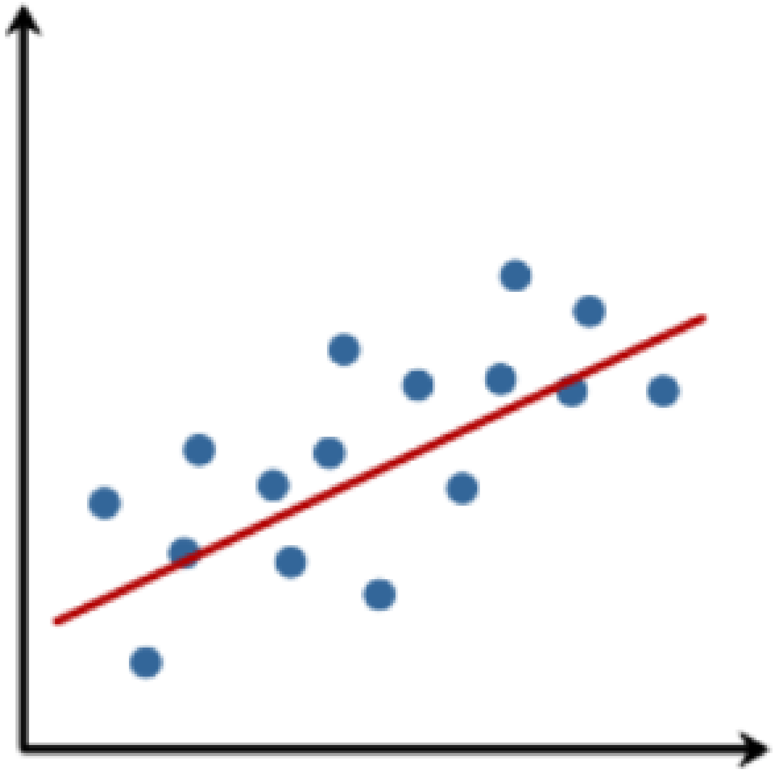
Regression

Koko Friansa
6-12-2024
(Sekolah Bisnis Manajemen ITB)

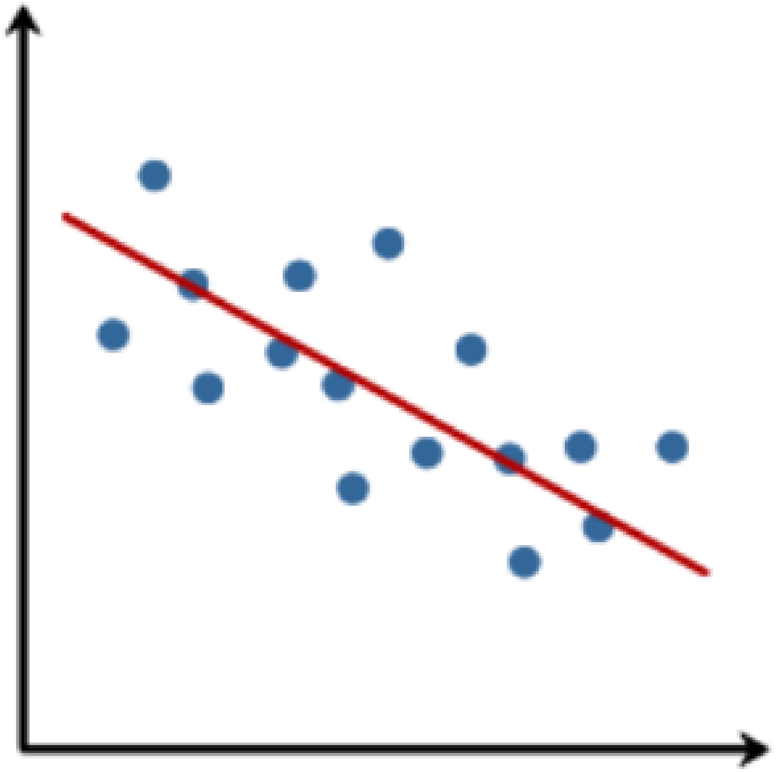
Introduction

- Regression seeks the best relationship between the independent variable (regressor) X and the dependent variable (response) Y, determines the strength of that relationship, and predicts the value of the response Y based on the regressor X.
- Simple linear regression applies only to cases with one regressor variable and assumes a linear relationship between X and Y.
- The relationship between variables is not deterministic (i.e., not exact). There is a random component in the equation.

