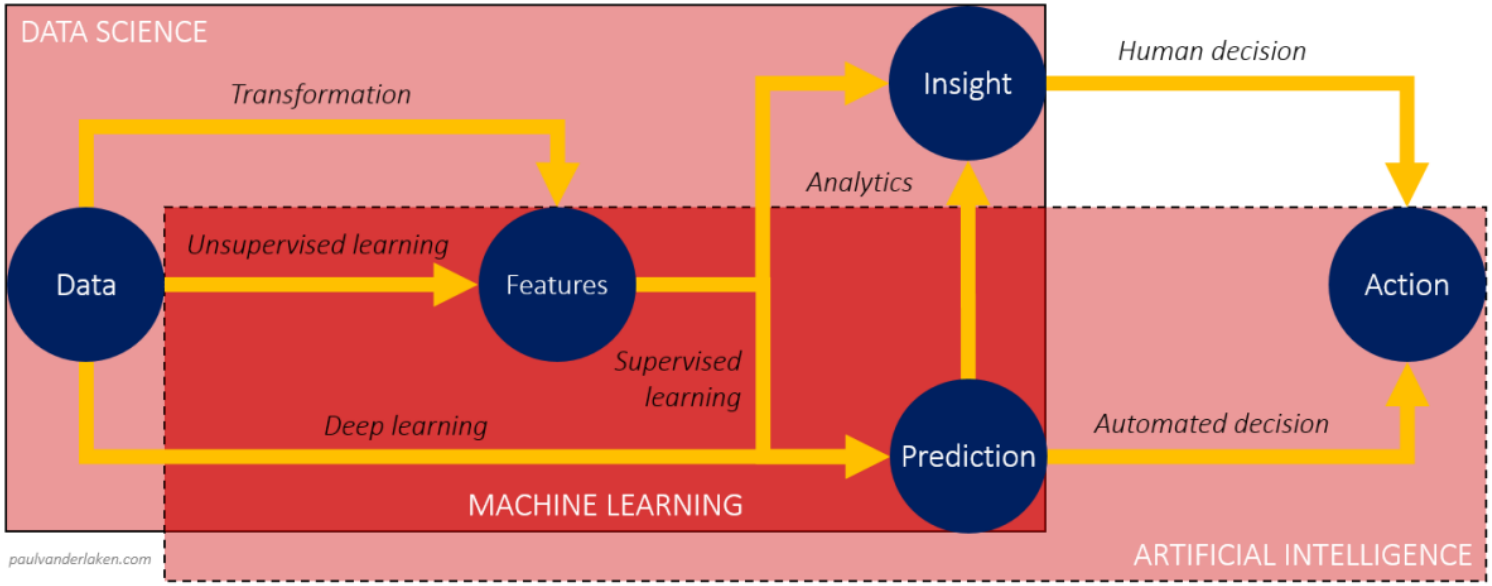
Supervised Learning for Regression



Overview: Linear Regression

- Linear regression is a simple approach to supervised learning, as it assumes that the dependence of Y on X_1, X_2,…,X_p is linear.

- Most modern machine learning approaches can be seen as generalizations or extensions of linear regression.

- When augmented with kernels or other forms of basis function expansion (which replace X with some non-linear function of the inputs), it can also model non-linear relationships.

- Goal: predict Y from X by $f(X)$

Linear Regression Model

- Input vector: $X^T = (X_1, X_2, \ldots, X_p)$
- Output Y is real-valued (quantitative response) and ordered
- We want to predict Y from X.
- Before we actually do the prediction, we have to *train* the function $f(X)$.
- By the end of training, we have a function $f(X)$ to map every X into an estimated Y (aka $\hat{Y}$).

Linear Regression Model (cont.)

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

- This is a linear combination of the measurements that are used to make predictions, plus a constant.

- No matter the source of the X_j, the model is linear in the parameters.

- β_0 is the intercept and β_j is the slope for the *j*th variable X_j, which is the **average** increase in Y when X_j is increased by one unit and all other X's are held constant.

Ordinary Least Squares Estimation

- Typically we have a set of *training data* $(X_1,Y_1)\ldots(X_n,Y_n)$ from which to estimate the parameters β.

- Each X_i is a vector of feature measurements for the ith case.

- We can apply the method of MLE to the linear regression setting (using the definition of the Gaussian), where the log-likelihood function is given by:

$$\ell(\boldsymbol{\theta}) = \frac{-1}{2\sigma^2} RSS(\boldsymbol{\beta}) - \frac{n}{2} \ln 2\pi\sigma^2$$

OLS Estimation (cont.)

