GENERAL ASSEMBLY

REGULARIZATION: RIDGE AND LASSO REGRESSION

Instructor: Aymeric Flaisler
(Adapted from Joseph Nelson - GA - DSI Washington)

AGENDA

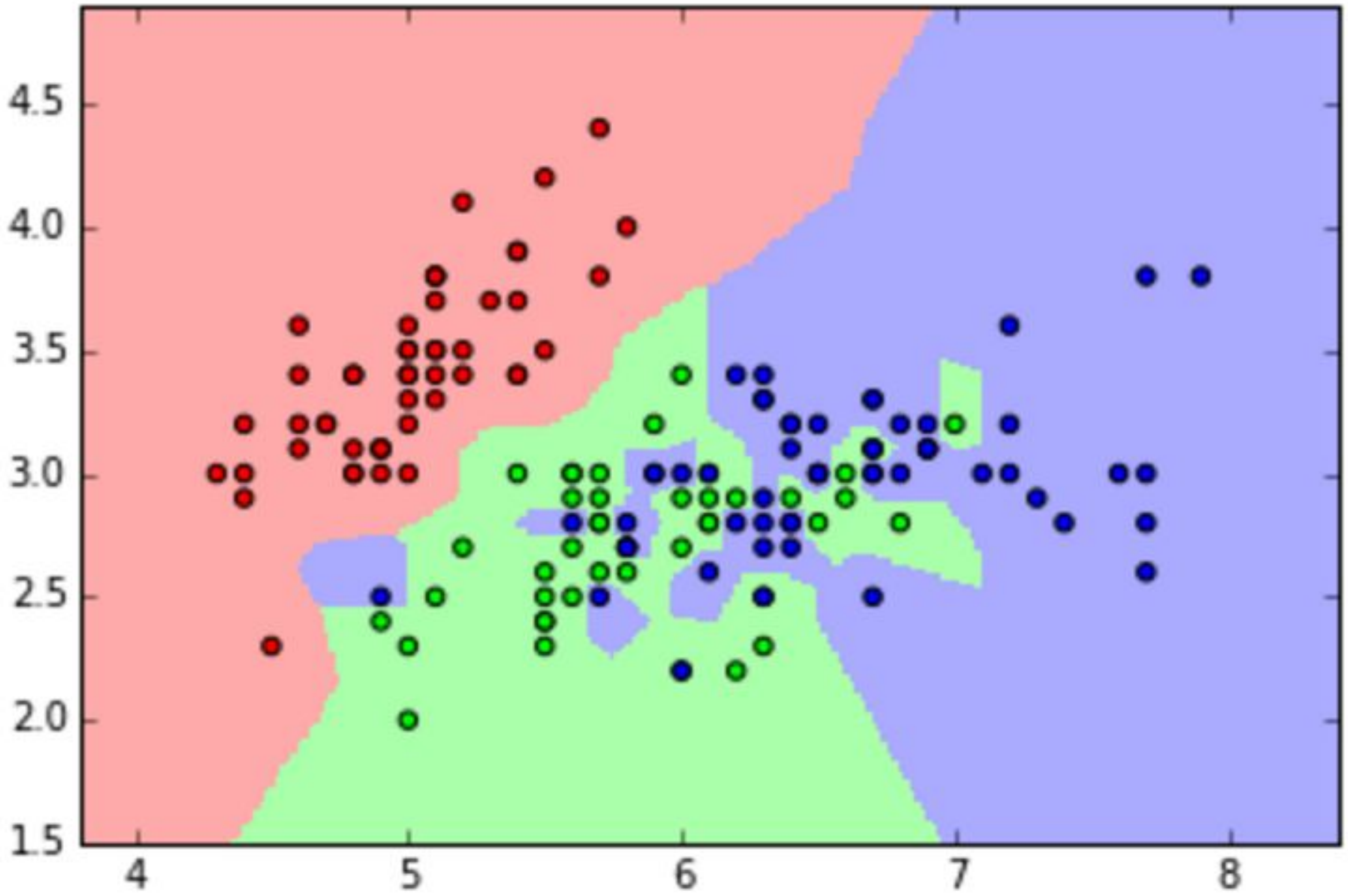
- What is overfitting?
- Overfitting with linear models
- Regularization of linear models
- Regularized regression in scikit-learn
- Comparing regularized linear models with unregularized linear models
- Coding implementation

