SAN JOSÉ STATE UNIVERSITY

Vector Data - Prediction

1

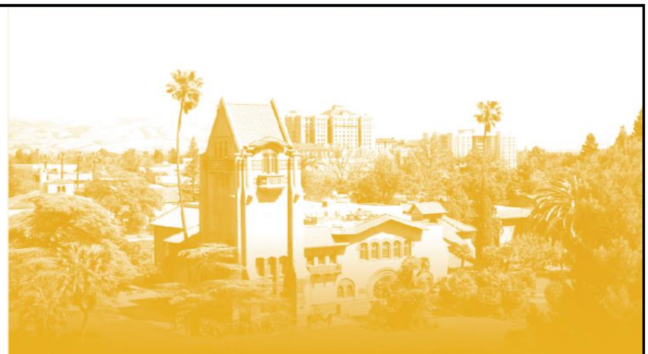
SAN JOSÉ STATE UNIVERSITY

Agenda

- Prediction with Linear Regression
- Model Evaluation & Selection
- Regularization

2

2



SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regression

3

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regression

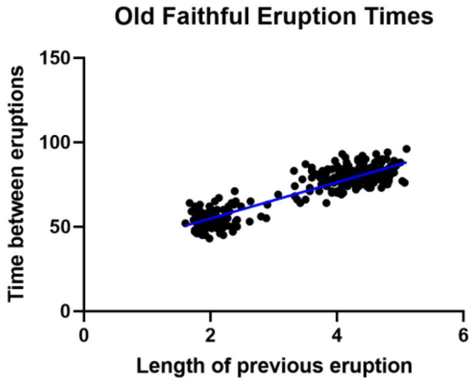
- Ordinary Least Square Regression
  - Closed form solution
- Gradient Descent
- Linear Regression with Probabilistic Interpretation

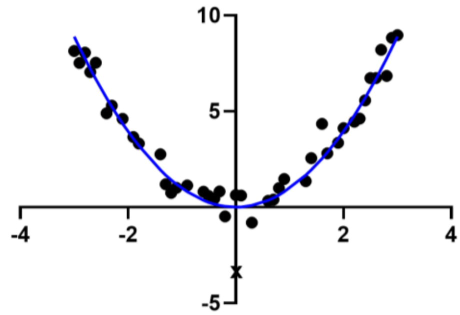
4

4

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regression Problems?





5

5

SJSU SAN JOSÉ STATE UNIVERSITY

The Linear Regression Problem

$$\mathbf{y} = \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \beta_3 \mathbf{x_3}$$

Dependent Variable
Outcome Variable
Response Variable

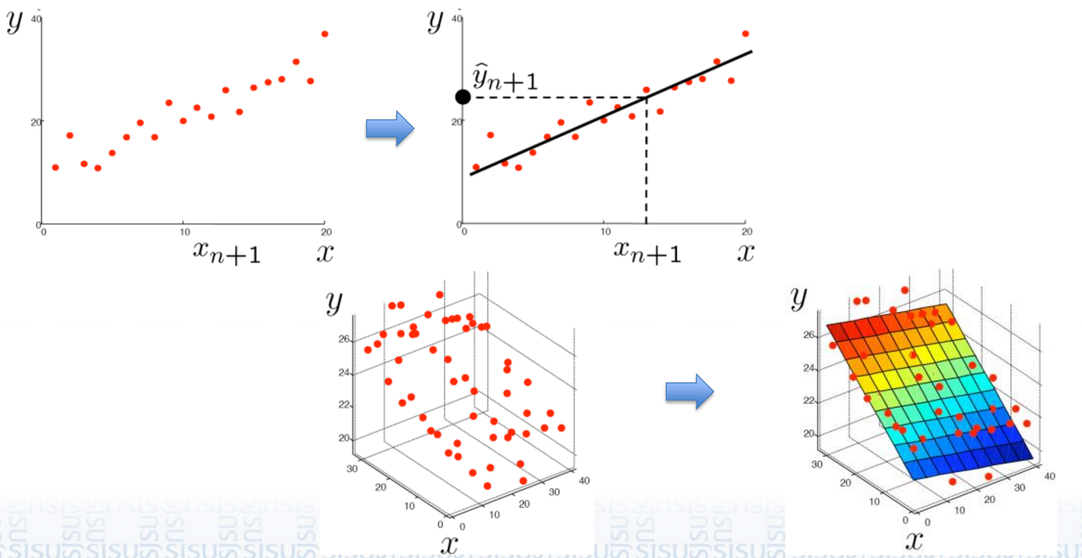
Independent Variable
Predictor Variable
Explanatory Variable