Relation Type

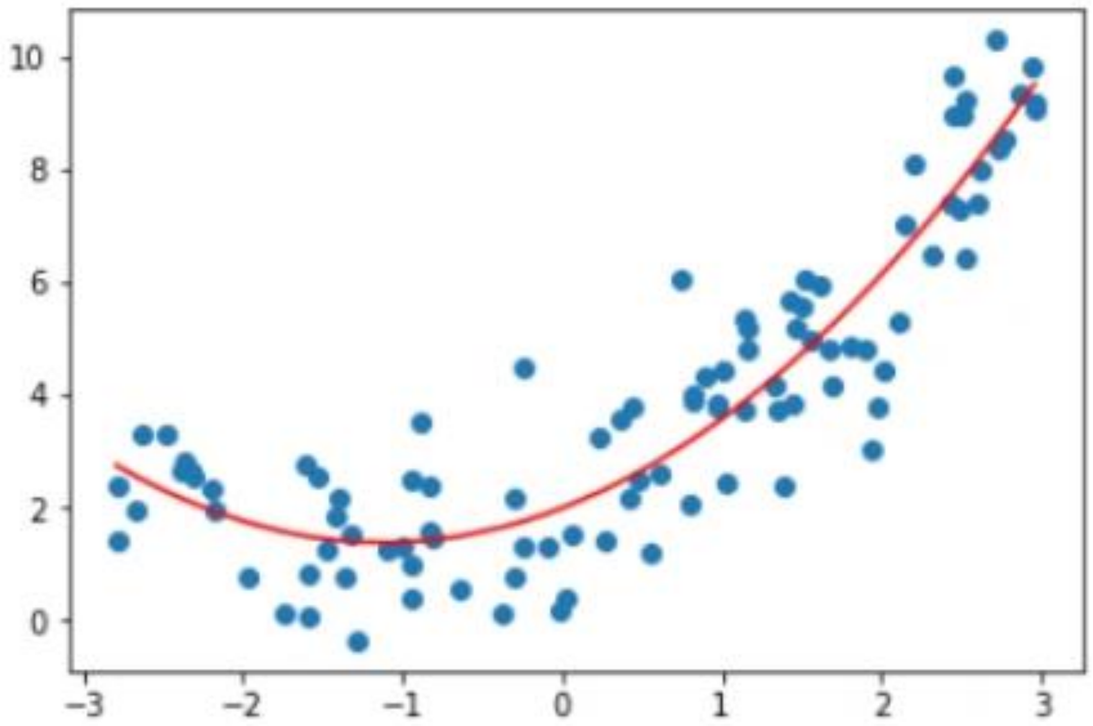


Positive Linear Relationship

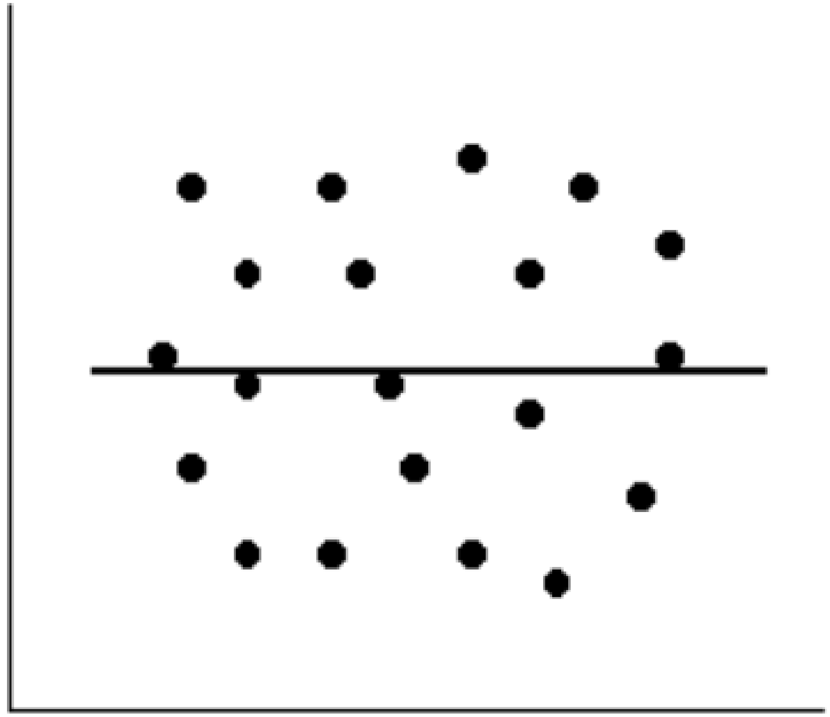


Negative Linear Relationship

Relation Type



Non-Linear Relationship



No Relationship

Strength of the Relationship

Correlation measures the strength of the relationship between two variables.

- **Correlation coefficient (r):**
  - The closer it is to -1, the stronger the negative relationship. If the value of one variable increases, the value of the other tends to decrease.
  - The closer it is to 1, the stronger the positive relationship. If the value of one variable increases, the value of the other also tends to increase.
  - The closer it is to 0, the weaker the relationship.

Strength of the Relationship

Correlation measures the strength of the relationship between two variables.

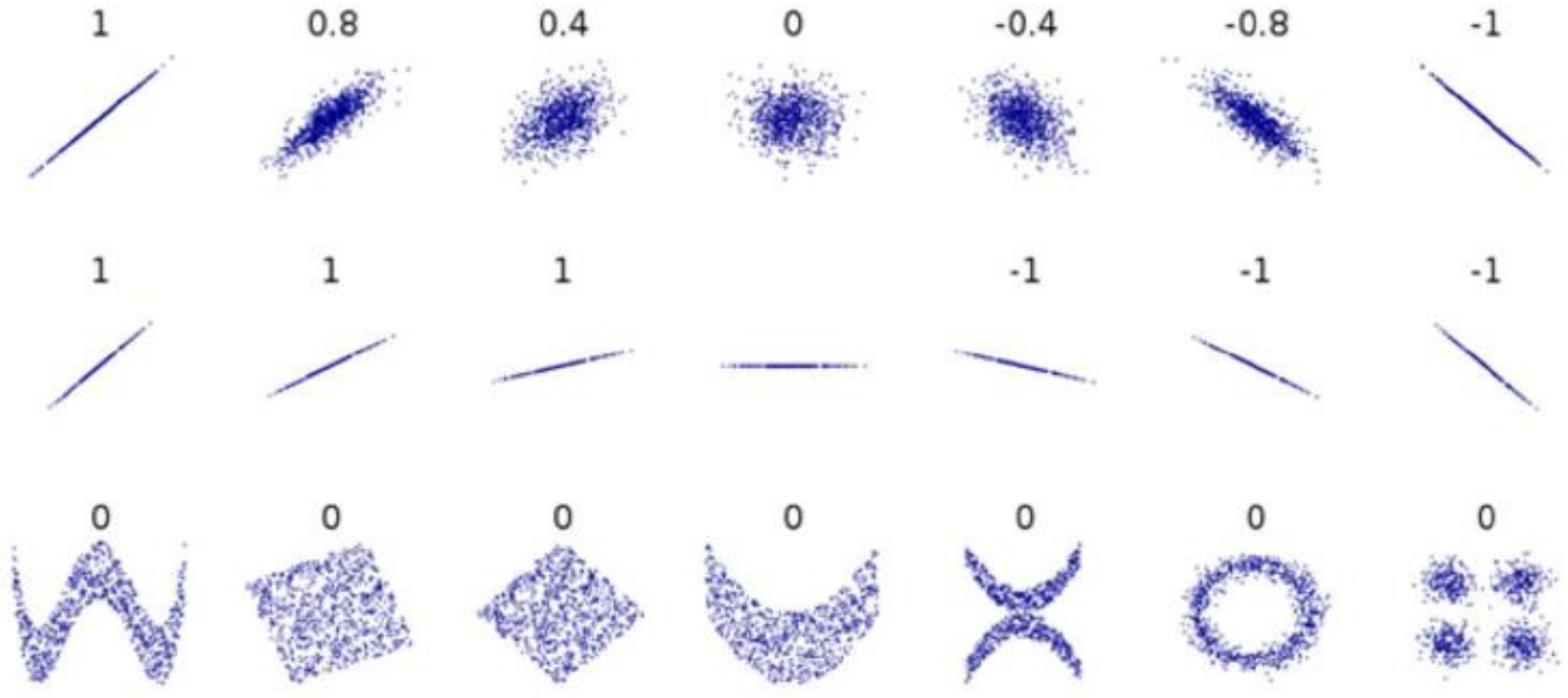
- **Correlation coefficient (r):**
  - The closer it is to -1, the stronger the negative relationship. If the value of one variable increases, the value of the other tends to decrease.
  - The closer it is to 1, the stronger the positive relationship. If the value of one variable increases, the value of the other also tends to increase.
  - The closer it is to 0, the weaker the relationship.

