Lecture 12: Linear Regression

Lecturer: Jie Fu

A motivating example

- The leaning tower of Pisa continuously tilts over time.
- Measurements between 1975 and 1987 of the "lean" of a fixed point on the tower (the distance in measures of the actual position of the point, and its position if the tower was straight).
- We are interested in predicting the lean of the tower for year 1988 based on the data.
- How to do that?

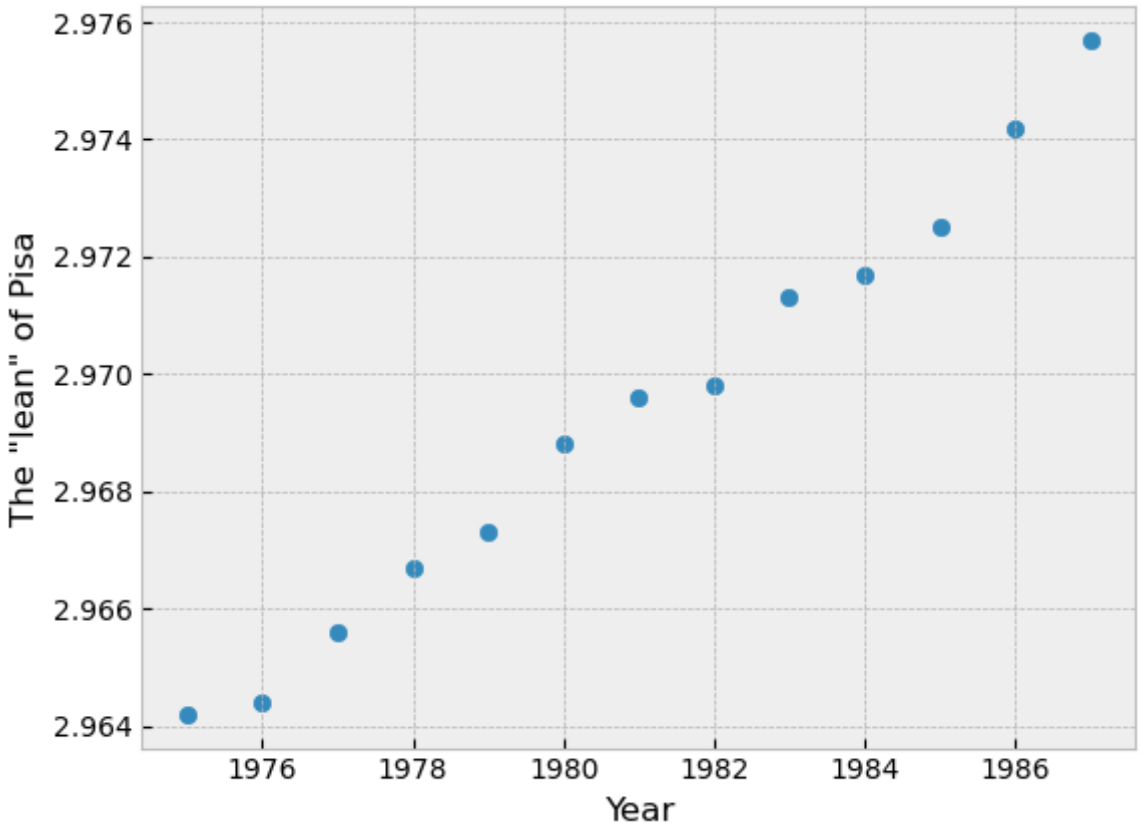
year	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
lean	2.9642	2.9644	2.9656	2.9667	2.9673	2.9688	2.9696	2.9698	2.9713	2.9717	2.9725	2.9742	2.9757

A general problem statement

- We consider the case of only two variables for illustration.
- We wish to model the relation between two variables of interest, x and y
  - e.g., year and the tilt of the Pisa tower.
- based on a collection of data pairs (x_i, y_i), $i = 1, \ldots, n$.
  - e.g. x_i is the year, and y_i the measured "lean" of the tower in that year.

