CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD COLLEGE

Admin

- Lab 2 posted, due Monday
- My office hours: 3-4:30pm on Monday (H110)

TA Hour Schedule (H204)

Sundays	**4-6pm**	**Grace**
Mondays	**7-9pm**	**Trinity**
Wednesdays	**6:30-8:30pm**	**Henry**
Thursdays	**7:30-9:30pm**	**Ella**

Command line arguments example

```
def parse_args():
    """Parse command line arguments (train and test data files)."""
    parser = optparse.OptionParser(description='climate change model analysis')

    # specify all command line options here
    parser.add_option('-r', '--train_filename', type='string', help='path to' +\
        ' train csv file')
    parser.add_option('-e', '--test_filename', type='string', help='path to' +\
        ' test csv file')

    (opts, args) = parser.parse_args()

    # mandatory arguments
    mandatories = ['train_filename']
    for m in mandatories:
        if not opts.__dict__[m]:
            print('mandatory option ' + m + ' is missing\n')
            parser.print_help()
            sys.exit()

    return opts

def main() :
    opts = parse_args()
    print(opts.train_filename)
```

Outline for September 14

- Why are models useful? (recap)
- Linear models
- Fitting a linear model (one feature)
- Model complexity and evaluation

Outline for September 14

- Why are models useful? (recap)
- Linear models
- Fitting a linear model (one feature)
- Model complexity and evaluation

Why are models useful?

- Understand/explain/interpret the phenomenon
- Predict outcomes for future examples

Outline for September 14

- Why are models useful? (recap)
- Linear models
- Fitting a linear model (one feature)
- Model complexity and evaluation

Goals of fitting a linear model

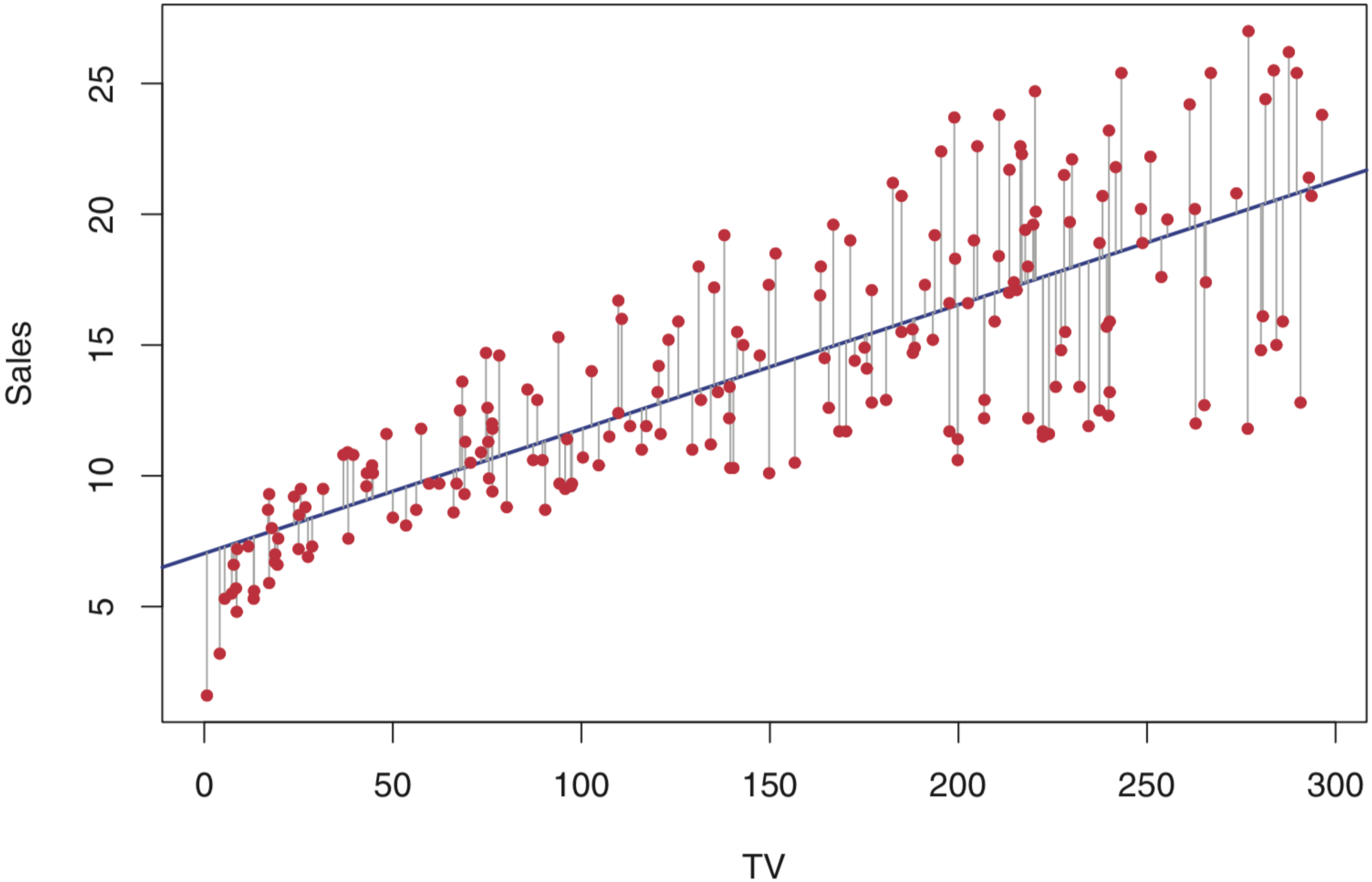
1) Which of the features/explanatory variables/predictors (x) are associated with the response variable (y)?