- Note that RSS stands for *residual sum of squares*:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - f(X_i))^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2$$

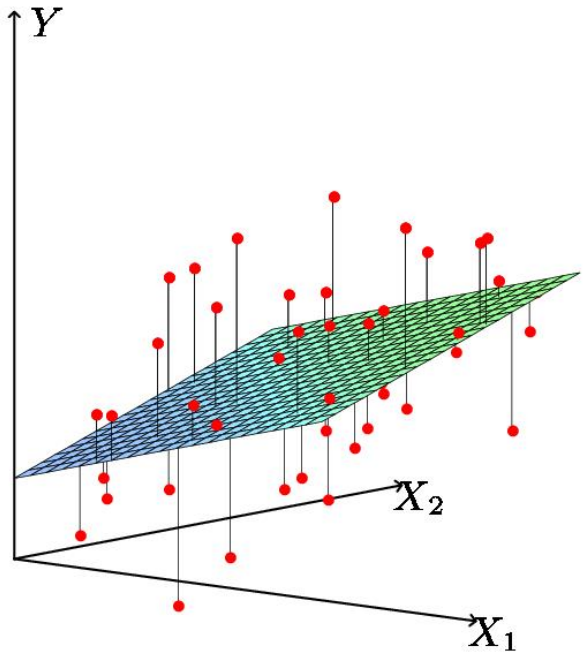
- The RSS is also called the *sum of squared errors* (SSE), where

$$MSE = \frac{SSE}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

- We see that the MLE for $\boldsymbol{\beta}$ is the one that minimizes the RSS.
- Thus, we estimate the parameters using *ordinary least squares* (OLS), which is identical to the MLE, to choose $\hat{\beta}_0$ through $\hat{\beta}_p$ as to minimize the RSS.

OLS Estimation (cont.)

- We illustrate the geometry of OLS fitting, where we seek the linear function of *X* that minimizes the sum of squared residuals from *Y*.



- The predictor function corresponds to a plane (hyper plane) in the 3D space.
- For accurate prediction, hopefully the data will lie close to this hyper plane, but they won't lie exactly in the hyper plane.

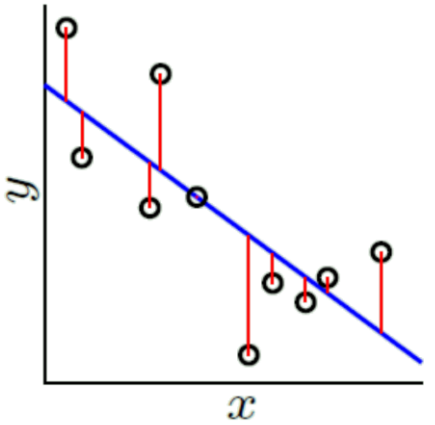
OLS Estimation (cont.)

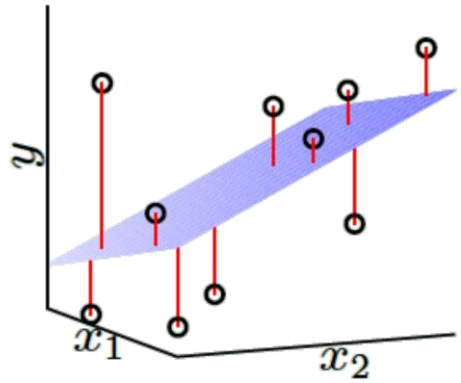
OLS Linear Regression Algorithm:

1. From the training data set, construct the input matrix $\boldsymbol{X}$ and the output vector $\boldsymbol{Y}$

2. Assuming $\boldsymbol{X}^T\boldsymbol{X}$ is invertible (positive definite and non-singular), compute $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