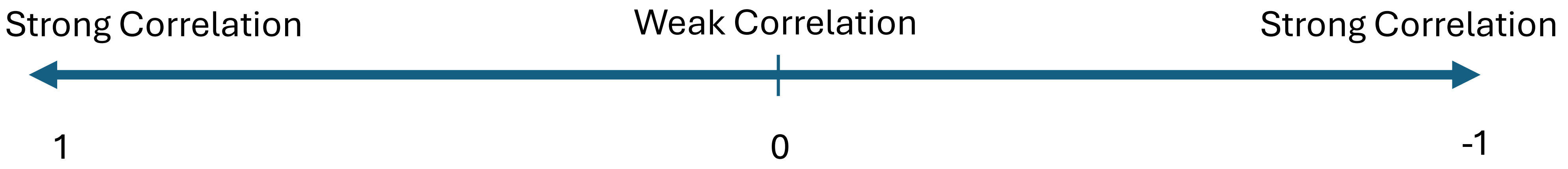
$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}}$$

Where:

- X_i and Y_i: Individual data points for variables X and Y, respectively.
- $\bar{X}$: Mean of X.
- $\bar{Y}$: Mean of Y.
- $\sum$: Summation symbol.

Strength of the Relationship





Regression Model

Regression Model

- Linear Regression
- Polynomial Regression
- Ridge Regression
- Lasso Regression

Simple Linear Regression

- A method to predict a quantitative output Y based on a single predictor variable X

- Assumes there is a linear (approximately) relationship between X and Y

$$Y \approx \beta_0 + \beta_1 X$$

- The coefficients β_0 and β_1 represent the *intercept* and *slope* terms in the linear model

Estimating Coefficients

- β_0 and β_1 and unknown. So we must use data to estimate the coefficients
- Supposing we have *n* data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, our goal is to obtain estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the data:

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

- Thus, we want to find an intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ such that the resulting line is as close as possible to the *n* data points

Estimating Coefficients

- Minimizing the *least squares* criterion is a way of measuring how close our line is from the data points

- Consider $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ the prediction for a measurement of Y based on the ith value of X

- Now $e_i = y_i - \hat{y}_i$ represents the ith *residual*: the difference between the ith observed value and the ith predicted value by the linear model

- Residual Sum of Squares: $RSS = \; e_1^2 + e_2^2 + \ldots + e_n^2$

The Least Squares

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the RSS:

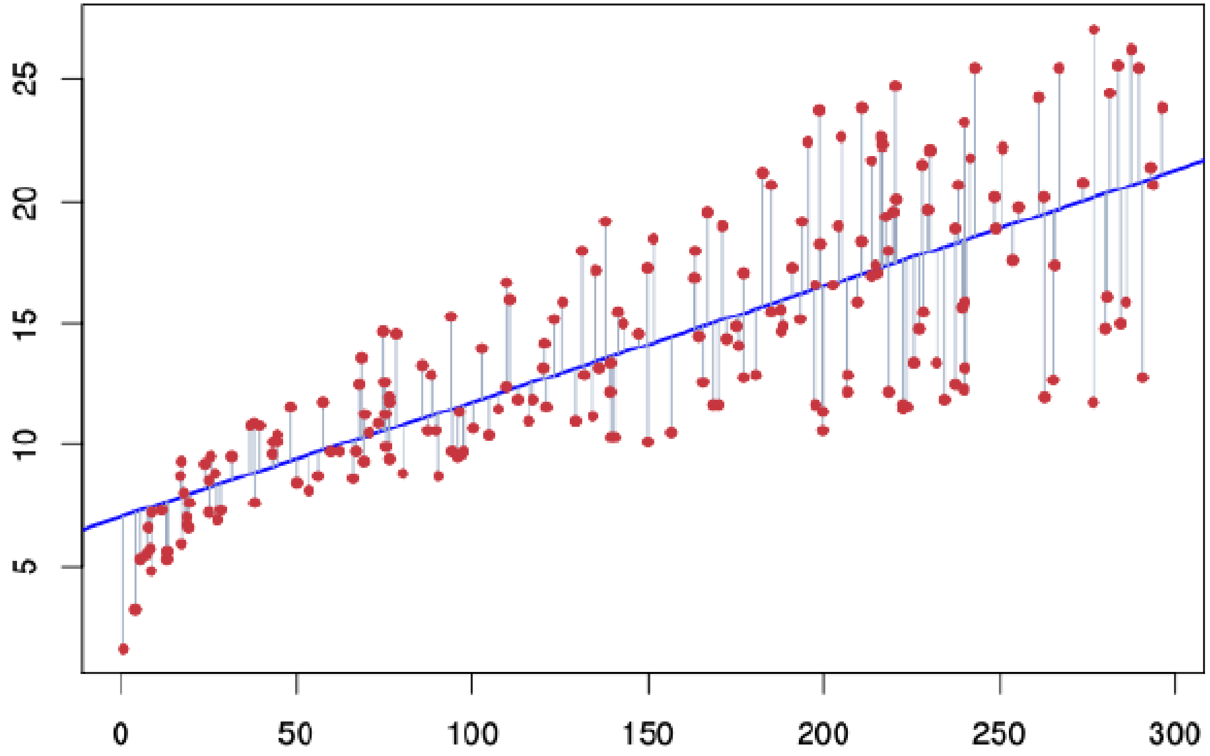
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

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

- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means

The Least Squares

- $\hat{\beta}_0$ and $\hat{\beta}_1$ (the fit) and found by minimizing the residual sum of squares
  - Each grey segment represents a residual