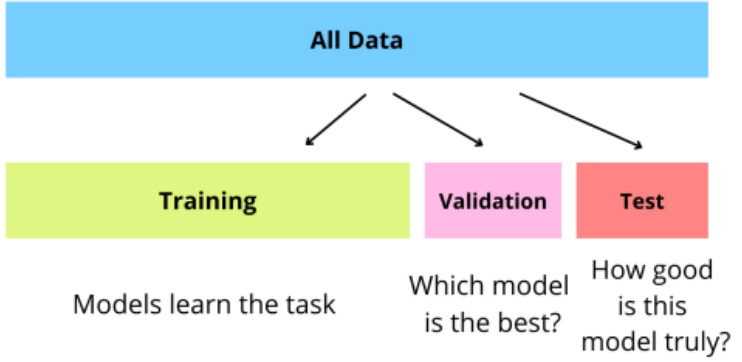
Visualize the data in python

```
x = range(1975, 1988)
print(len(x))
y = np.array([2.9642, 2.9644 ,2.9656,
2.9667, 2.9673, 2.9688, 2.9696,
2.9698,  2.9713, 2.9717,  2.9725,
2.9742, 2.9757])
print(len(y))
plt.scatter(x,y, label='Data')
plt.xlabel('Year')
plt.ylabel('The "lean" of Pisa')
```



Data

- Training data:
  - Used to train the model by adjusting its parameters. The model learns patterns from this data.
- Validation data:
  - Used to evaluate model performance during training.
- Test data:
  - Used to assess the final model's performance on unseen data. Provides an unbiased evaluation of its generalization ability.
  - Not available during any part of the learning process



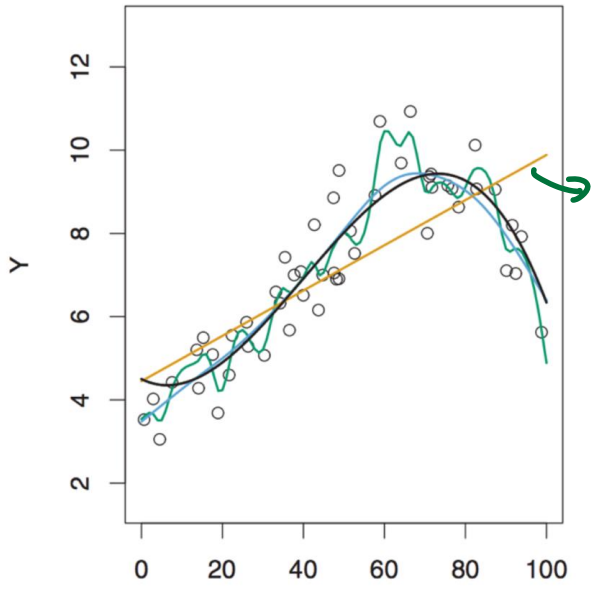
Model Flexibility

As model flexibility increases, training error will decrease.

The model can fit more and more of the variance in the training set.

The test error may or may not decrease.

If training error is much larger than test error, the model is overfitting.



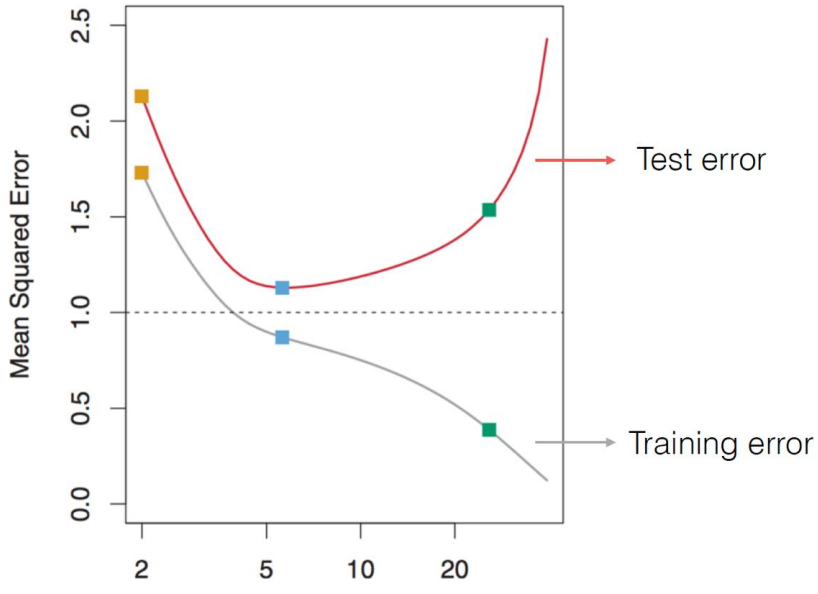


FIGURE 2.9, ISL (8th printing 2017)

different predictors $f(x)$ with different degree of polynomial.

Linear regression

- Linear regression assumes that the relationship between the input variables (features) and the output variable (target) can be described by a straight line:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y: the dependent variable (output)
- X is the independent variable (input)
- β_0 is the intercept
- β_1 is the slope
- ε is the error term

Simple Linear Regression

- Estimate βs using training data:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

The estimated value of βs are $\hat{\beta}$s

For a particular realization of X, aka $X = x$, the predicted output is denoted $\hat{y}$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Goal: Pick $\hat{\beta}_0, \hat{\beta}_1$, such that the model is a good fit to the training data.