3. Return $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$





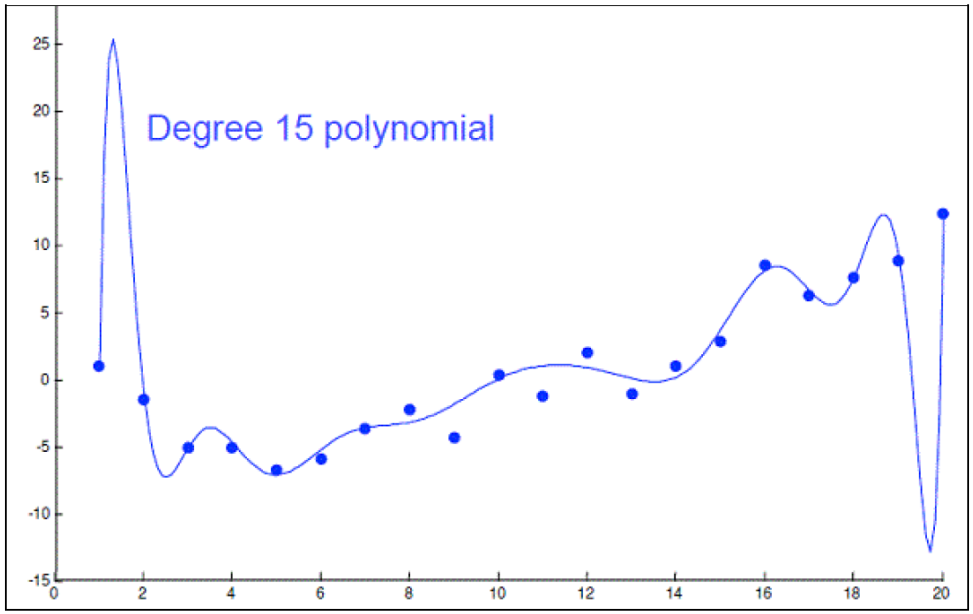
Improving the Linear Model

- We may want to improve the simple linear model by replacing OLS estimation with some alternative fitting procedure.

- Why use an alternative fitting procedure?
  - Prediction Accuracy
  - Model Interpretability

Feature/Variable Selection

- Carefully selected features can improve model accuracy, but adding too many can lead to overfitting.
  - Overfitted models describe random error or noise instead of any underlying relationship.
  - They generally have poor predictive performance on test data.



- For instance, we can use a 15-degree polynomial function to fit the following data so that the fitted curve goes nicely through the data points.
- However, a brand new dataset collected from the same population may not fit this particular curve well at all.

Feature/Variable Selection (cont.)

- Subset Selection
  - Identify a subset of the *p* predictors that we believe to be related to the response; then, fit a model using OLS on the reduced set.
  - Methods: best subset selection, stepwise selection

- Shrinkage (Regularization)
  - Involves shrinking the estimated coefficients toward zero relative to the OLS estimates; has the effect of reducing variance and performs variable selection.
  - Methods: ridge regression, lasso

Shrinkage (Regularization) Methods

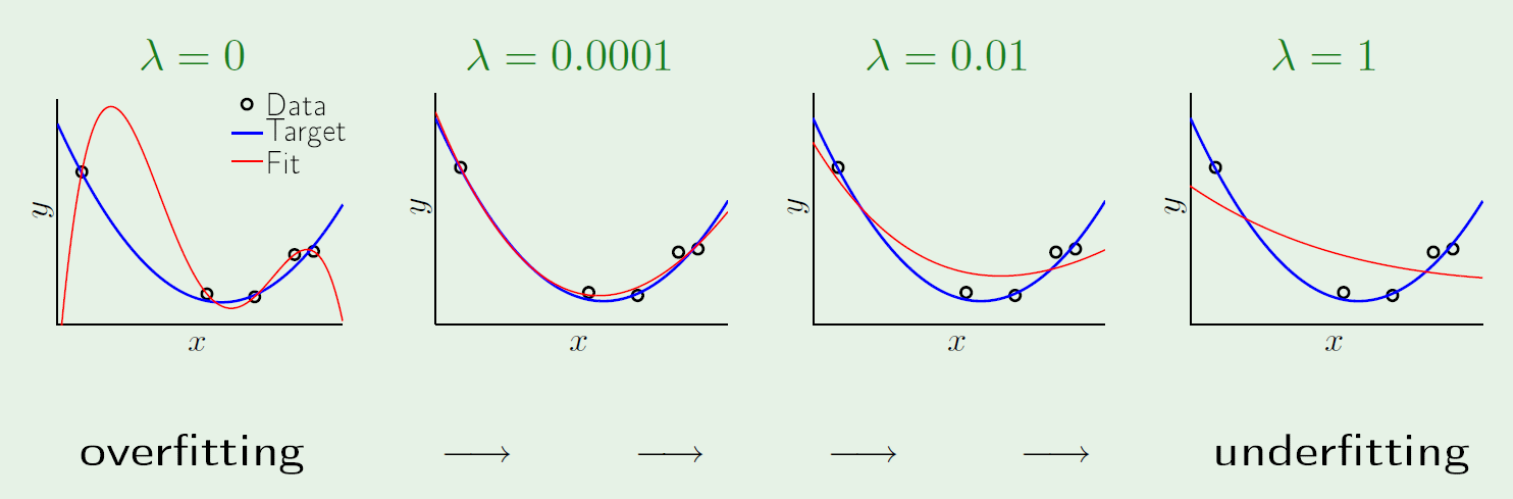
- The subset selection methods use OLS to fit a linear model that contains a subset of the predictors.