Evaluating the Coefficient Estimates

- How close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values β_0 and β_1?
  - The standard errors associated to $\hat{\beta}_0$ and $\hat{\beta}_1$ are:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})}\right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

  - where $\sigma^2 = Var(\epsilon)$, and ϵ contains the errors for each observation

Confidence Intervals

- Standard errors can be used to compute confidence intervals
  - A 95% confidence interval is the range of values such that with 95% probability, the range will contain the true unknown value of the parameter

- For linear regression, $\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$ is the 95% confidence interval for $\hat{\beta}_1$

Confidence Intervals

- With the 95% confidence interval $\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$:
  - There is approximately a 95 % chance that the interval

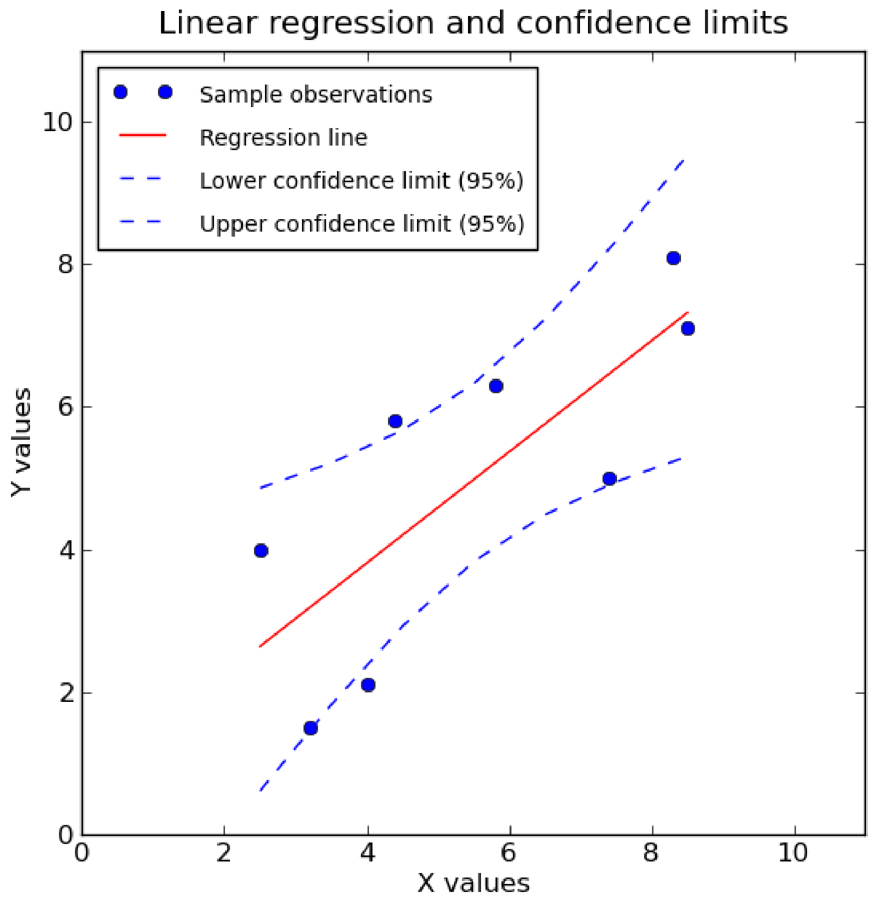
$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)]$$

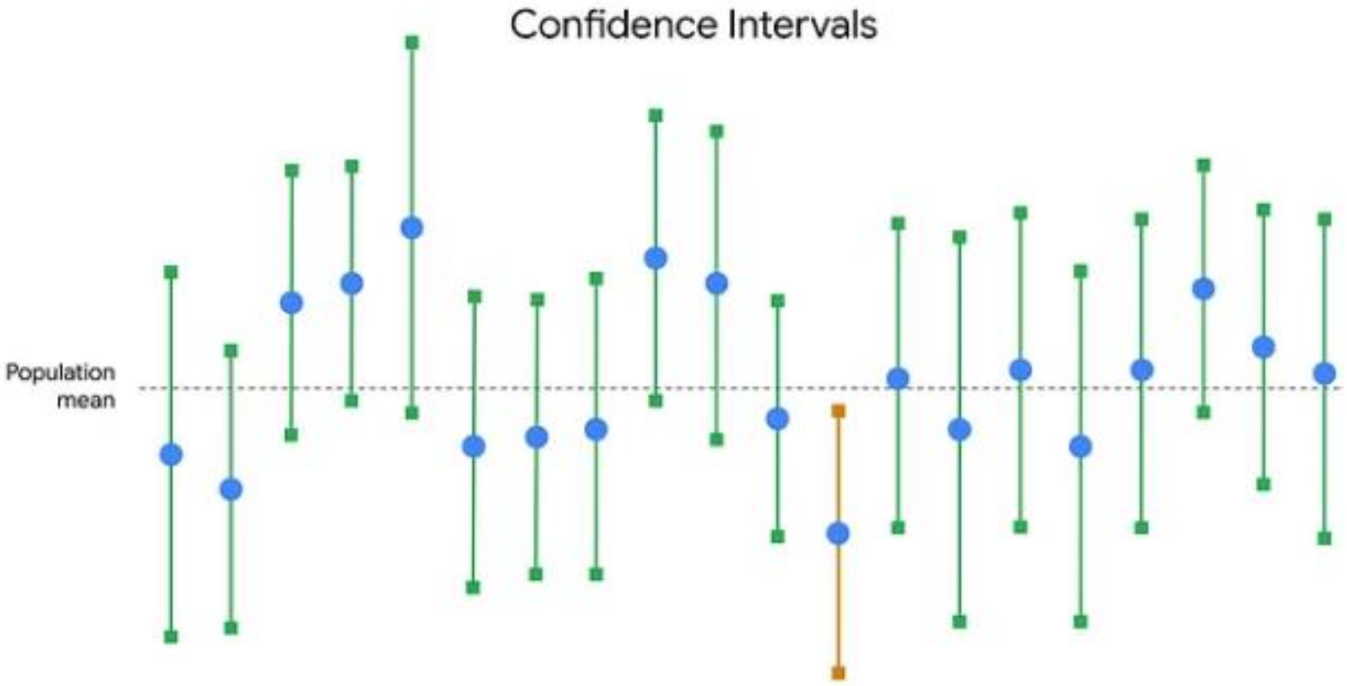
  - will contain the true value of $\hat{\beta}_1$

- A confidence interval for $\hat{\beta}_0$ takes the form

$$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0)$$

Confidence Intervals





Exercise

From the dataset below, find the intercept and slope! Then create a linear model

Month	Income ($)	Occupancy
January	16923	100
February	15797	100
March	17609	110
April	13399	70
May	18252	110

Exercise

From the linear model, predict the income ($) from this dataset!