2) What is the relationship between x and y?

3) Is a linear model enough?

4) Can we predict y given a new x?

Example: predict sales from TV advertising budget



ISL: Figure 3.1

Maybe a linear model is not enough

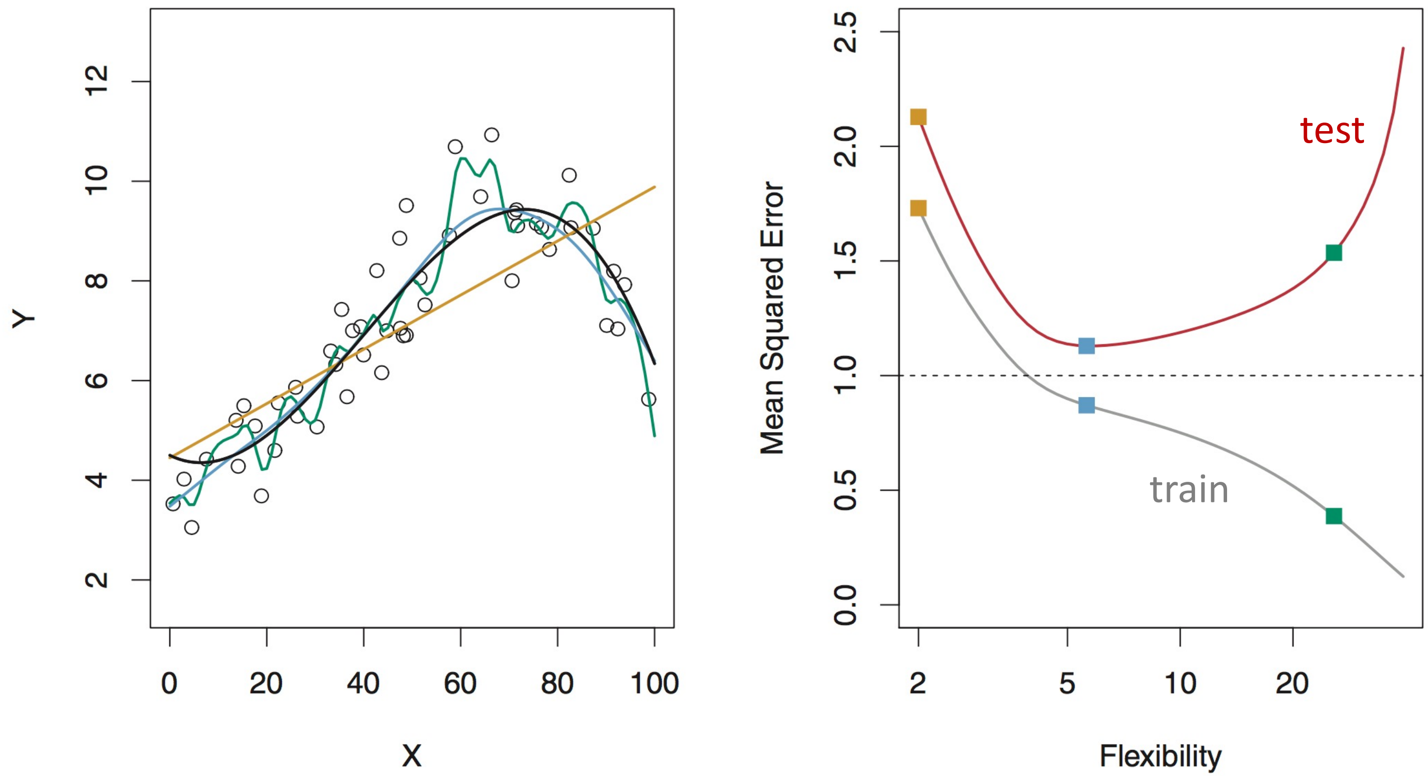
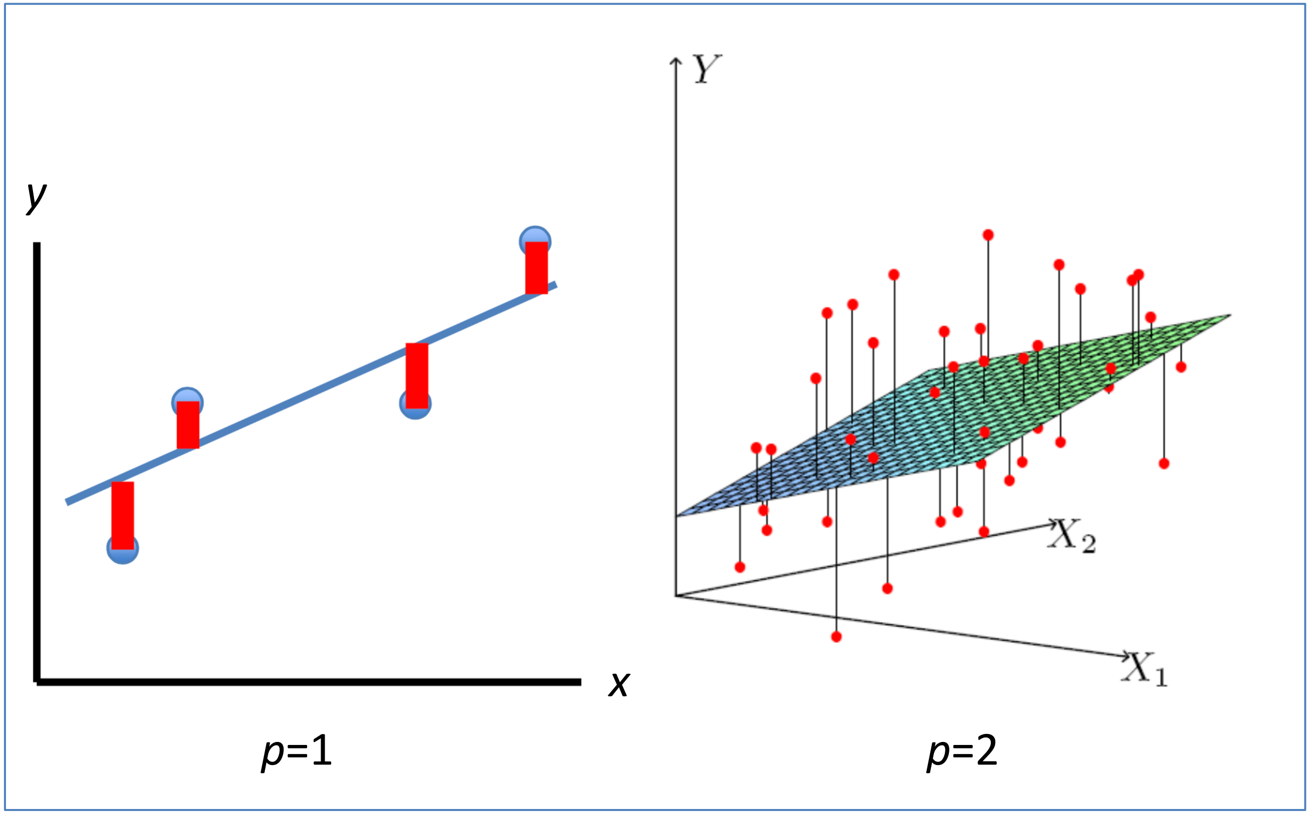


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

ISL, Fig 2.9

Linear model with 1 or 2 features



Slide: modified from Jessica Wu
Original: Eric Eaton

Linear Regression

- Output (*y*) is continuous, not a discrete label

- <u>Learned model</u>: *linear function* mapping input to output (a *weight* for each feature + *bias*)