OVERFITTING

- **What is overfitting?**
  - Building a model that matches the training data "too closely”
  - Learning from the noise in the data, rather than just the signal
- **How does overfitting occur?**
  - Evaluating a model by testing it on the same data that was used to train it
  - Creating a model that is "too complex"
- **What is the impact?**
  - Model will do well on the training data, but won't generalize to out-of-sample data
  - Model will have low bias, but high variance

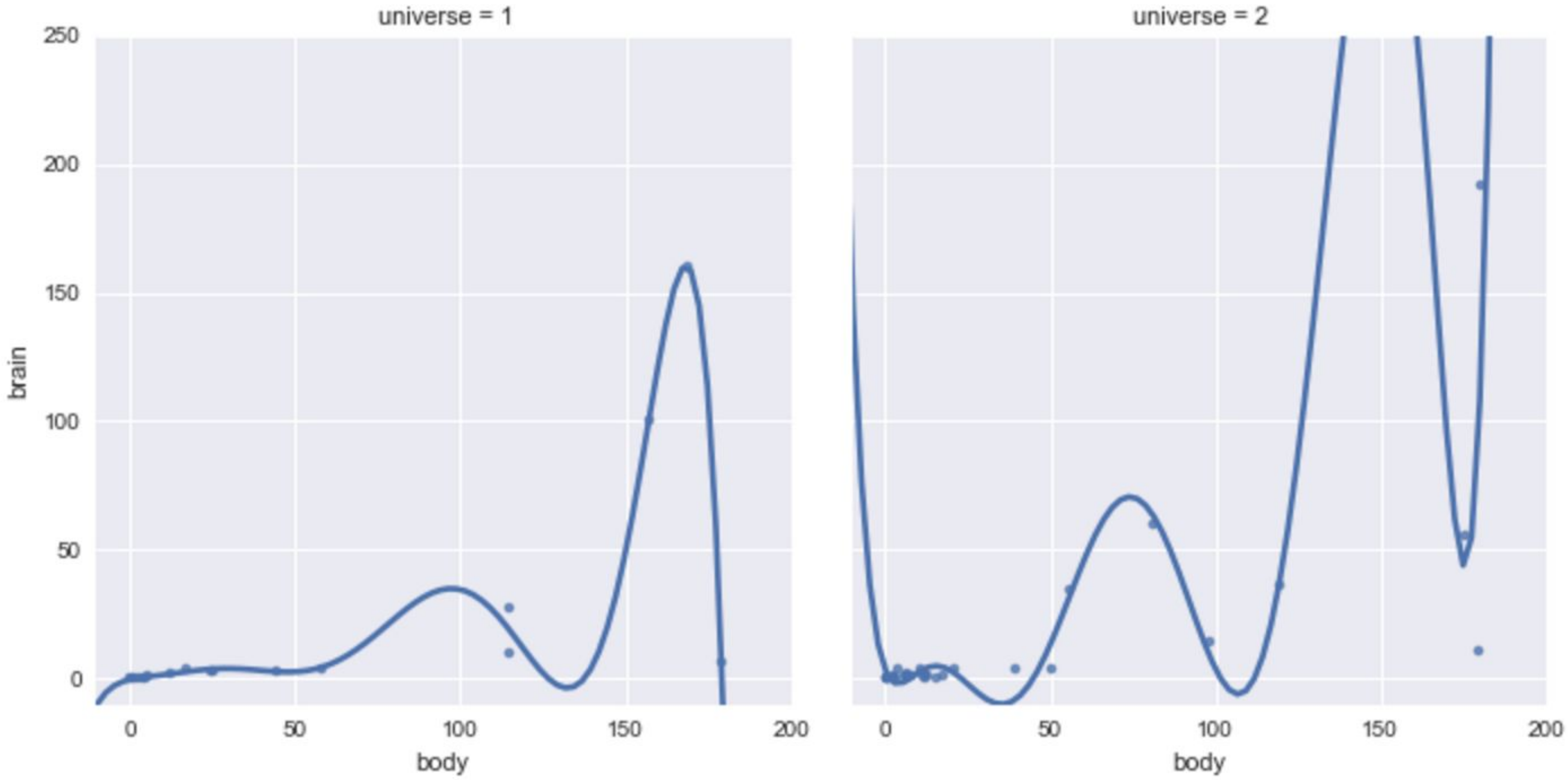
OVERFITTING

- Overfitting with KNN



REVIEW: OVERFITTING

- Overfitting with polynomial regression



OVERFITTING WITH LINEAR MODELS

- **What are the general characteristics of linear models?**

    Low model complexity

    High bias, low variance

    Does not tend to overfit

- Nevertheless, **overfitting can still occur** with linear models if you allow them to have **high variance**. Here are some common causes:

CAUSE 1: IRRELEVANT FEATURES

- Linear models can overfit if you include "irrelevant features", meaning features that are unrelated to the response. Why?