Date	Occupancy	Income ($)
Jan-25	95	
Feb-25	98	
Mar-25	106	
Apr-25	79	

Evaluating the Model

- Mean Squared Error (MSE)
- Pearson Correlation (R)
- R Squared (R^2)

Mean Squared Error

- **Definition:** MSE measures the average squared difference between the observed (y_i) and predicted ($\hat{y}_i$) values.

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

- **Interpretation:**
  - A smaller MSE indicates better model performance.
  - MSE penalizes larger errors more heavily because of squaring.

Pearson Correlation (*r*)

- **Definition:** Pearson's correlation coefficient quantifies the linear relationship between the observed (y) and predicted ($\hat{y}$) values.

$$r = \frac{\mathrm{Cov}(y, \hat{y})}{\sqrt{\mathrm{Var}(y) \cdot \mathrm{Var}(\hat{y})}}$$

  Where:

  - Cov($y, \hat{y}$): Covariance between y and $\hat{y}$.
  - Var(y): Variance of y, and similarly for $\hat{y}$.

- **Interpretation:**
  - r ranges from -1 to 1.
    - $r = 1$: Perfect positive linear relationship.
    - $r = 0$: No linear relationship.
    - $r = -1$: Perfect negative linear relationship.
  - A higher r value suggests stronger correlation.

R-Squared (R^2)

- **Definition:** R^2 measures the proportion of variance in the dependent variable (y) explained by the independent variable (x) through the model.

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

Where:

- $SSE = \sum(y_i - \hat{y}_i)^2$: Sum of squared errors.
- $SST = \sum(y_i - \bar{y})^2$: Total sum of squares (variability of y around its mean).

or

$$R^2 = \left(\frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x) \cdot \mathrm{Var}(y)}} \right)^2$$

Multiple Linear Regression

- How can we extend our analysis in order to accommodate these additional predictors?
  - We can give each predictor a separate slope coefficient in a single model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- X_j represents the jth predictor and β_j quantifies the association between that variable and the response

Estimating the Regression Coefficients

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$, predictions can me made using the formula:

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

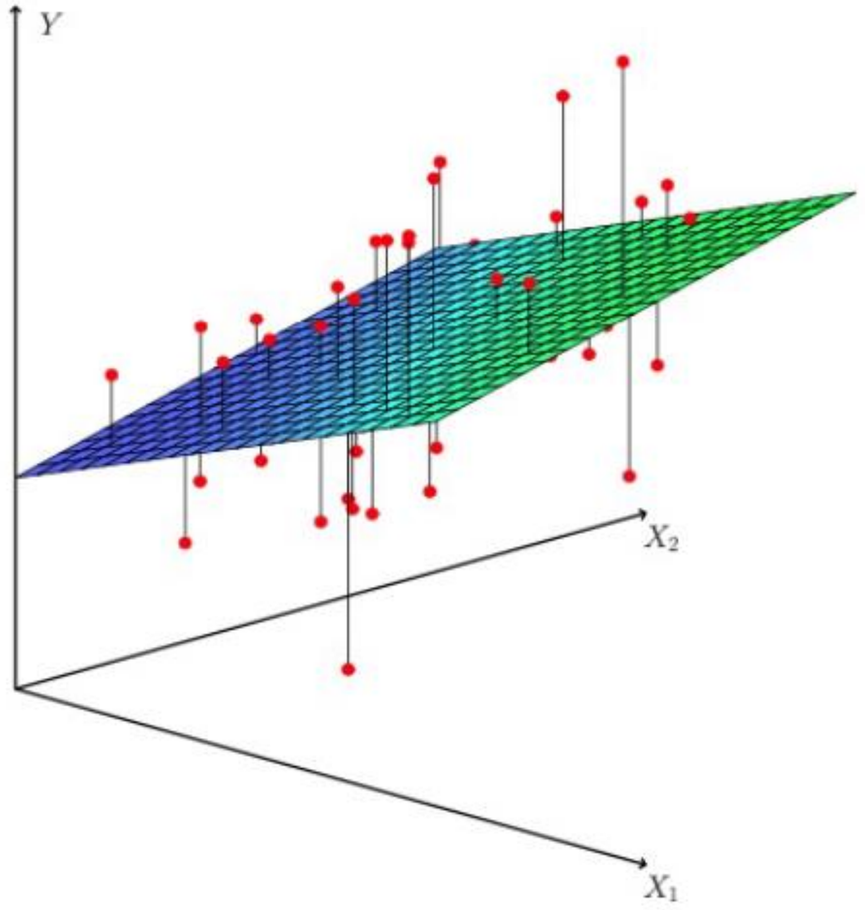
- The parameters are estimated using the same least squares approach used in the context of simple linear regression

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

$$= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}\right)^2$$

Estimating the Regression Coefficients

- The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.



Polynomial Regression

The simplest non-linear model we can consider, for a response Y and a predictor X, is a polynomial model of degree M,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_M x^M + \epsilon.$$

Just as in the case of linear regression with cross terms, polynomial regression is a special case of linear regression - we treat each x^m as a separate predictor. Thus, we can write:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \ldots & x_1^M \\ 1 & x_2^1 & \ldots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^M \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$

Generalized Polynomial Regression

We can generalize polynomial models:

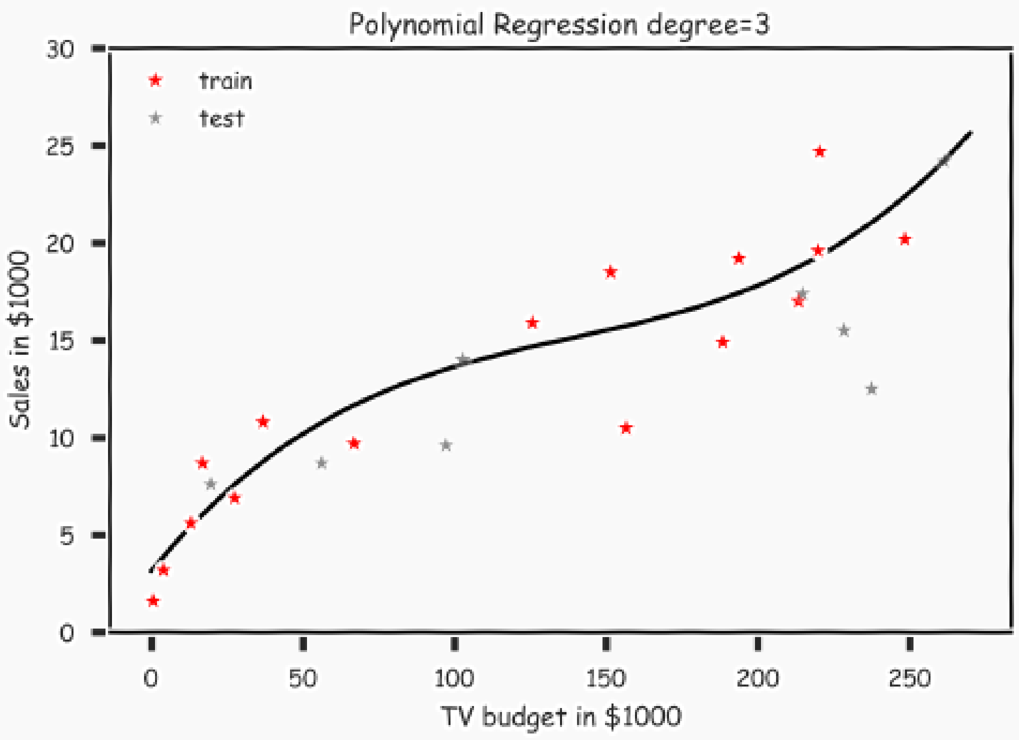
1. consider polynomial models with multiple predictors $\{X_1, \ldots, X_j\}$:

$$
\begin{aligned}
y = &\beta_0 + \beta_1 x_1 + \ldots + \beta_M x_1^M \\
&+ \beta_{M+1} x_2 + \ldots + \beta_{2M} x_2^M \\
&+ \ldots \\
&+ \beta_{M(J-1)+1} x_J + \ldots + \beta_{MJ} x_J^M
\end{aligned}
$$

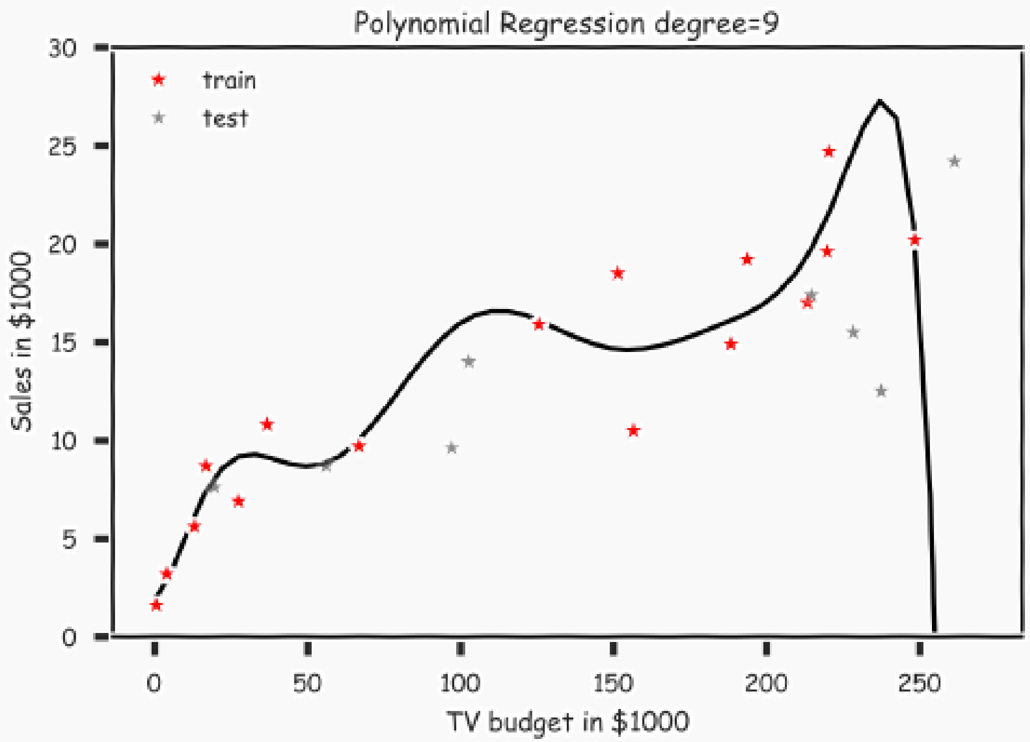
2. consider polynomial models with multiple predictors $\{X_1, X_2\}$ and cross terms:

$$
\begin{aligned}
y = &\beta_0 + \beta_1 x_1 + \ldots + \beta_M x_1^M \\
&+ \beta_{1+M} x_2 + \ldots + \beta_{2M} x_2^M \\
&+ \beta_{1+2M} (x_1 x_2) + \ldots + \beta_{3M} (x_1 x_2)^M
\end{aligned}
$$