6

6

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regression Examples



7

SJSU SAN JOSÉ STATE UNIVERSITY

Why Linear Regression?

- Suppose we want to model the dependent variable Y in terms of three variables, x_1, x_2, x_3

$$\mathbf{y = f(x_1, x_2, x_3)}$$

- Typically, we won't have sufficient data to estimate **$f(x_1, x_2, x_3)$**
- Therefore, we usually have to assume that it has some restricted form, such as linear:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

8

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regression Example

Living Area (sqft)	# of Beds	Has pool	Price (1000$)
2104	3	Yes	400
1600	3	No	330
2400	3	No	369
1416	2	No	232
3000	4	Yes	540

x_1, x_2, x_3 y

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

9

9

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regression: Advantages

- Simplicity & Interpretability - easy to understand & clear insights coefficients
- Efficiency - computationally less intensive compared to more complex models, making it suitable for large datasets
- Predictive Power – good for prediction as well as baseline
- Statistical Significance – hypothesis testing and confidence intervals
- Flexibility – can be extended to multiple linear regression as well as including regularization
- Diagnostics – residual analysis and goodness of fit

10

10

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regressions: Assumptions

Linear regression relies on several key assumptions:

- Linearity: The relationship between the dependent and independent variables is linear
- Independence: The residuals (errors) are independent
- Homoscedasticity: The residuals have constant variance at every level of (x)
- Normality: The residuals of the model are normally distributed

11

11

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regressions: Limitations

- Linearity Assumption – model won't capture true pattern of data if relationship not linear
- Sensitivity to Outliers – highly sensitive to outliers ➔ affect slope of regression line
- Multicollinearity – can inflate the variance of coefficient estimates if independent variables are highly correlated
- Independence of Errors – assume error terms are independent of each other
- Overfitting & Underfitting – potentially high bias and variance
- Limited Explanatory Power – only explains the relationship between the mean of the dependent and independent variable. Doesn't capture full distribution of depenent variable.

12

12

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regression

- **Data**: n independent data points $\boldsymbol{x}_i$, y_i $i = 1..n$
  - y_i dependent variable
  - $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ independent or explanatory variables
- **Model**:
  - For any data point $(\boldsymbol{x}, y)$
    - Shared weight vector: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$
    - Predicted outcome: $y = \boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$
  - For convenience, can include bias term β_0 into $\boldsymbol{\beta}$
    - $\boldsymbol{x} = (1, x_1, x_2, \ldots, x_p)^T$
    - $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$
    - $y = \boldsymbol{x}^T\boldsymbol{\beta}$

13

13

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regression Process

- Model Construction
  - Use **training data** to find the best parameter $\boldsymbol{\beta}$, denoted as $\hat{\boldsymbol{\beta}}$
- Model Selection
  - Use **validation data** to select the best model

    e.g. feature selection
- Model Usage
  - Apply the model to the unseen data (test data): $\hat{y}_{new} = \boldsymbol{x}_{new}^T\hat{\boldsymbol{\beta}}$