- <u>Goal</u>: minimize the *RSS* (residual sum of squared errors) or *SSE* (sum of squared errors)

Outline for September 14

- Why are models useful? (recap)
- Linear models
- Fitting a linear model (one feature)
- Model complexity and evaluation

model

$$h_{\vec{w}} = w_0 + w_1 x = \hat{y}$$

$$\vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

Goal minimize

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

y_i ↑ truth, $\hat{y}_i$ ↑ prediction

SSE: sum of squared errors

RSS: residual sum of squares

cost function

$$J(w_0, w_1) = \frac{1}{2}\sum_{i=1}^{n}(\overbrace{y_i - w_0 - w_1 x_i}^{\text{model}})^2$$

take derivative & set to 0

$$\frac{\partial J}{\partial w_0} = 0, \quad \frac{\partial J}{\partial w_1} = 0$$

(a) $$\frac{\partial J}{\partial w_0} = -\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i) = 0$$

$$-\frac{1}{n}\sum_{i=1}^{n} y_i + w_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n w_0}{n}$$

$\bar{x}$ avg of x's

$\bar{y}$ avg of y's

$$\boxed{\hat{w}_0 = \bar{y} - w_1 \bar{x}}$$

res

(b) $$\frac{\partial J}{\partial w_1} = -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) x_i = 0$$

$$= -\sum_{i=1}^{n} (y_i - \bar{y} + w_1 \bar{x} - w_1 x_i) x_i = 0$$

$$= \sum_{i=1}^{n} (y_i x_i - \bar{y} x_i) = w_1 \sum_{i=1}^{n} (x_i^2 - \bar{x} x_i)$$

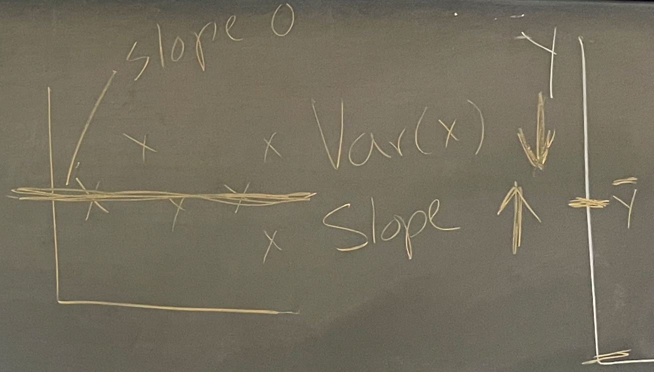
$$= \hat{w}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \bar{y} x_i}{\sum_{i=1}^{n} (x_i^2 - \bar{x} x_i)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{Cov(X, y)}{Var(X)}$$