  Because it will learn a coefficient for every feature you include in the model, regardless of whether that feature has the signal or the noise.

  This is especially a problem when **m (number of features) is close to n (number of observations),** because that model will naturally have high variance.

CAUSE 2: CORRELATED FEATURES

Linear models can overfit if the included features are highly correlated with one another. Why?

From the scikit-learn documentation:

> "...coefficient estimates for Ordinary Least Squares rely on the independence of the model terms. When terms are correlated and the columns of the design matrix X have an approximate linear dependence, the design matrix becomes close to singular and as a result, the least-squares estimate becomes highly sensitive to random errors in the observed response, producing a large variance."
>
> http://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares

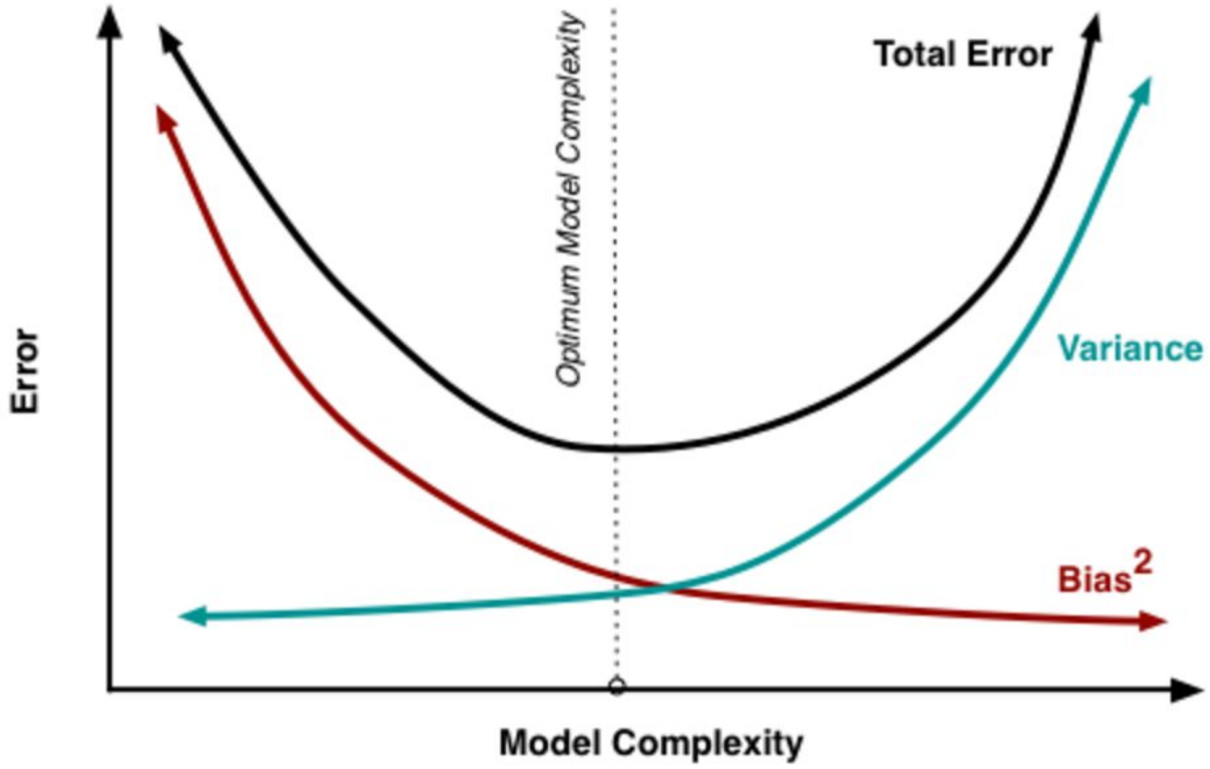
CAUSE 3: LARGE COEFFICIENTS

- Linear models can overfit if the coefficients (after feature standardization) are too large. Why?

