikit-learn

Activity

- Data: https://www.kaggle.com/datasets/shivachandel/kc-house-data
- Build a model to predict the house price
- Evaluate the model
- Explain relevant variables

Thank you

Question?

BIG DATA
EXPERIENCE
CENTER