14

14

SJSU SAN JOSÉ STATE UNIVERSITY

Least Square Estimation

- Cost function (Mean Square Error):

$$J(\boldsymbol{\beta}) = \frac{1}{2n} \sum_i (\boldsymbol{x}_i^T \boldsymbol{\beta} - y_i)^2$$

- Matrix form: $J(\boldsymbol{\beta}) = \frac{1}{2n}(\boldsymbol{X\beta} - \boldsymbol{y})^T(\boldsymbol{X\beta} - \boldsymbol{y}) = \frac{1}{2n}\|\boldsymbol{X\beta} - \boldsymbol{y}\|$

$$\begin{bmatrix} 1, x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1, x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1, x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \qquad \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

$\boldsymbol{X}: \boldsymbol{n} \times (\boldsymbol{p}+\boldsymbol{1})$ matrix $\qquad$ y: $\boldsymbol{n} \times \boldsymbol{1}$ vector

SJSU SAN JOSÉ STATE UNIVERSITY

Ordinary Least Squares (OLS)

- Goal: Find $\boldsymbol{\beta}$ that minimizes $J(\boldsymbol{\beta})$:

$$J(\boldsymbol{\beta}) = \frac{1}{2n}(\boldsymbol{X\beta} - \boldsymbol{y})^T(\boldsymbol{X\beta} - \boldsymbol{y}) = \frac{1}{2n}(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta} - \boldsymbol{y}^T\boldsymbol{X\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y})$$

- Ordinary least squares: set first derivative of $J(\boldsymbol{\beta}) = \boldsymbol{0}$:

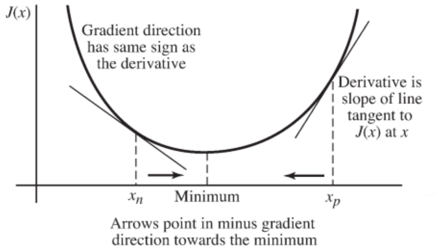
$$\frac{\partial J}{\partial \boldsymbol{\beta}} = \frac{1}{n}(\boldsymbol{X}^T\boldsymbol{X\beta} - \boldsymbol{X}^T\boldsymbol{y}) = \boldsymbol{0} \quad \Rightarrow \quad \boxed{\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}$$

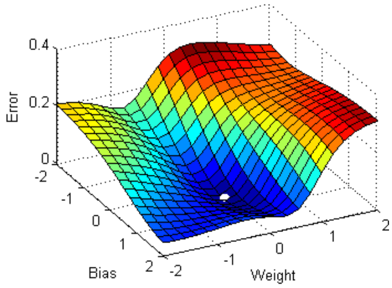
- What if $(\boldsymbol{X}^T\boldsymbol{X})$ is not invertible?

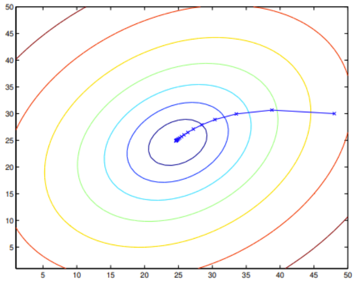
SJSU SAN JOSÉ STATE UNIVERSITY

Gradient Descent

- Minimize the cost function by moving down in the steepest direction.

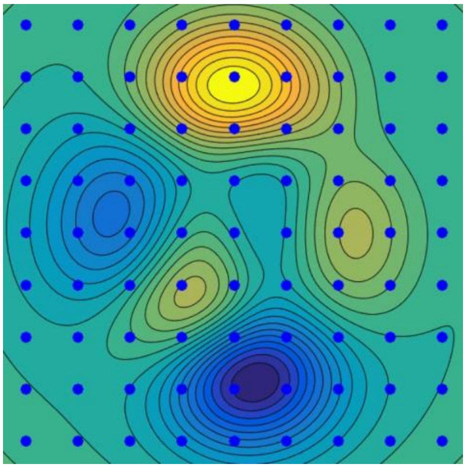






SJSU SAN JOSÉ STATE UNIVERSITY

Gradient Descent in 2D Illustration



SJSU SAN JOSÉ STATE UNIVERSITY

Batch Gradient Descent

- Move in the direction of **steepest** descent (all data points considered)

