Applied Machine Learning

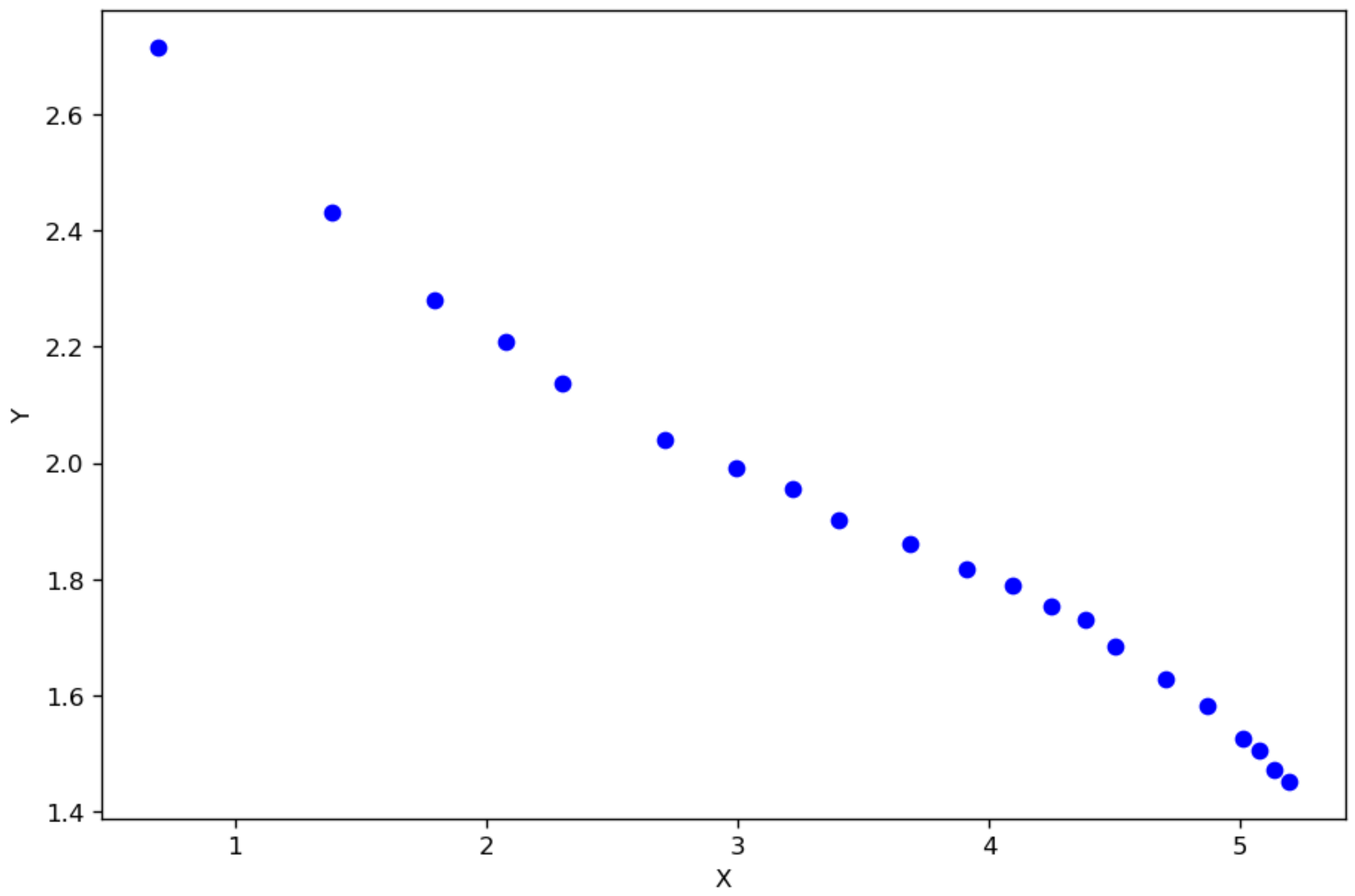
Linear Regression - Performance

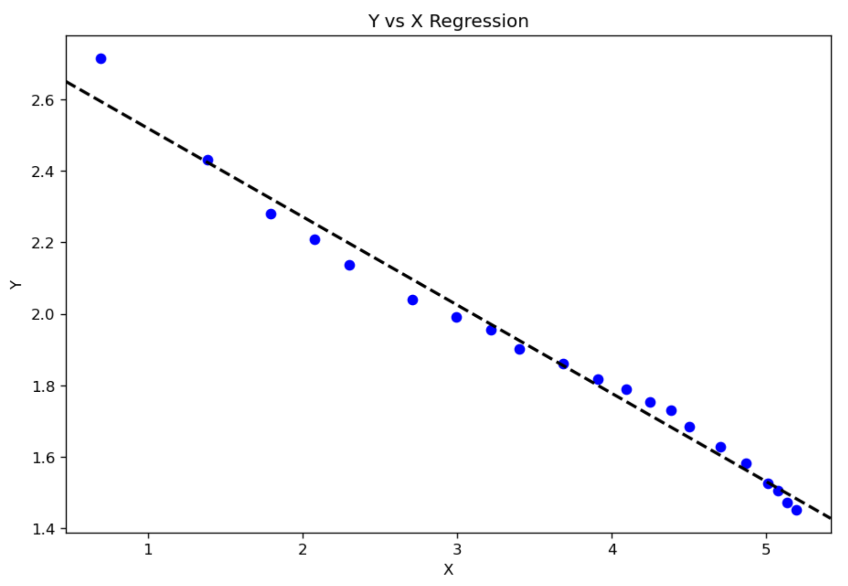
Linear Regression - Performance

- Residuals and Standardized Residuals
- R^2
- Cook's distance
- Outliers

Linear Regression

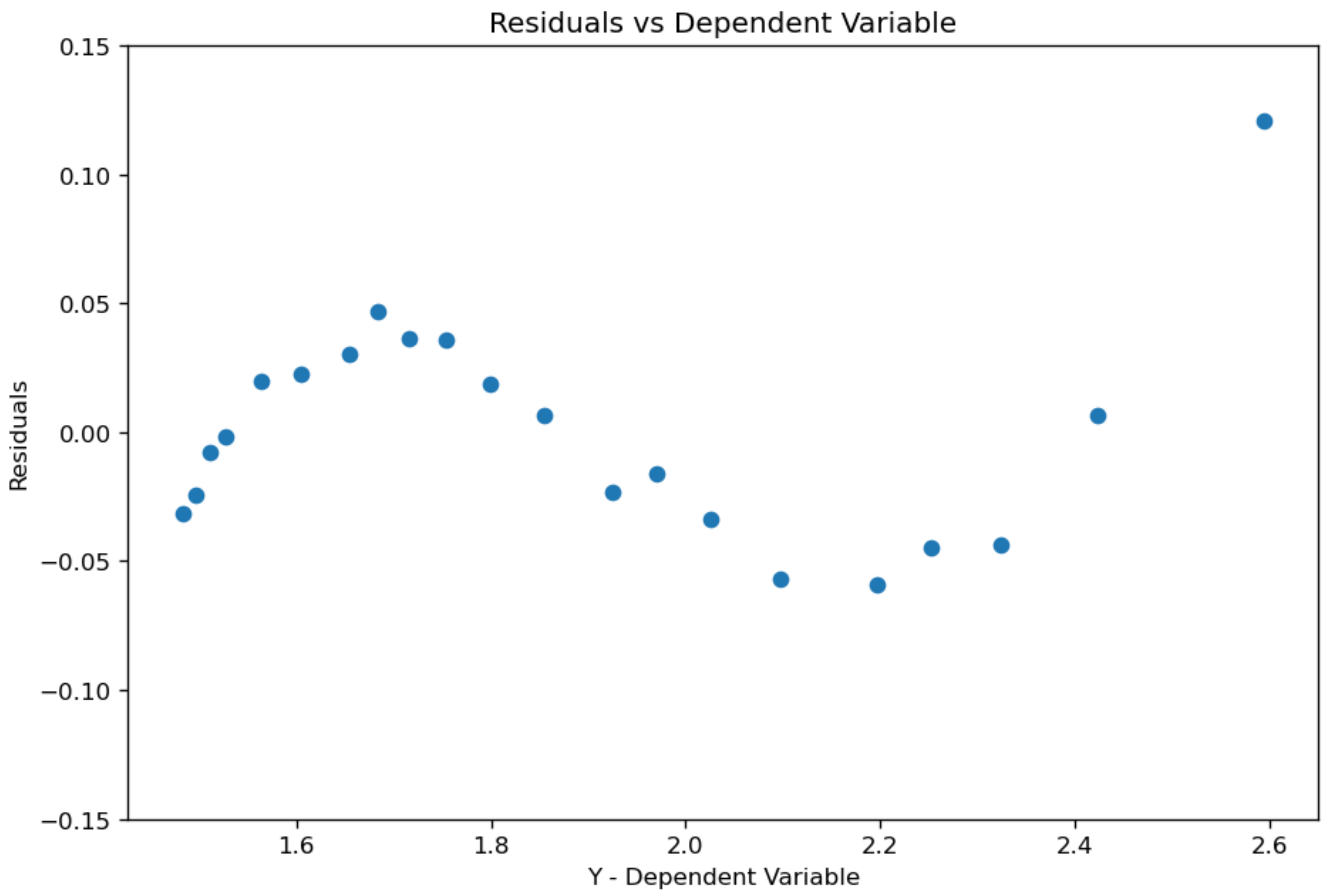
- Linear classifier
  - N pairs of $(\mathbf{x}_i, y_i)$ items
    - $\mathbf{x}_i$: feature vector
    - y_i: numerical value of function evaluated at $\mathbf{x}_i$
- Regressing dependent variable against explanatory variable
  - $y = \mathbf{x}^\top \beta + \xi$
  - $\beta = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$