Polynomial Regression Degree



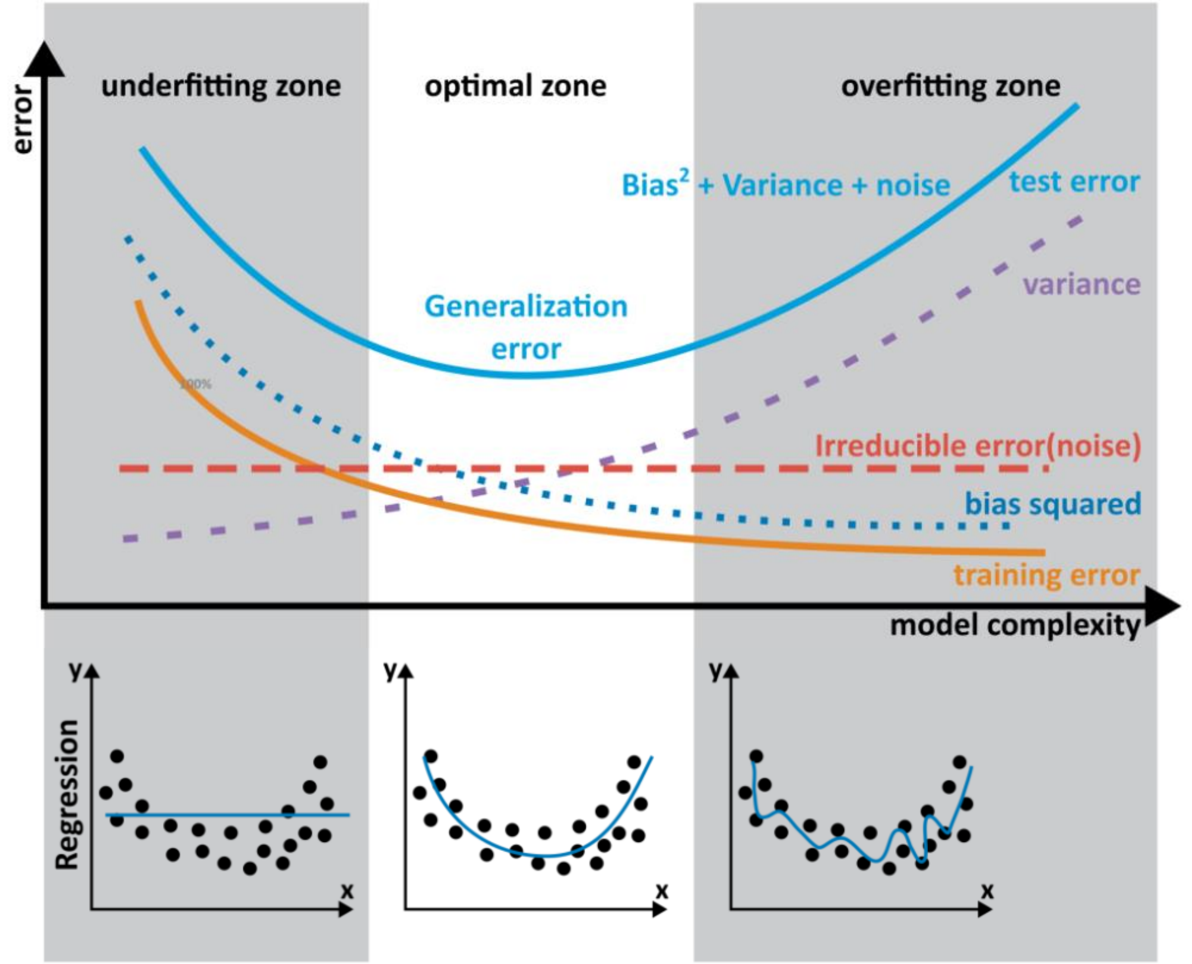
Polynomial Regression Degree



Shrinkage Methods

- Shrinking the coefficient estimates can significantly reduce variance

- Two best-known techniques:
  - Ridge Regression
  - Lasso Regression



Ridge Regression

- The least squares fitting procedure estimates the coefficients using the values that minimize:

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

- Ridge regression is similar, choosing $\hat{\beta}_\lambda^R$ that minimize:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2 = RSS + \lambda\sum_{j=1}^{p}{\beta_j}^2$$

Ridge Regression

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2 = RSS + \lambda\sum_{j=1}^{p}{\beta_j}^2$$

- In this equation, $\lambda \geq 0$ is a tuning parameter.
- We seek for coefficients that lead to small RSS. The second term, $\lambda \sum_{j=1}^{p} {\beta_j}^2$, called shrinkage penalty, is small when the coefficients are close to zero
- The second term has the effect of shrinking the estimates of β_j towards zero

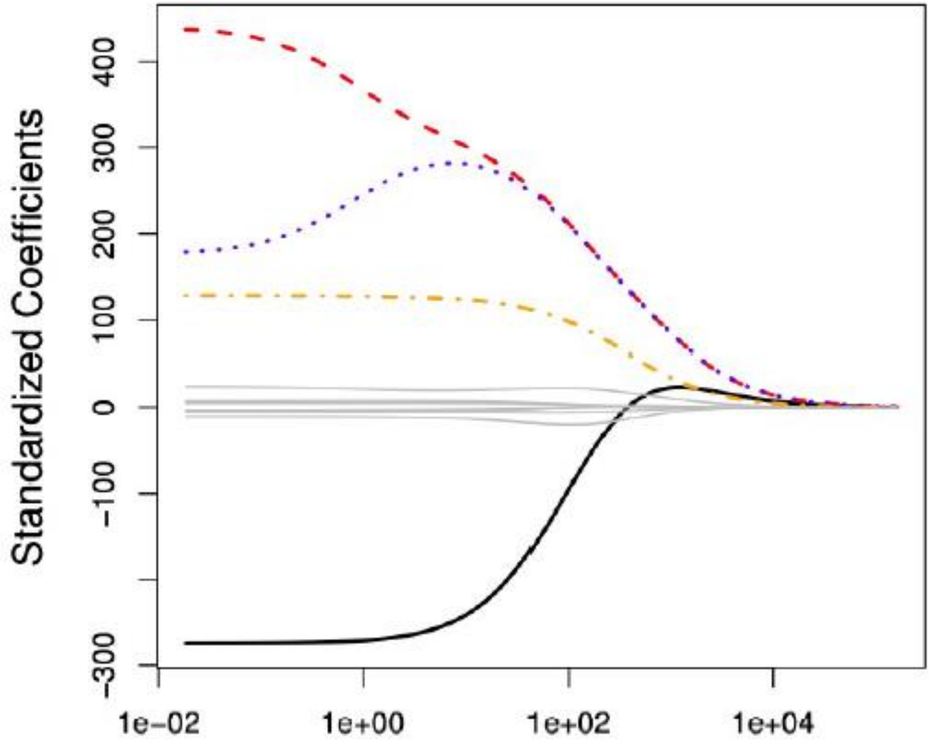
Ridge Regression

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2 = RSS + \lambda\sum_{j=1}^{p}{\beta_j}^2$$

- The tuning parameter λ controls the relative impact of the two terms on the regression coefficient estimates
- When $\lambda = 0$ the penalty term has no effect, and ridge regression produces the least squares estimates
- As $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero

Ridge Regression

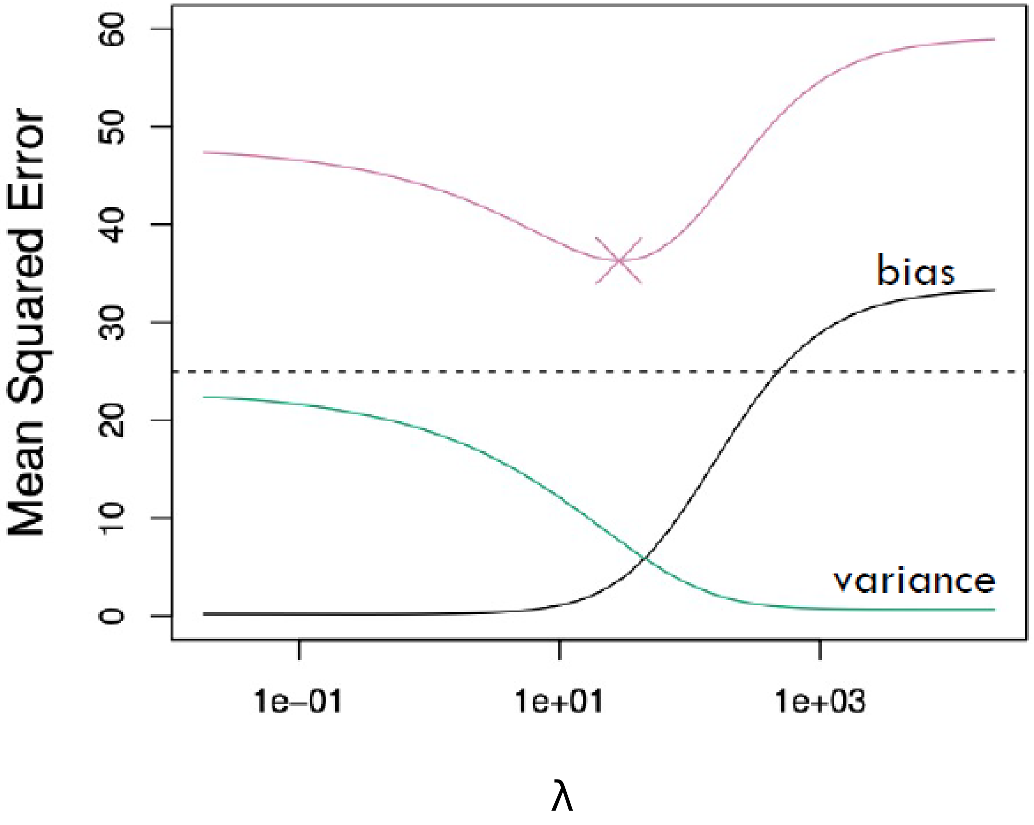
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2 = RSS + \lambda\sum_{j=1}^{p}{\beta_j}^2$$



λ

Ridge Regression

- The advantage of ridge regression's over least squares is rooted in the bias-variance trade-off



Lasso Regression

- Ridge regression includes all p predictors in the final model
  - The penalty term $\lambda \sum \beta_j^{\,2}$ will shrink all the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$)
  - This can be a problem in model interpretation, when the number of variables p is very large
  - We might wish to build a model including just the predictors considered more important for the desired outcome

Lasso Regression

- Lasso is an alternative to ridge regression that overcomes this disadvantage, choosing $\hat{\beta}_\lambda^L$ to minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

- We just replaced ${\beta_j}^2$ by $|\beta_j|$ in the penalty
- We now have an ℓ_1 penalty instead of an ℓ_2 penalty
- The ℓ_1 norm of a coefficient vector β is $\|\beta\|_1 = \sum|\beta_j|$

Lasso Regression

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

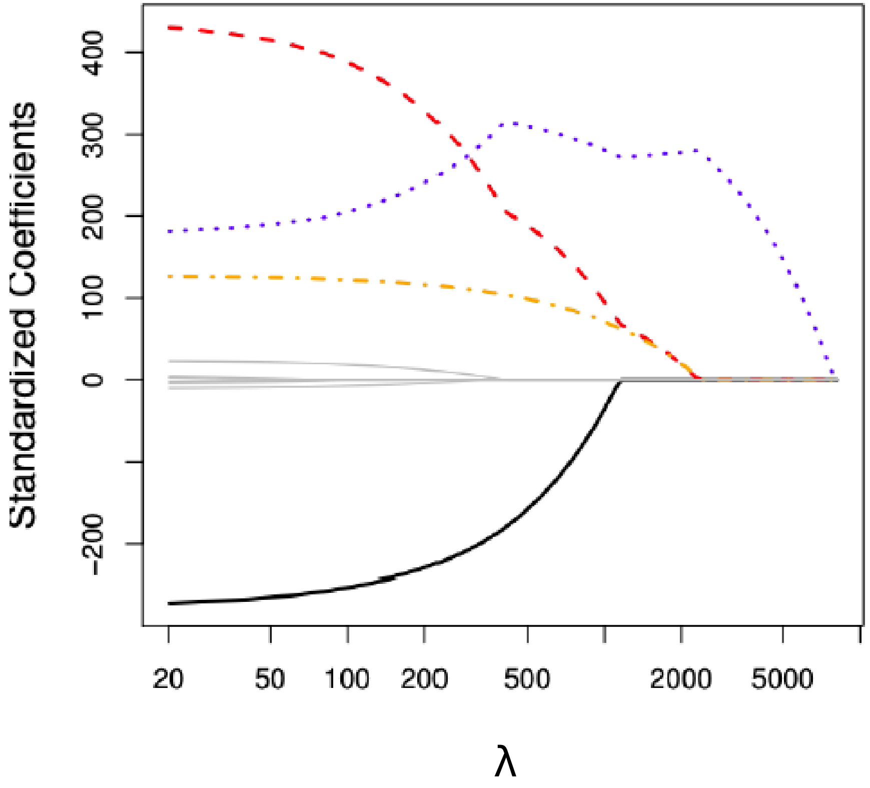
- The ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large

- Thus, lasso performs variable selection

- Models generated from the lasso are generally much easier to interpret than those produced by ridge regression

Lasso Regression

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = RSS+\lambda\sum_{j=1}^{p}|\beta_j|$$



Thank you