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i, \forall i = 1, \ldots, n$$

Example:

House sizes:

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178

Hypothesis:

$$h_\beta(x) = -40 + 0.25x$$

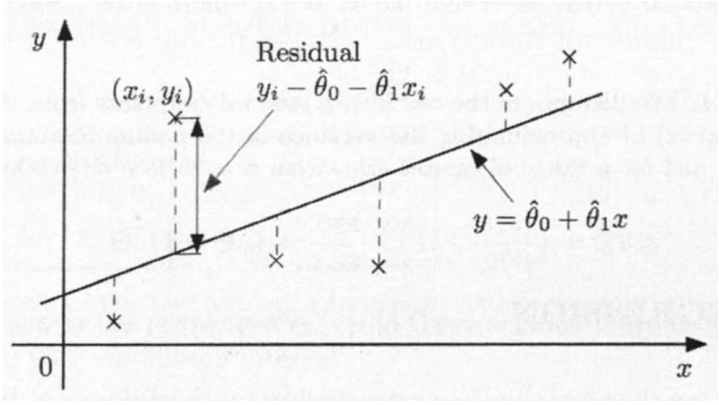
Use matrix-vector multiplication:

$$\begin{bmatrix} h_\beta(x^{(1)}) \\ h_\beta(x^{(2)}) \\ \vdots \\ h_\beta(x^{(n)}) \end{bmatrix} = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(n)} \end{bmatrix} \cdot \underbrace{\begin{bmatrix} -40 \\ 0.25 \end{bmatrix}}_{\beta}$$

- Residual for the i-th sample:

$$\varepsilon_i = y^{(i)} - \hat{y}^{(i)}$$

$$= y^{(i)} - (\hat{\beta}_0 + \hat{\beta}_1 x^{(i)})$$



- Residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2$$

- The model fit using least squares finds the parameters, that minimize the RSS:

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{\beta_0, \beta_1} \left[\sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x^{(i)} \right) \right)^2 \right]$$

Taking gradient $= 0$

$$\begin{cases} \dfrac{\partial RSS}{\partial \beta_0} = 0 \\ \dfrac{\partial RSS}{\partial \beta_1} = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} 2\left(y_i - (\beta_0 + \beta_1 x_i)\right) \cdot (-1) = 0 \\ \sum_{i=1}^{n} 2\left(y_i - (\beta_0 + \beta_1 x_i)\right) \cdot (-x_i) = 0 \end{cases}$$

- It can be verified that:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X,Y)}{Var(X)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\bar{x}$: mean of $\{x^{(1)}, x^{(2)} \cdots x^{(n)}\}$

$\bar{y}$: mean of $\{y^{(1)}, y^{(2)} \cdots y^{(n)}\}$.

How good is the model fit?

- Residual standard error (RSE):

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where:

- n is the number of observations,
- p is the number of predictors (independent variables). (features) used in a regression model to predict the dependent variable. In **simple linear regression**, there is only **one predictor** (p = 1)

- It represents the standard deviation of residuals and helps assess the goodness of fit for a regression model.

How good is the model fit?

- R2 statistic:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

the total variance of y explained by x.

- RSS: Residual sum of squares (RSS)

$$\text{RSS} = \epsilon^{(1)2} + \epsilon^{(2)2} + \ldots + \epsilon^{(n)2} = \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2$$

- TSS: Total sum of squares (TSS)

$$\text{TSS} = \sum_{i=1}^{n} (y^{(i)} - \bar{y})^2$$

How good is the model fit?

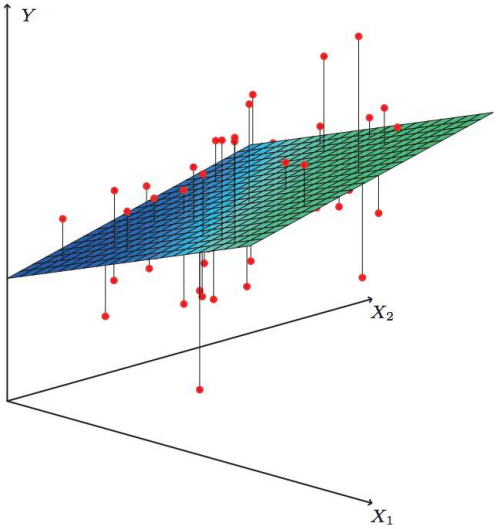
- If we know the data comes from a linear model, then *R2 close to 1 in order to call the fit good.*

- Otherwise, if the linear model is only a crude approximation, then *with R2 of 0.4 may be considered good.*

Multiple Linear Regression

- Predict the response variable using more than one predictor variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad (1)$$



$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$

$\hat{y}$: predicted value.