Residuals

- Linear classifier
  - N pairs of $(\mathbf{x}_i, y_i)$ training set items
  - Find coefficients of linear function $\hat{\beta}$
- Residual
  - $$\begin{aligned} \mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\hat{\beta} \end{aligned}$$
  - and $\mathbf{y}$ measured in the same units
- Mean Square Error of training examples:
  - $m = \frac{\mathbf{e}^\top \mathbf{e}}{N}$

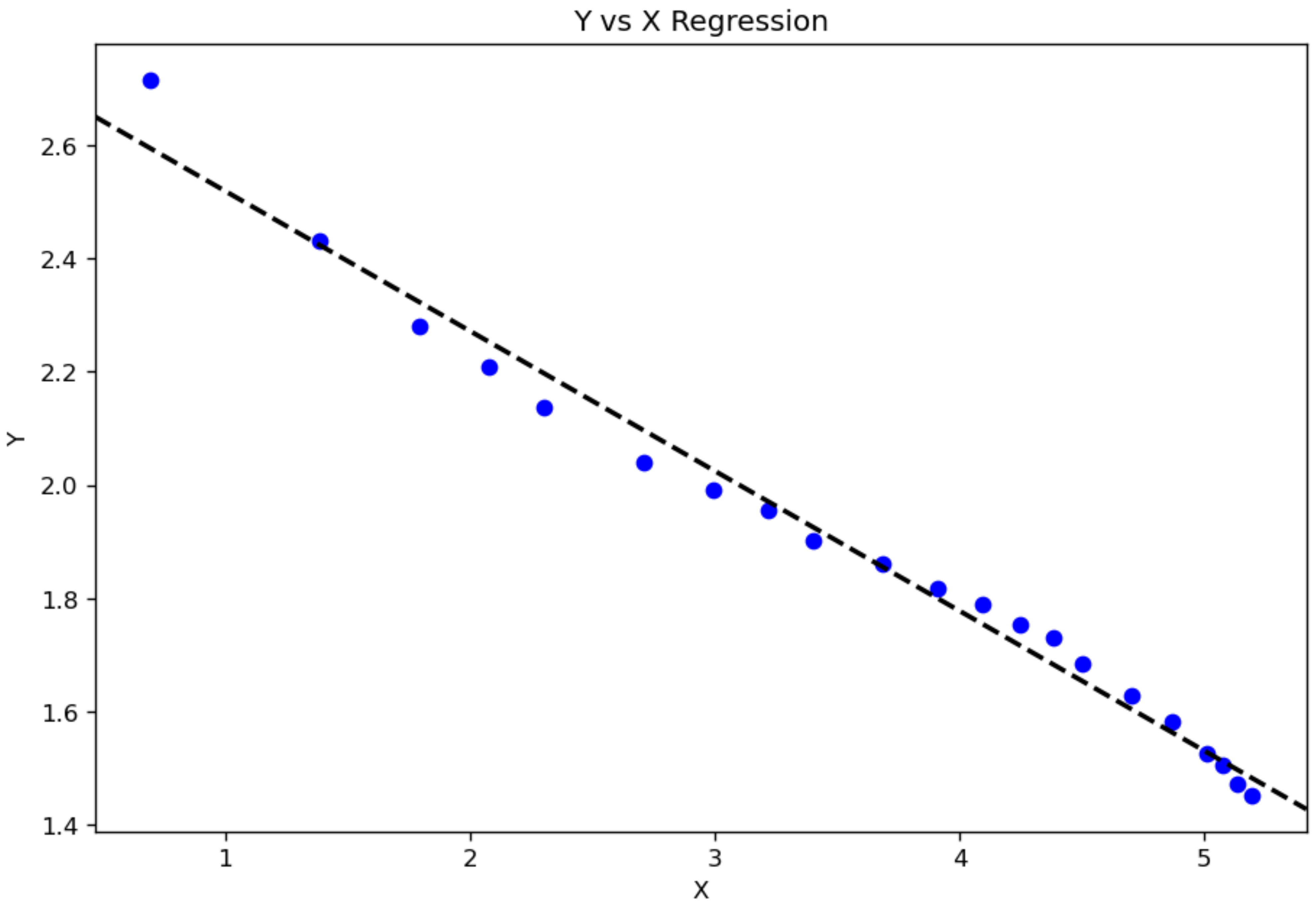


Residuals

- Residual $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta}$
- Some properties of residuals:
  - $\mathbf{e}$ is orthogonal to feature columns in $\mathbf{X}$: $\mathbf{e}^T\mathbf{X} = 0$
  - sum of residuals:
    $$\begin{aligned} \mathbf{e}^\top \mathbf{1} &= 0 \\ \mathbf{1}^\top(\mathbf{y} - \mathbf{X}\hat{\beta}) &= 0 \end{aligned}$$
  - sum of product of individual errors and their corresponding predictions:
    - $\mathbf{e}^T\mathbf{X}\hat{\beta} = 0$

R^2

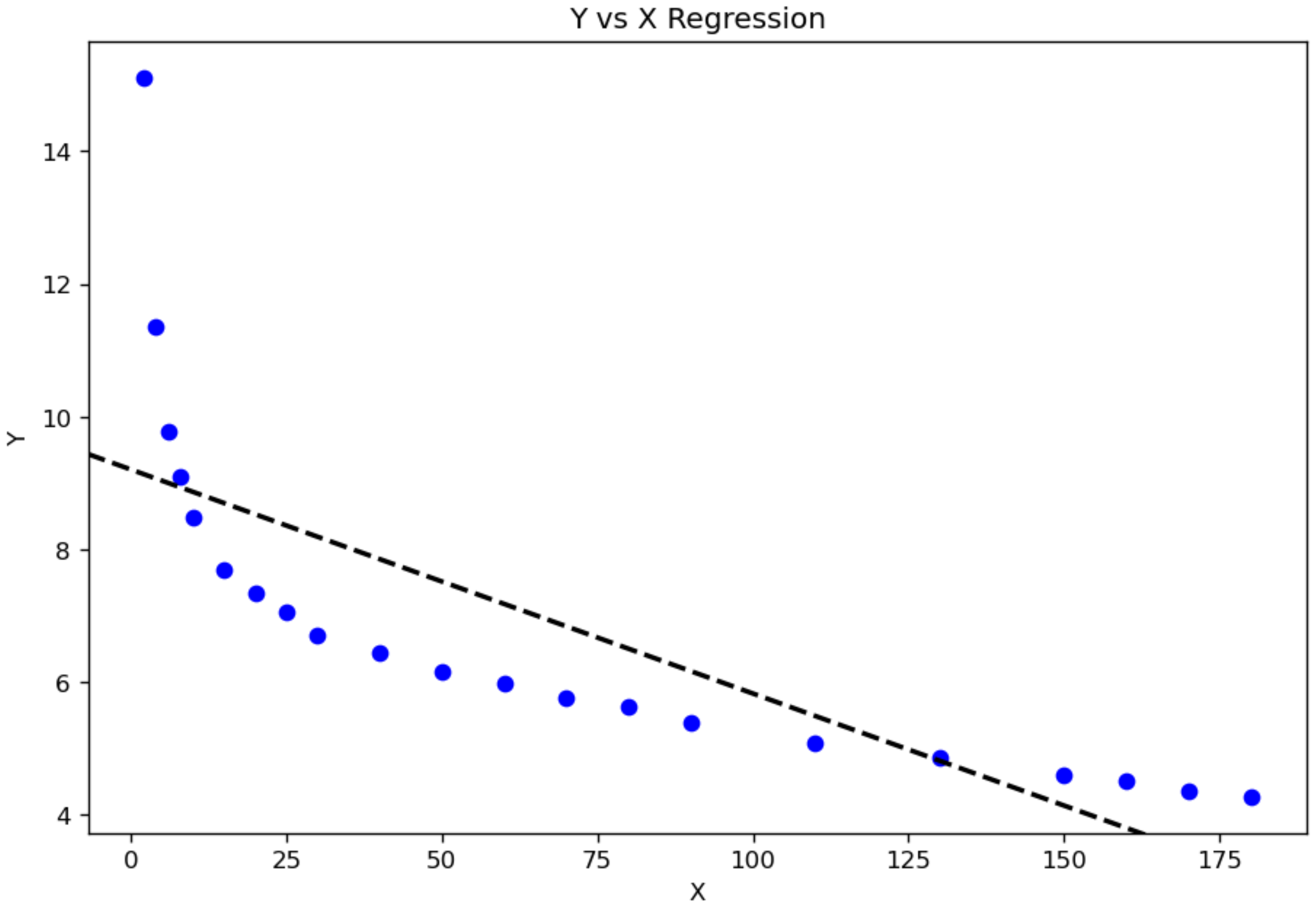
- $\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{e}$
- Applying properties of residuals
  - $\text{var}(\mathbf{y}) = \text{var}(\mathbf{X}\hat{\beta}) + \text{var}(\mathbf{e})$
- $R^2 = \dfrac{\text{var}(\mathbf{X}\hat{\beta})}{\text{var}(\mathbf{y})}$
  - $0 \leq R^2 \leq 1$
  - $R^2 \to 1$: $\text{var}(\mathbf{e}) \to 0$: Good regression
  - $R^2 \to 0$: $\text{var}(\mathbf{e}) \to \text{var}(\mathbf{X}\hat{\beta})$