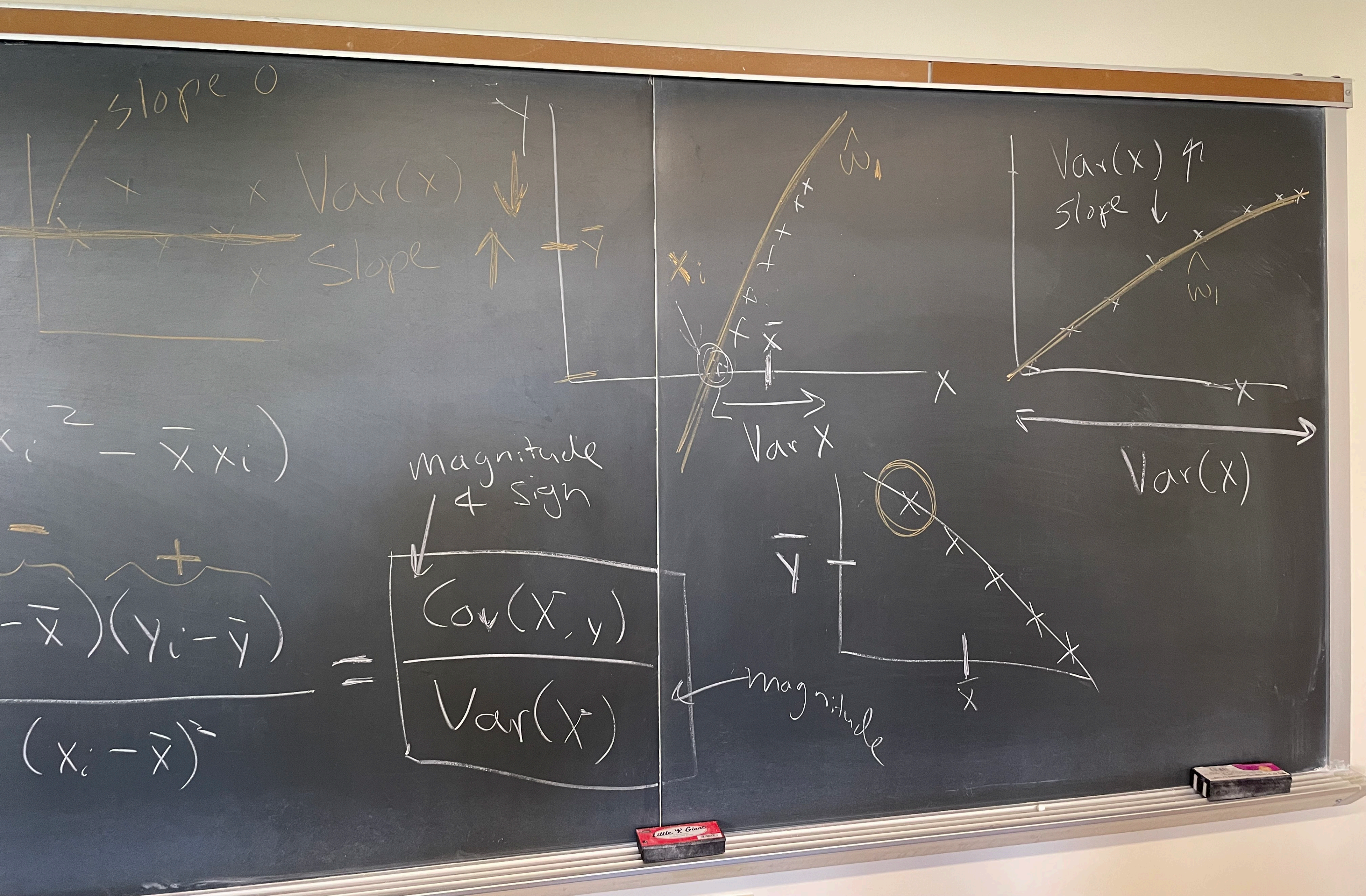
neg slope

add 0

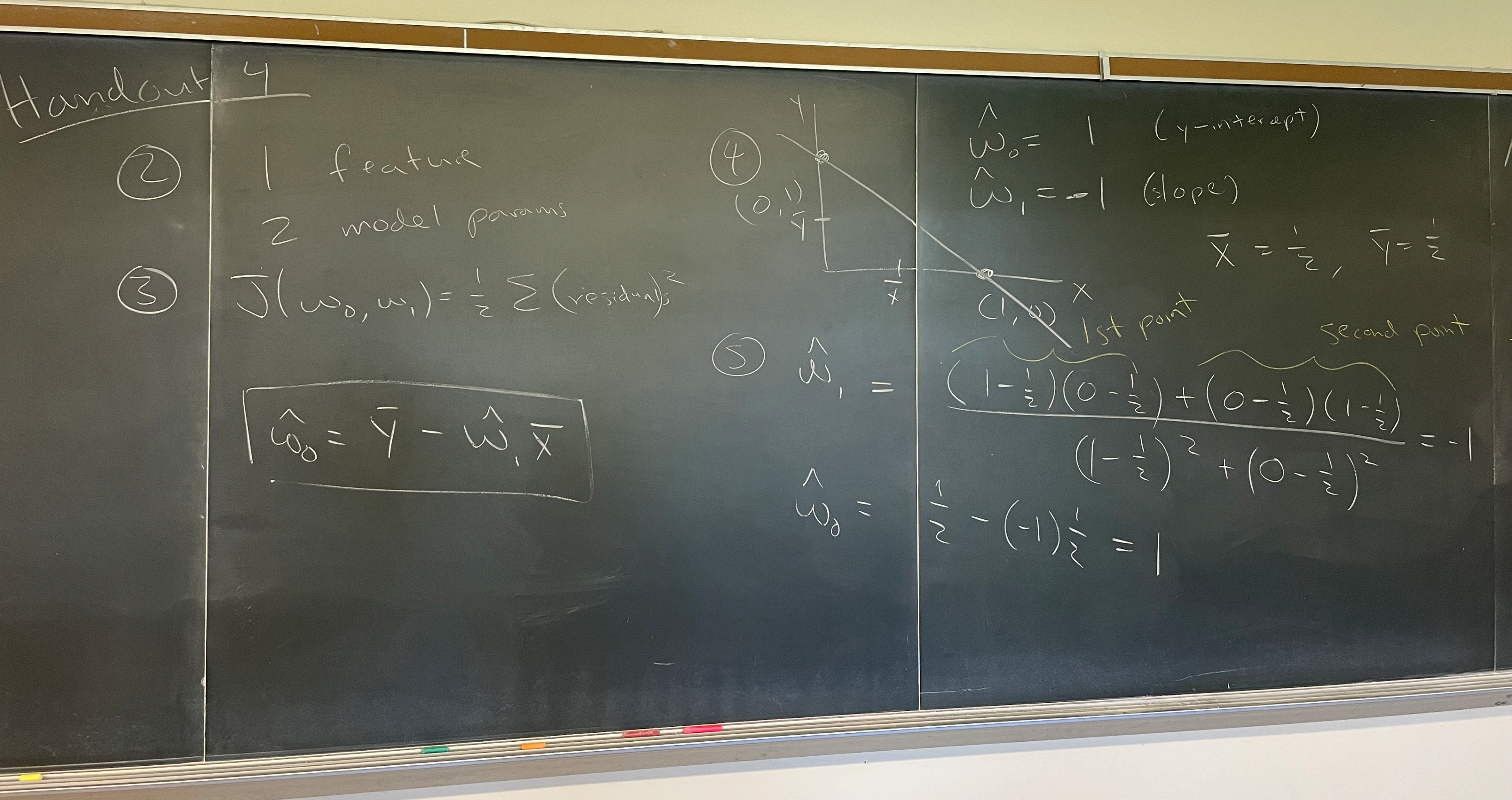
Skip

magnitude & sign




slope 0
Var(x)
Slope
magnitude
& sign
Cov(X, y)
Var(X)
magnitude
Var X
Var(x)
Slope
Var(x)

Handout 4

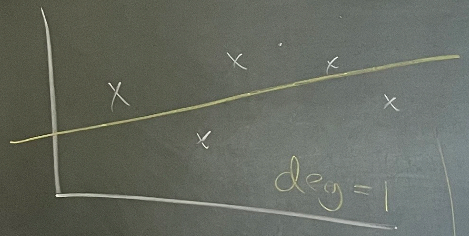

Handout 4
(2) 1 feature
2 model params
(3) $J(w_0, w_1) = \frac{1}{2} \sum (\text{residuals})^2$
$\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}$
(4)
y
(0,1)
$\bar{y}$
$\bar{x}$
x
(1,0)
$\hat{w}_0 = 1$ (y-intercept)
$\hat{w}_1 = -1$ (slope)
$\bar{x} = \frac{1}{2}$, $\bar{y} = \frac{1}{2}$