$\hat{\beta}_i$: parameters from L.R.

Data: $(x_1^{(1)}, x_2^{(1)}, \cdots x_p^{(1)}, y^{(1)})$

$(x_1^{(2)}, x_2^{(2)} \cdots x_p^{(2)}, y^{(2)})$

⋮

} n data points

Residual sum of squares.

A company wants to predict the **price of a house** based on the following features:

- **Size** (x_1) in square feet
- **Number of bedrooms** (x_2)
- **Age of the house** (x_3) in years

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

Size (x_1)	**Bedrooms** (x_2)	**Age** (x_3)	**Price** (y) **(in 1000s)**
1500	3	10	300
2000	4	5	450
2500	4	2	500
1800	3	20	280
2200	5	7	480

Table 1: House Price Prediction Data

Multiple Linear Regression

- Using matrix notation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad (1)$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

n data point

$$\vec{x}^{(k)} : \left(x_1^{(k)}, x_2^{(k)} \cdots x_p^{(k)} \right), \quad y^{(k)}$$

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ 1 & & & & \\ \vdots & & & & \\ 1 & x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

linear matrix equation (LME)

n equations;

$$X \cdot \vec{\beta} = \vec{y} \qquad \vec{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

Least squares

- RSS: Residual sum of squares (RSS)

$$\text{RSS} = \epsilon^{(1)2} + \epsilon^{(2)2} + \ldots + \epsilon^{(n)2} = \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2$$

- The model fit using least squares finds that minimize the RSS.

predicted.

$\hat{y} = X \cdot \hat{\beta}$

$\vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$

$\sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2 = \| \vec{y} - \hat{\vec{y}} \|_2^2$ L-2 norm

$= \| \vec{y} - X\hat{\beta} \|_2^2$

$\|X\|_2^2 = X^T X$

$\hat{\beta}^* = \underset{\hat{\beta}}{\arg\min} \| \vec{y} - X\hat{\beta} \|_2^2$

$\Rightarrow \frac{\partial}{\partial \hat{\beta}} \underbrace{\left((\vec{y} - X\hat{\beta})^T (\vec{y} - X\hat{\beta}) \right)}_{f(\hat{\beta})} = 0$

- Analytical solution to least squares

$$\mathbf{X}^\top \mathbf{X}\hat{\beta} = \mathbf{X}^\top \mathbf{Y} \quad \Rightarrow \quad \hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$$

$n \times (p+1)$

$\frac{\partial (f(\hat{\beta}))}{\partial \beta_0} = 0$

$\frac{\partial f(\hat{\beta})}{\partial \beta_1} = 0$

$\vdots$

$\frac{\partial f(\hat{\beta})}{\partial \beta_p} = 0$

$\Rightarrow$

X^T: $(p+1) \times n$ · X: $n \times (p+1)$ $\Rightarrow (p+1) \times (p+1)$

A: square: inverse of A: $A^{-1} = B$

$B \cdot A = I$

$A^{-1} \cdot A = I$

Non-linearity of Data

2nd: $X = [1, 2, \cdots 10]$

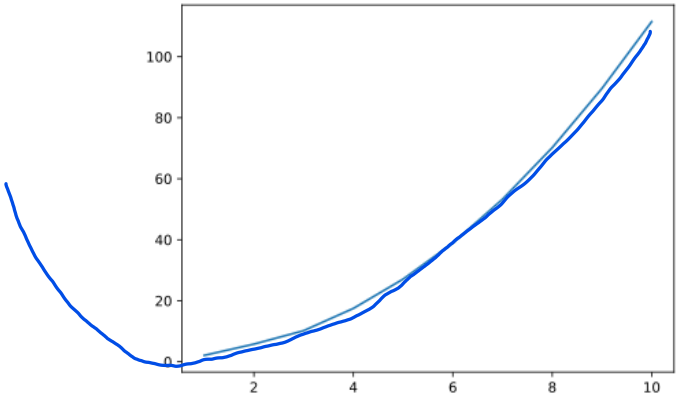
Augment $X_1 = X = [1, 2, \cdots 10]$

$X_2 = X^2 = [1, 4, 9, \cdots 100]$

- A simple way to extend linear regression to model non-linear relationships is via polynomial regression.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \epsilon \quad (1)$$

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$,



x = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
y = [2.1, 5.8, 10.2, 17.5, 27.1, 38.9, 53.3, 70.2, 89.6, 111.5])

- Which degree of polynomials to select? intuition, try
- Can we derive the analytical solution? Yes.

Feature: x

multiple features: $X_1 = x$, $X_2 = x^2$, $X_3 = x^3$ ... $X_n = x^n$,