$R^2 \approx 0.98$

R^2

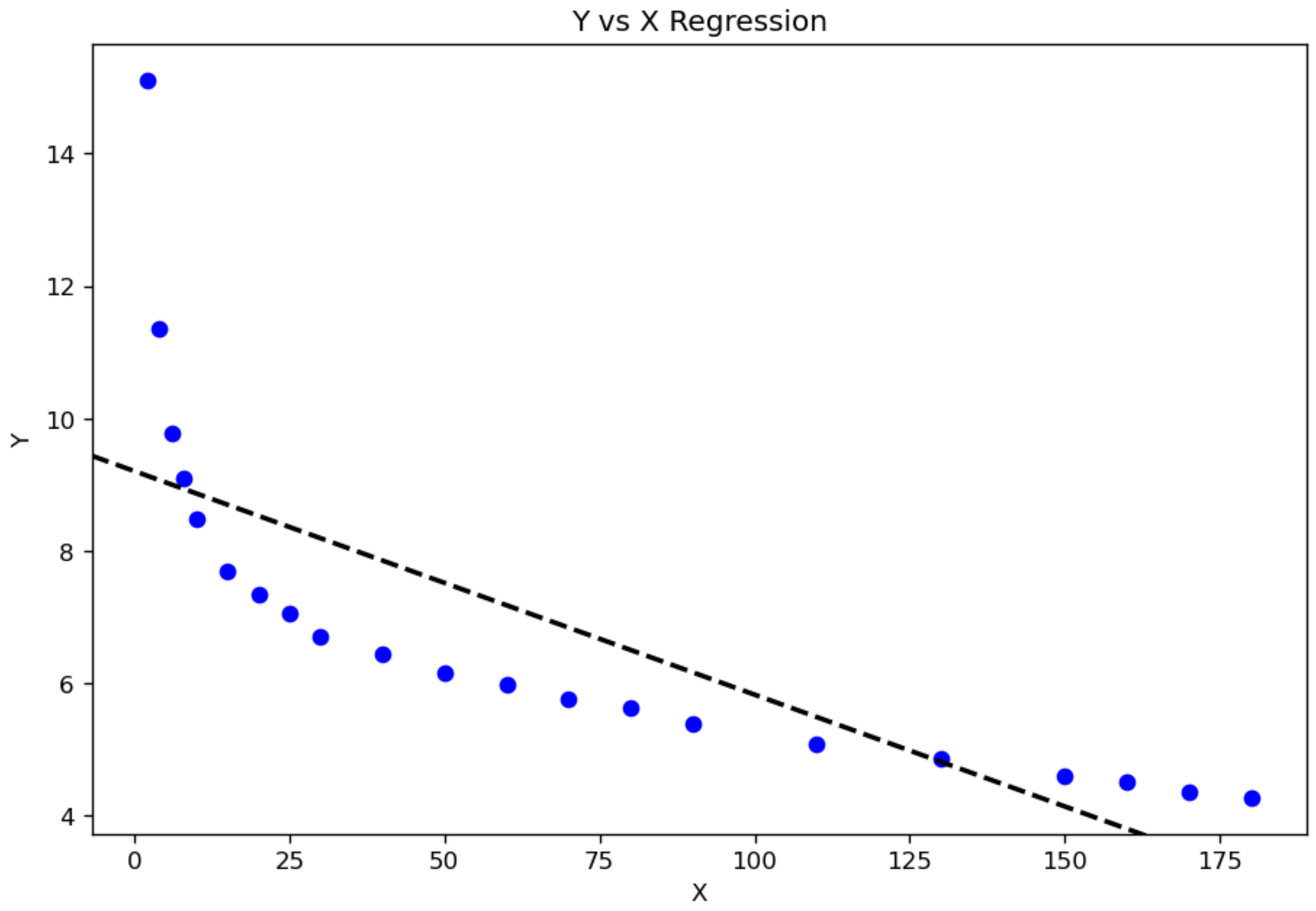
- $\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{e}$
- Applying properties of residuals
  - $\text{var}(\mathbf{y}) = \text{var}(\mathbf{X}\hat{\beta}) + \text{var}(\mathbf{e})$
- $R^2 = \dfrac{\text{var}(\mathbf{X}\hat{\beta})}{\text{var}(\mathbf{y})}$
  - $0 \leq R^2 \leq 1$
  - $R^2 \to 1$: $\text{var}(\mathbf{e}) \to 0$: Good regression
  - $R^2 \to 0$: $\text{var}(\mathbf{e}) \to \text{var}(\mathbf{X}\hat{\beta})$