  Because the **larger** the absolute value of the coefficient, the more **power** it has to change the predicted response, resulting in a higher variance.

REGULARIZATION OF LINEAR MODELS

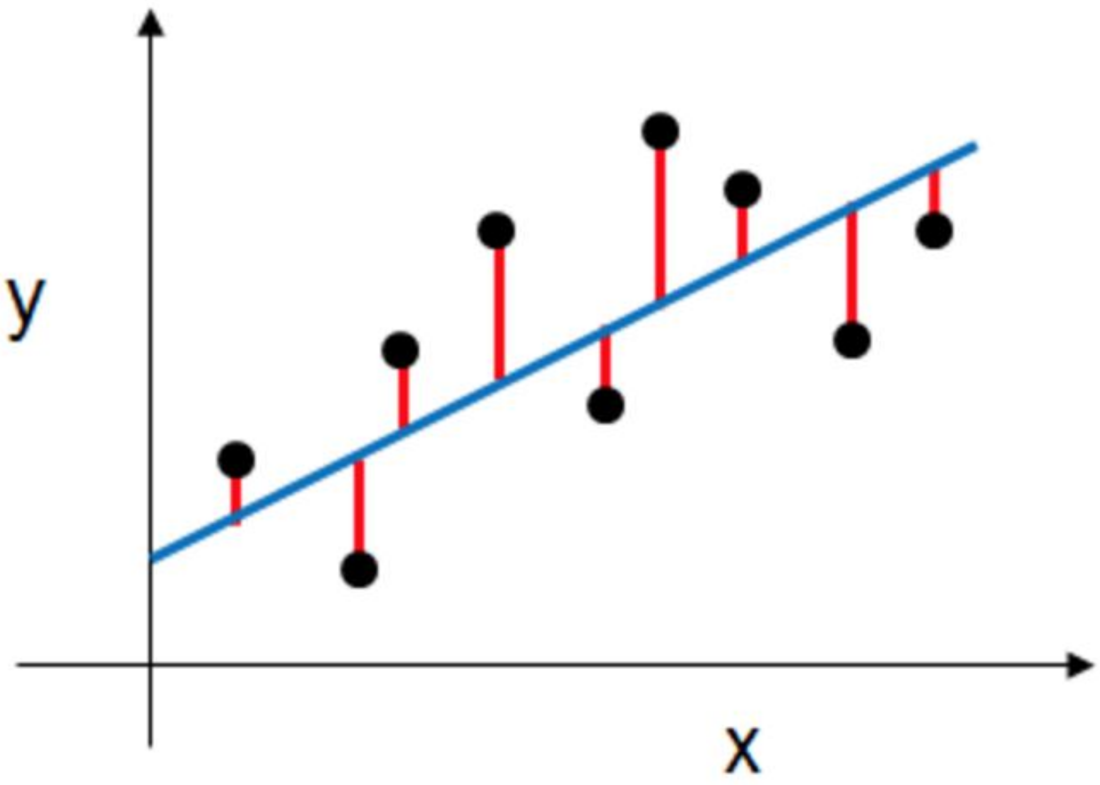
- Regularization is a method for "constraining" or "regularizing" the size of the coefficients, thus "shrinking" them towards zero.
- It reduces model variance and thus **minimizes overfitting**.
- If the model is too complex, it tends to reduce variance more than it increases bias, resulting in a model that is more likely to generalize.



- Our goal is to locate the **optimum model complexity**, and thus regularization is useful when we believe our model is too complex.

HOW DOES REGULARIZATION WORK?