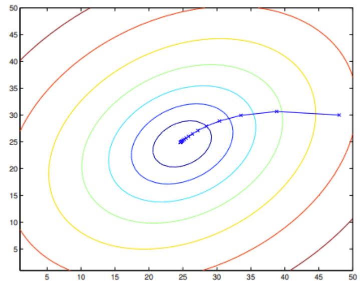
Repeat until converge {

$$\boldsymbol{\beta}^{t+1} := \boldsymbol{\beta}^{t} - \eta \left.\frac{\partial J}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{t}}$$

}

$$J(\boldsymbol{\beta}) = \frac{1}{2n}(\boldsymbol{X\beta} - \boldsymbol{y})^{T}(\boldsymbol{X\beta} - \boldsymbol{y})$$

$$\frac{\partial J}{\partial \boldsymbol{\beta}} = \frac{1}{n}(\boldsymbol{X}^{T}\boldsymbol{X\beta} - \boldsymbol{X}^{T}\boldsymbol{y}) = \boldsymbol{0}$$



SJSU SAN JOSÉ STATE UNIVERSITY

Stochastic (Incremental) Gradient Descent

- When a new observation, i, comes in, update weights $\boldsymbol{\beta}$ immediately (extremely useful for large-scale datasets):

Repeat {

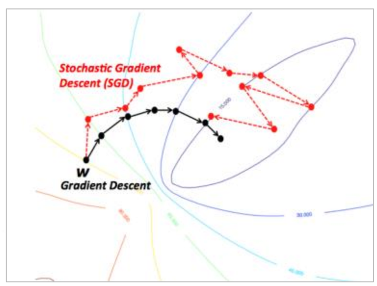
for i =1 to n {

$$\boldsymbol{\beta}^{t+1} := \boldsymbol{\beta}^{t} + \eta\left(y_i - \boldsymbol{x}_i^{T}\boldsymbol{\beta}^{t}\right)\boldsymbol{x}_i$$

}

}

If the prediction for object i is smaller than the real value, $\boldsymbol{\beta}$ should move forward to the direction of $\boldsymbol{x}_i$

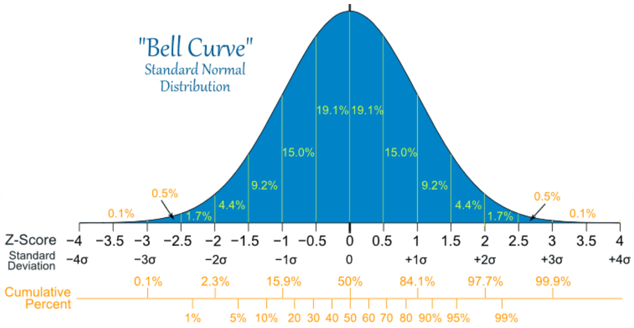


SJSU SAN JOSÉ STATE UNIVERSITY

Probabilistic Interpretation

- Recall that for normal distribution

$$X \sim N(\mu, \sigma^2) \Rightarrow f(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



SJSU SAN JOSÉ STATE UNIVERSITY

Probabilistic Interpretation

- Model: $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$
  - $\varepsilon_i \sim N(0, \sigma^2)$
  - $y_i | \boldsymbol{x}_i, \beta \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$
  - $E[y_i|\boldsymbol{x}_i] = \boldsymbol{x}_i^T\boldsymbol{\beta}$

$$p(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)$$

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})}{2\sigma^2}\right)$$

- Likelihood:

$$L(\boldsymbol{\beta}) = \prod_i p(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})}{2\sigma^2}\right)$$

- Maximum Likelihood Estimation (MLE)
  - find $\widehat{\boldsymbol{\beta}}$ that maximizes $L(\boldsymbol{\beta})$
  - arg max $L(\boldsymbol{\beta})$ = arg min $J(\boldsymbol{\beta})$ ← equivalent to OLS!