$R^2 \approx 0.59$

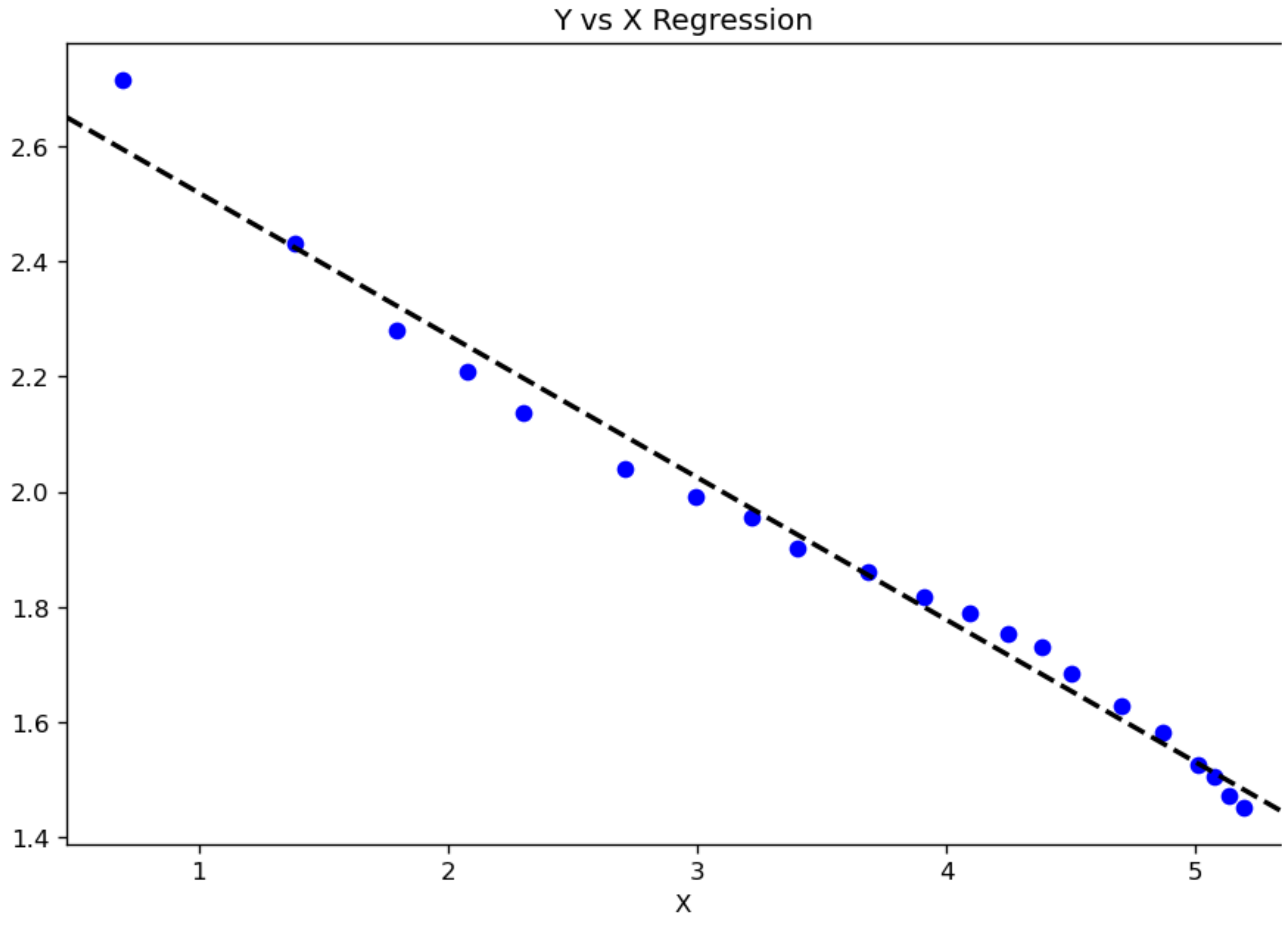
Potential holdups for regressions

- Outliers
- Linear function may not explain the data
- Insufficient number of features