- For a normal linear regression model, we estimate the coefficients using the least squares criterion, which minimizes the residual sum of squares (RSS):



Model Prediction

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

Observed Result

RIDGE REGRESSION

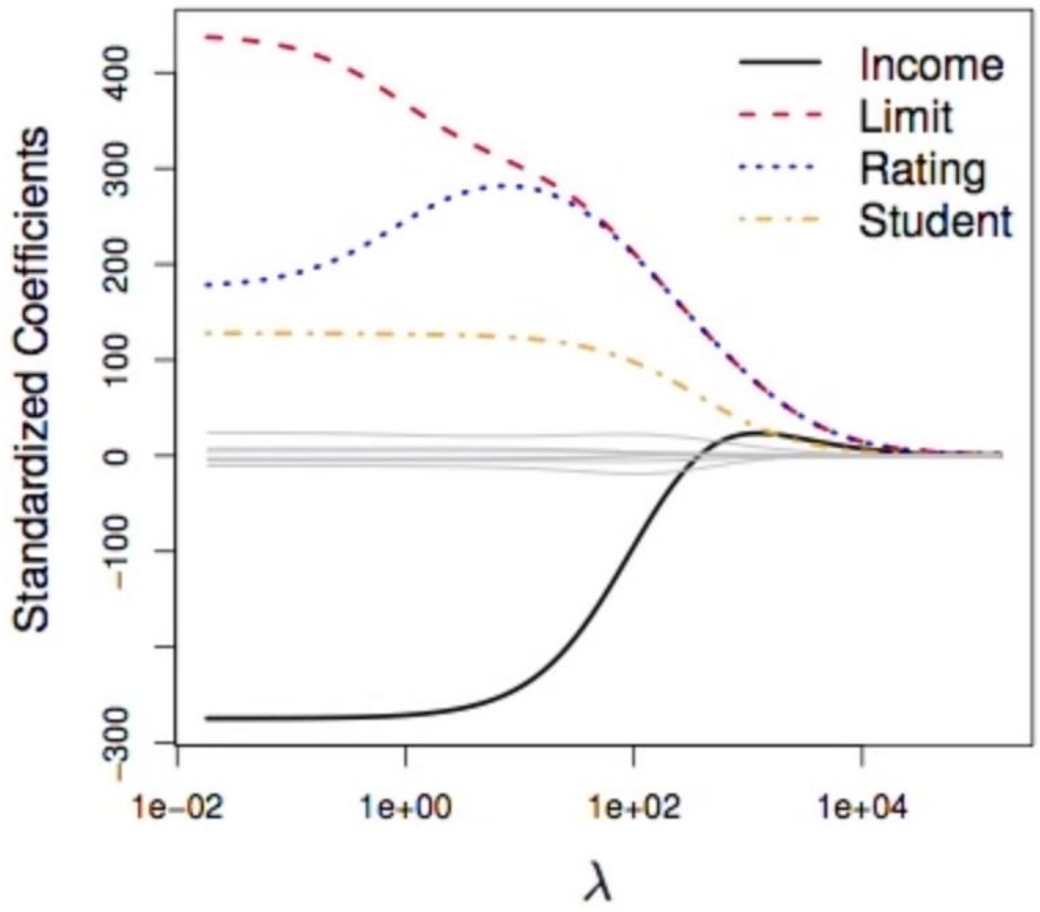
- We seek to minimize the squared errors AND some penalty term, whose power is equal to lambda.

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

where $\lambda \geq 0$ is a *tuning parameter*, to be determined separately.

RIDGE REGRESSION

- Ridge coefficients as a function of our regularization penalty:



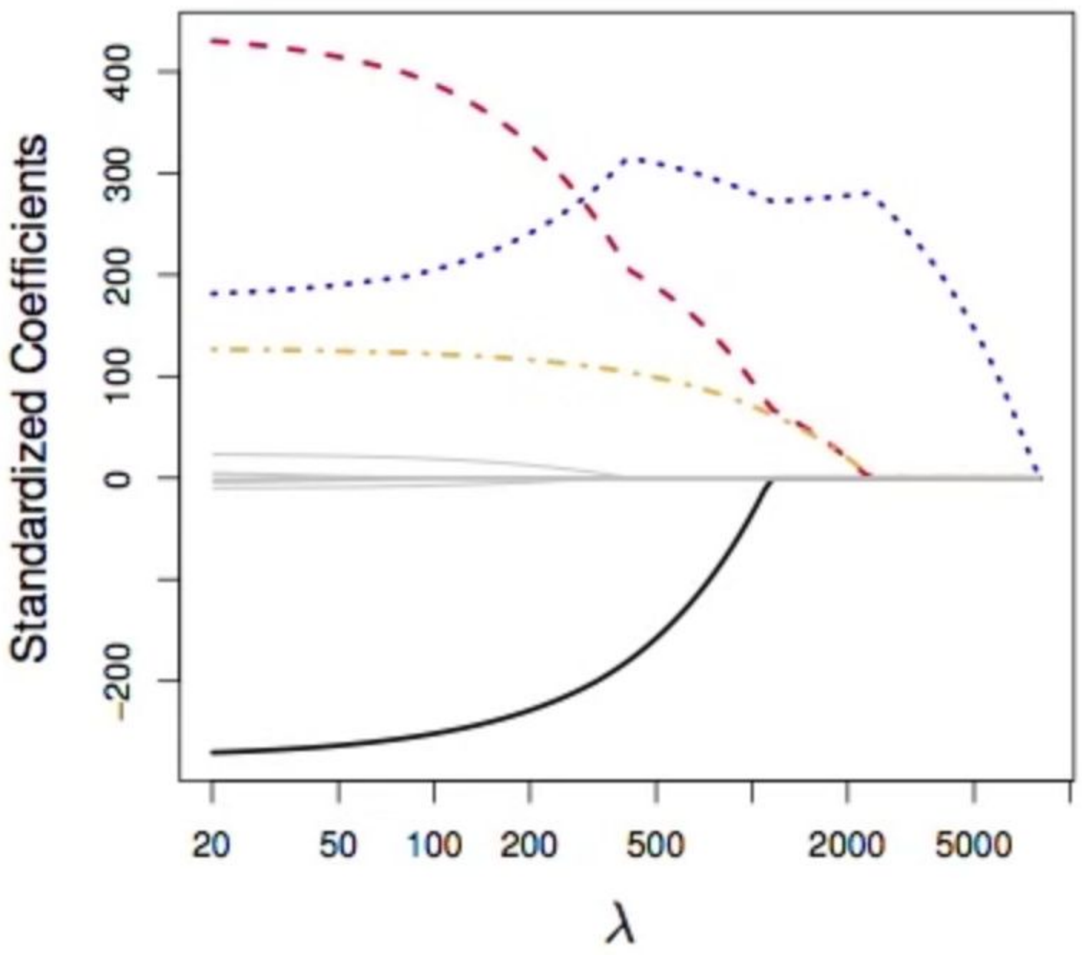
LASSO REGRESSION REGRESSION

- We seek to minimize the squared errors AND some penalty term, whose power is equal to lambda.

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

LASSO REGRESSION REGRESSION

- Lasso coefficients as a function of our regularization penalty:



RIDGE VS LASSO REGRESSION

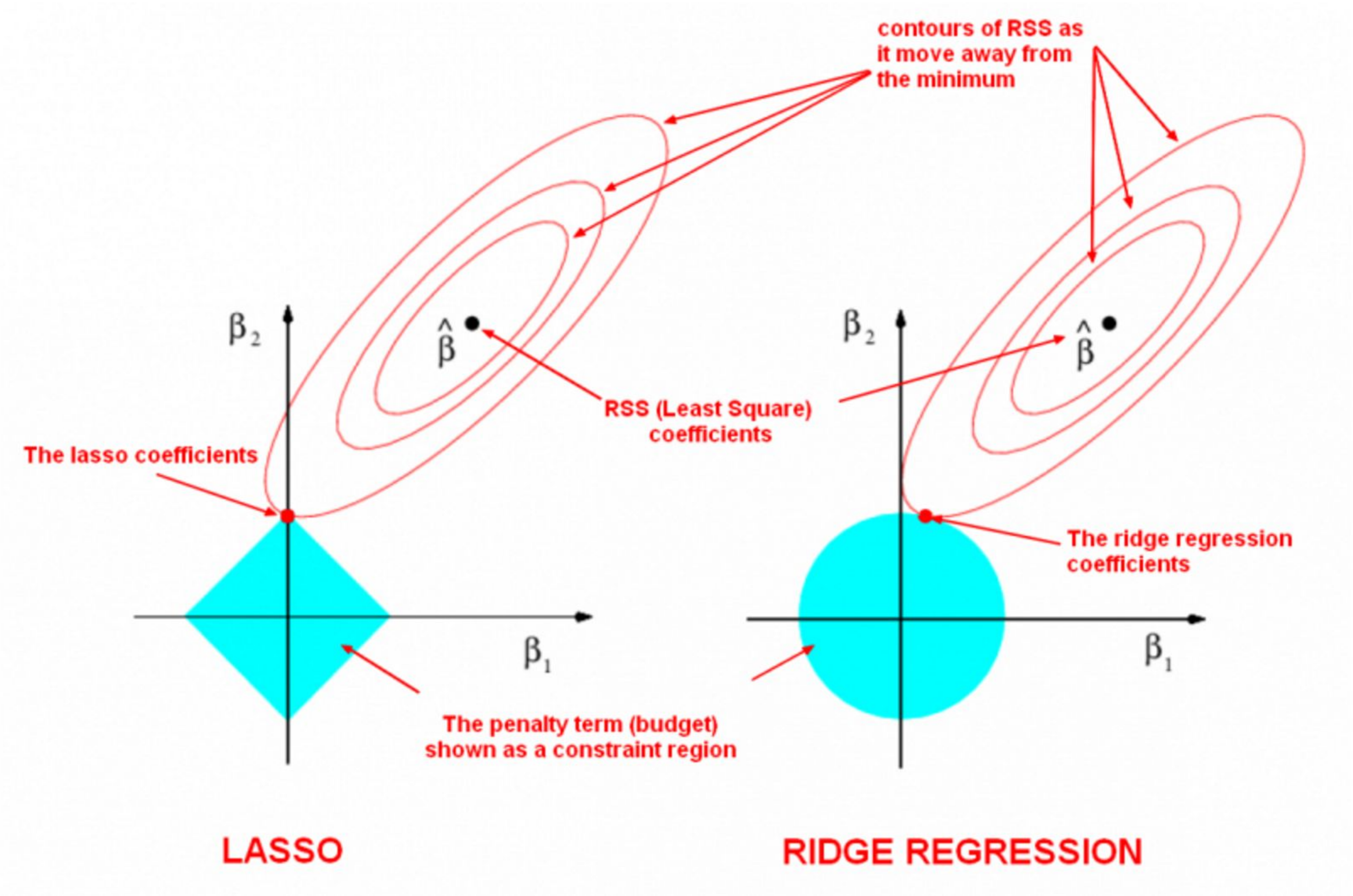
- Lasso Regression (L1 norm): shrink towards 0 using the sum of the absolute value of our coefficients as a constraint
- Ridge Regression (L2 norm): shrink the squares of our our coefficients

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

RIDGE VS LASSO REGRESSION



RIDGE VS LASSO REGRESSION

- Previous slide:
- We are fitting a linear regression model with two features, x1 and x2.
- $\hat{\beta}$ represents the set of two coefficients, β_1 and β_2, which minimize the RSS for the **unregularized model**.
- Regularization restricts the allowed positions of $\hat{\beta}$ to the blue constraint region:
- For lasso, this region is a diamond because it constrains the absolute value of the coefficients.
- For ridge, this region is a circle because it constrains the square of the coefficients.
- When α is zero, the blue region is infinitely large, and thus the coefficient sizes are not constrained.

RIDGE VS LASSO REGRESSION

- **Should features be standardized?**

  Yes, because otherwise, features would be penalized simply because of their scale.

  Also, standardizing avoids penalizing the intercept, which wouldn't make intuitive sense.

- **How should you choose between Lasso regression and Ridge regression?**

  Lasso regression is preferred if we believe many features are irrelevant or if we prefer a sparse model.

  If model performance is your primary concern, it is best to try both.

  ElasticNet regression is a combination of lasso regression and ridge Regression.