SJSU SAN JOSÉ STATE UNIVERSITY

Other Practical Issues

- Handle different scales of numerical attributes
  - Standardization with Z-score: $z = \frac{x-\mu}{\sigma}$
  - x (raw score to be standardized), μ (mean of the population), σ (standard deviation)
- What if some attributes are nominal?
  - Binary values – convert to a binary number/boolean
    e.g. x = 1, if gender = F; x = 0, if gender = M
  - Nominal variable with multiple values?
    - Create more dummy variables for one variable
- What if some attributes are ordinal?
  - replace x_{if} by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing i^{th} object in the f^{th} variable by $z_{if} = \frac{r_{if}-1}{M_f-1}$

23

23

SJSU SAN JOSÉ STATE UNIVERSITY

Other Practical Issues

- What if $X^T X$ is not invertible?
  - Add a small portion of identity matrix, λI, to it
    - ridge regression or linear regression with L_2 norm regularization

$$\sum_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

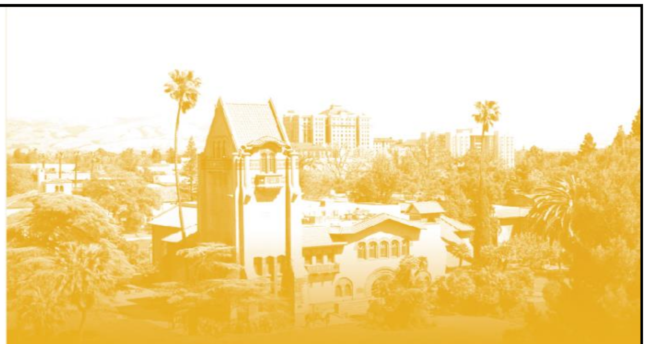
- What if non-linear correlation exists?
  - Transform features, say, x to x^2

24

24

SJSU SAN JOSÉ STATE UNIVERSITY

Model Evaluation & Selection



25

SJSU SAN JOSÉ STATE UNIVERSITY

Model Evaluation

- Mean Squared Error (MSE)

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

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

- (square) Root of the Mean of the Squared Errors (RMSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$$

26

26

SJSU SAN JOSÉ STATE UNIVERSITY

Model Selection Problem

- Basic problem:
  - how to choose between competing linear regression models
- Model too simple:
  - "underfit" the data; poor predictions; high bias; low variance
- Model too complex:
  - "overfit" the data; poor predictions; low bias; high variance
- Model just right:
  - balance bias and variance to get good predictions

SJSU SAN JOSÉ STATE UNIVERSITY

Bias

Bias refers to the error introduced by approximating a complex problem with a simplified model.

  - High Bias: Models with high bias are often too simple and do not capture the underlying patterns of the data well. (underfitting ➔ performs poorly on both training and test data)
  - Low Bias: Models with low bias make fewer assumptions about the data, allowing them to capture more complex patterns.

- Example: A linear regression model applied to a non-linear dataset will have high bias because it cannot capture the non-linear relationship.

SJSU SAN JOSÉ STATE UNIVERSITY

Variance

Variance refers to the error introduced by the model's sensitivity to small fluctuations in the training data. It measures how much the model's predictions would change if it were trained on a different dataset.

  - High Variance: Models with high variance are highly sensitive to the specific training data they were trained on. (overfitting ➔ performs well on training data but poorly on test data)
  - Low Variance: Models with low variance are more stable and less sensitive to changes in the training data.