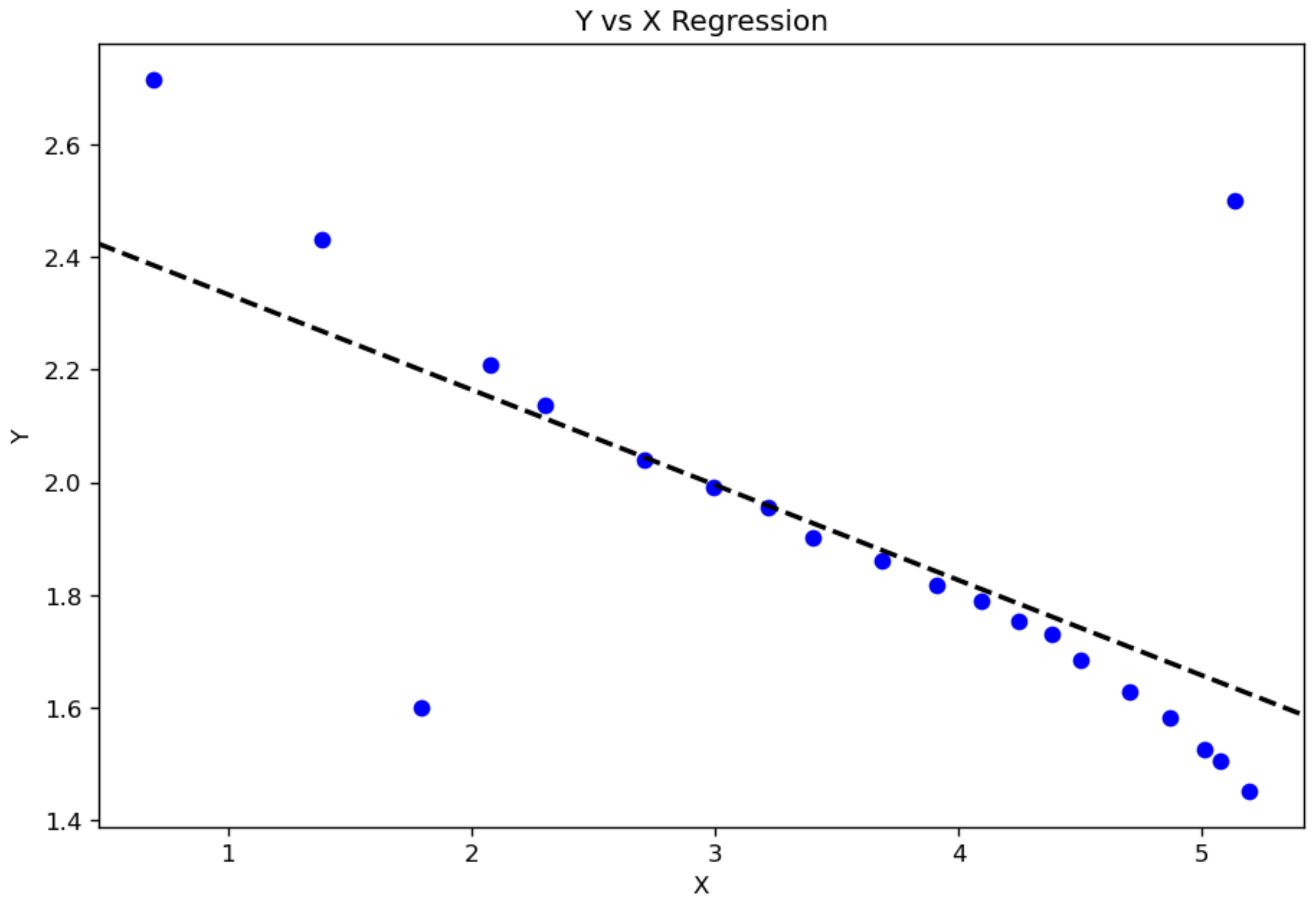
Outliers

- Items that are for from most other items in dataset
  - rare occurrences that do impact the process
  - errors when collecting data
- large residuals with heavy pull on the regression
  - minimization of $\sum_i e^2$



Outliers

- Options
  - keep outliers => poor regression
  - remove outliers
  - discount effect of outliers
  - transform data
- Outlier detection
  - Leverage
  - Changes after removing points



Leverage - Hat Matrix

- Estimated coefficients $\hat{\beta} = (X^\top X)^{-1}(X^\top \mathbf{y})$
- Predictions
$$\begin{aligned} \mathbf{y}^{(p)} &= X\hat{\beta} \\ &= X(X^\top X)^{-1}(X^\top \mathbf{y}) \\ &= (X(X^\top X)^{-1}X^\top)\mathbf{y} \end{aligned}$$
- Hat Matrix $H = X(X^\top X)^{-1}X^\top$ $\mathbf{y}^{(p)} = H\mathbf{y}$
  - elements $h_{i,j}$
  - symmetric with eigenvalues that are either 0 or 1

Leverage - Hat Matrix

- Hat Matrix $H = X(X^\top X)^{-1}X^\top$ $\qquad \mathbf{y}^{(p)} = H\mathbf{y}$

- $\sum_j h_{i,j}^2 \leq 1$

- leverage of training point i: $h_{i,i}$

$$
\begin{aligned}
y_i^{(p)} &= \sum_j h_{i,j} y_j \\
&= h_{i,i} y_i + \sum_{j \neq i} h_{i,j} y_j
\end{aligned}
$$