1st point
second point
(5) $\hat{w}_1 = \frac{(1-\frac{1}{2})(0-\frac{1}{2}) + (0-\frac{1}{2})(1-\frac{1}{2})}{(1-\frac{1}{2})^2 + (0-\frac{1}{2})^2} = -1$
$\hat{w}_0 = \frac{1}{2} - (-1)\frac{1}{2} = 1$

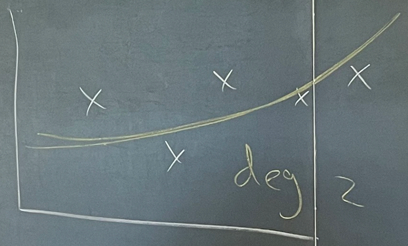
Outline for September 14

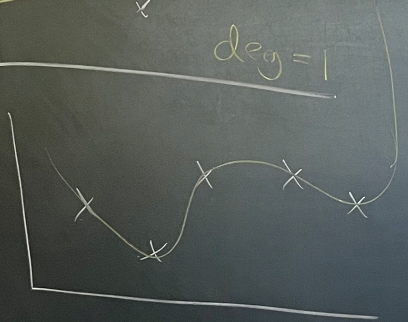
- Why are models useful? (recap)
- Linear models
- Fitting a linear model (one feature)
- Model complexity and evaluation

Model complexity

→ why stop at linear?







n points

n-1 degree will have $J = 0$

Elbow Plot