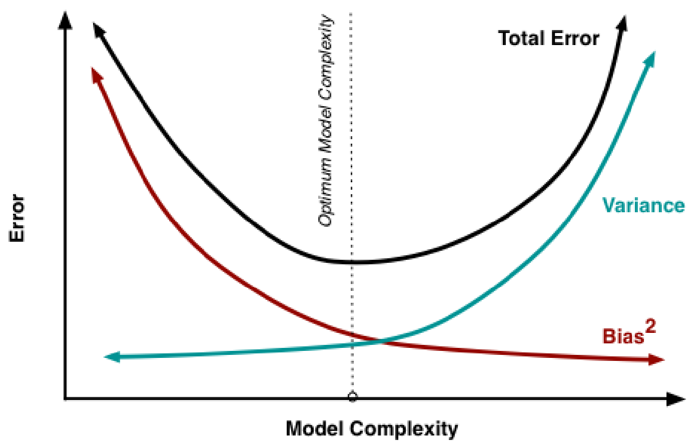
- Example: A decision tree with many branches can have high variance because it fits the training data very closely, including noise.

29

29

SJSU SAN JOSÉ STATE UNIVERSITY

Bias-Variance Trade-off

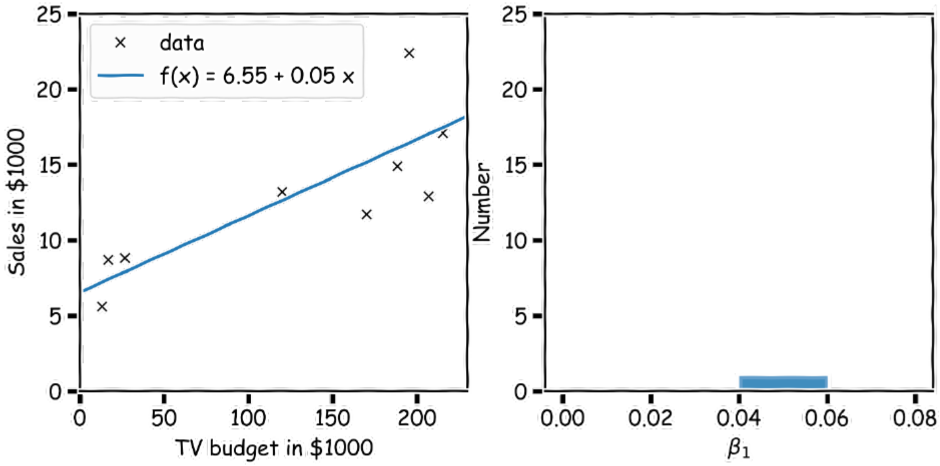


30

30

SJSU SAN JOSÉ STATE UNIVERSITY

Illustration



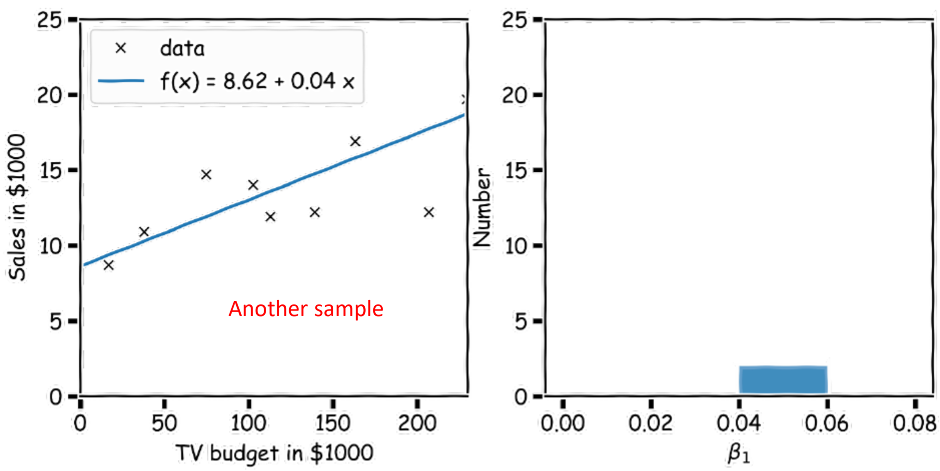
Our training data is only one possible sample