- dataset items with high leverage have a high pull on the prediction
    - may be outliers

Leverage - Standardized Residuals

- Hat Matrix $H = X(X^\top X)^{-1}X^\top$

$$\sigma_i^2 = m(1 - h_{i,i}) = \frac{\mathbf{e}^\top \mathbf{e}}{N}(1 - h_{i,i})$$

Standardized residual for data item i

$$s_i = \frac{e_i}{\sigma} = \frac{e_i}{\sqrt{\frac{\mathbf{e}^\top \mathbf{e}}{N}(1 - h_{i,i})}}$$

s_i away from normal: data item i may be an outlier

$$\sim 68\,\% \in [-1,1]$$
$$\sim 95\,\% \in [-2,2]$$
$$\sim 99\,\% \in [-3,3]$$

Cook's Distance for data item i

- Coefficients and prediction with full training set
  - $\mathbf{y}^{(p)} \quad = \quad X\hat{\beta}$
- Coefficients and prediction excluding item i from training set
  - $\mathbf{y}_{\hat{i}}^{(p)} = X\hat{\beta}_{\hat{i}}$
- Cook's distance for point i: $$\frac{(\mathbf{y}^{(p)} - \mathbf{y}_{\hat{i}}^{(p)})^{\top}(\mathbf{y}^{(p)} - \mathbf{y}_{\hat{i}}^{(p)})}{dm}$$
  - for a dataset with N items, model with d coefficients, and $m = \frac{\mathbf{e}^T\mathbf{e}}{N}$

Linear Regression - Performance

- Residuals and Standardized Residuals
- R^2
- Cook's distance
- Outliers

Applied Machine Learning

Linear Regression - Performance