31

31

SJSU SAN JOSÉ STATE UNIVERSITY

Illustration

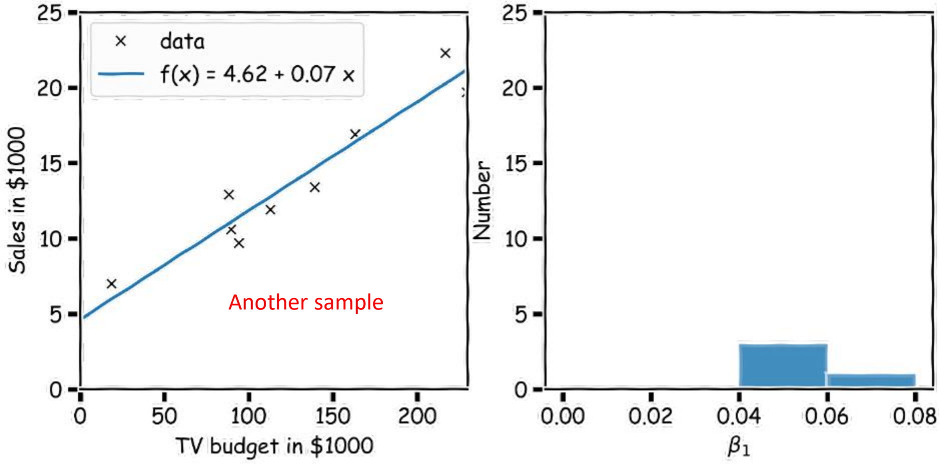


32

32

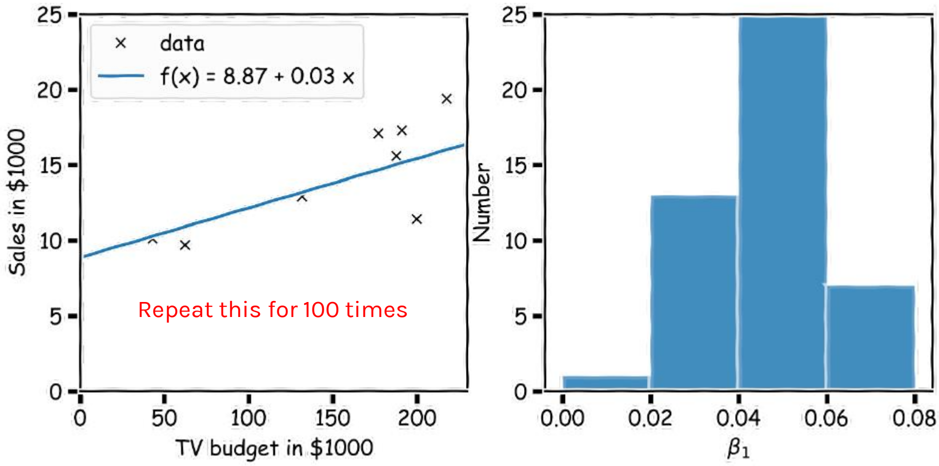
SJSU SAN JOSÉ STATE UNIVERSITY

Illustration



SJSU SAN JOSÉ STATE UNIVERSITY

Illustration



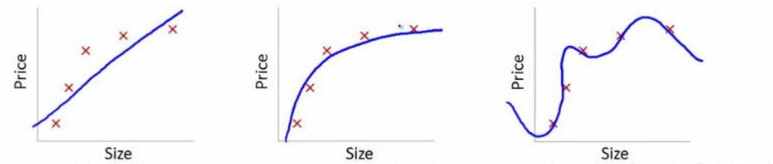
SJSU SAN JOSÉ STATE UNIVERSITY

Higher Degrees in Regression

1. $h_\theta(x) = \theta_0 + \theta_1 x$
2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$
3. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3$

$\vdots$

10. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}$



$\theta_0 + \theta_1 x$	$\theta_0 + \theta_1 x + \theta_2 x^2$	$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$
High bias (underfit)	"Just right"	High variance (overfit)

35

35

SJSU SAN JOSÉ STATE UNIVERSITY

Regularization

36

SJSU SAN JOSÉ STATE UNIVERSITY

L_1 or Lasso Regularization

- Lasso Regularization: adds the absolute value of the coefficients to the loss function:

$$J(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i}\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i\right)^2 + \lambda\sum_{i}^{p}|\beta_i|$$

- Encourages sparsity ➔ shrink some coefficients to exactly zero, effectively performing feature selection.
- Useful when you have a large number of features and suspect only a few are important.
- The optimization problem is less smooth due to the absolute values, which can make it harder to solve.

37

37

SJSU SAN JOSÉ STATE UNIVERSITY

L_2 or Ridge Regularization

- Ridge Regularization: adds the squared value of the coefficients to the loss function:

$$J(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i}\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i\right)^2 + \lambda\sum_{i}^{p}\beta_i^2$$

- Shrinks coefficients uniformly but does not set them to zero. It helps in reducing the impact of collinear features.
- Useful when you have many features that are all potentially useful, and want to reduce the model complexity without eliminating any features.
- The optimization problem is smoother due to the squared terms, making it easier to solve.

38

38

SJSU SAN JOSÉ STATE UNIVERSITY

Elastic Net (L_1 + L_2)

- Combines both L_1 and L_2 regularization:

$$J(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i}\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i\right)^2 + \lambda_1 \sum_{i}^{p}|\beta_i| + \lambda_2 \sum_{i}^{p}\beta_i^2$$

- Balances the benefit of both approaches ➔ more flexible regularization

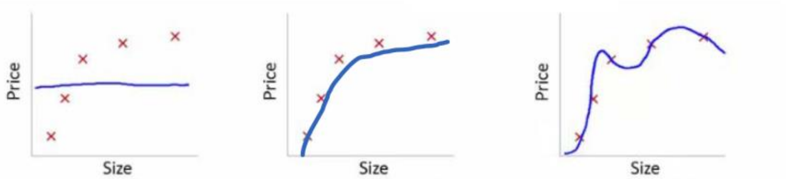
39

39

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Regression with Regularization

- Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$
  $J(\theta) = \frac{1}{2n}\sum_i (h_\theta(x_i) - y_i)^2 + \lambda \sum_j \theta_j^2$



Large λ
High bias (underfit)
$\lambda = 10000.\ \theta_1 \approx 0, \theta_2 \approx 0, \ldots$
$h_\theta(x) \approx \theta_0$

Intermediate λ
"Just right"

Small λ
High variance (overfit)
$\lambda \approx 0$

40

40

SJSU SAN JOSÉ STATE UNIVERSITY

Factors to Consider for Choosing Regularization

- Feature Selection
  - L_1 produces sparse models (few features). L_2 includes all features but shrinks their impact and are more stable and reliable when all features are important.
- Multicollinearity
  - L_2 is more effective for highly correlated features
- Model Interpretability
  - L_1 produces sparse models that are easier to interpret
- Computational Efficiency
  - L_1 is more intense computationally while L_2 is more efficient

41

41

SJSU SAN JOSÉ STATE UNIVERSITY

Which Regularization to Use?

- Financial Data Modeling
- Image Recognition
- Sports Analytics
- Robotics
- Healthcare
- Natural Language Processing

42

42

SJSU SAN JOSÉ STATE UNIVERSITY

Summary

- Linear Regression
  - Ordinary Linear Regression
  - Probabilistic interpretation → same as MLE
- Model Evaluation and Selection
  - Bias-Variance Trade-off
  - Error metrics
- Regularization

43