- As an alternative, we can fit a model containing all *p* predictors using a technique that constrains or *regularizes* the coefficient estimates (i.e., *shrinks* the coefficient estimates towards zero).

- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that *shrinking* the coefficient estimates can significantly reduce their variance.

Shrinkage (Regularization) Methods (cont.)

- Regularization is our first weapon to combat overfitting.
- It constrains the prediction algorithm to improve out-of-sample error (i.e., test error), especially when noise is present.
- Look at what a little regularization can do:



Ridge Regression

- Recall that the OLS fitting procedure estimates the beta coefficients using the values that minimize:

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

- Ridge regression is similar to OLS, except that the coefficients are estimated by minimizing a slightly different quantity:

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

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

Ridge Regression (cont.)

- The effect of this equation is to add a shrinkage penalty of the form

$$\lambda \sum_{j=1}^{p} \beta_j^2,$$

where the tuning parameter λ is a positive value.

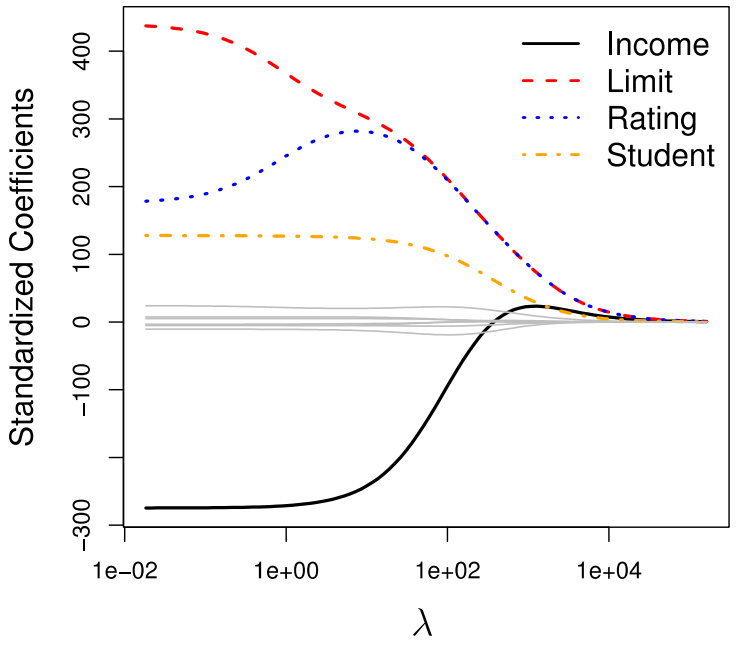
- This has the effect of shrinking the estimated beta coefficients towards zero. It turns out that such a constraint should improve the fit, because shrinking the coefficients can significantly reduce their variance.

- Note that when λ = 0, the penalty term has no effect, and ridge regression will procedure the OLS estimates. Thus, selecting a good value for λ is critical (can use cross-validation for this).

Ridge Regression (cont.)

- As λ increases, the standardized ridge regression coefficients shrinks towards zero.

- Thus, when λ is extremely large, then all of the ridge coefficient estimates are basically zero; this corresponds to the *null model* that contains no predictors.



Ridge Regression (cont.)

Computational Advantages of Ridge Regression

- If p is large, then using the best subset selection approach requires searching through enormous numbers of possible models.

- With ridge regression, for any given λ we only need to fit one model and the computations turn out to be very simple.

- Ridge regression can even be used when $p > n$, a situation where OLS fails completely (i.e., OLS estimates do not even have a unique solution).

The Lasso

- One significant problem of ridge regression is that the penalty term will never force any of the coefficients to be exactly zero.
- Thus, the final model will include all *p* predictors, which creates a challenge in model interpretation
- A more modern alternative is the *lasso*.
- The lasso works in a similar way to ridge regression, except it uses a different penalty term that shrinks some of the coefficients exactly to zero.

The Lasso (cont.)

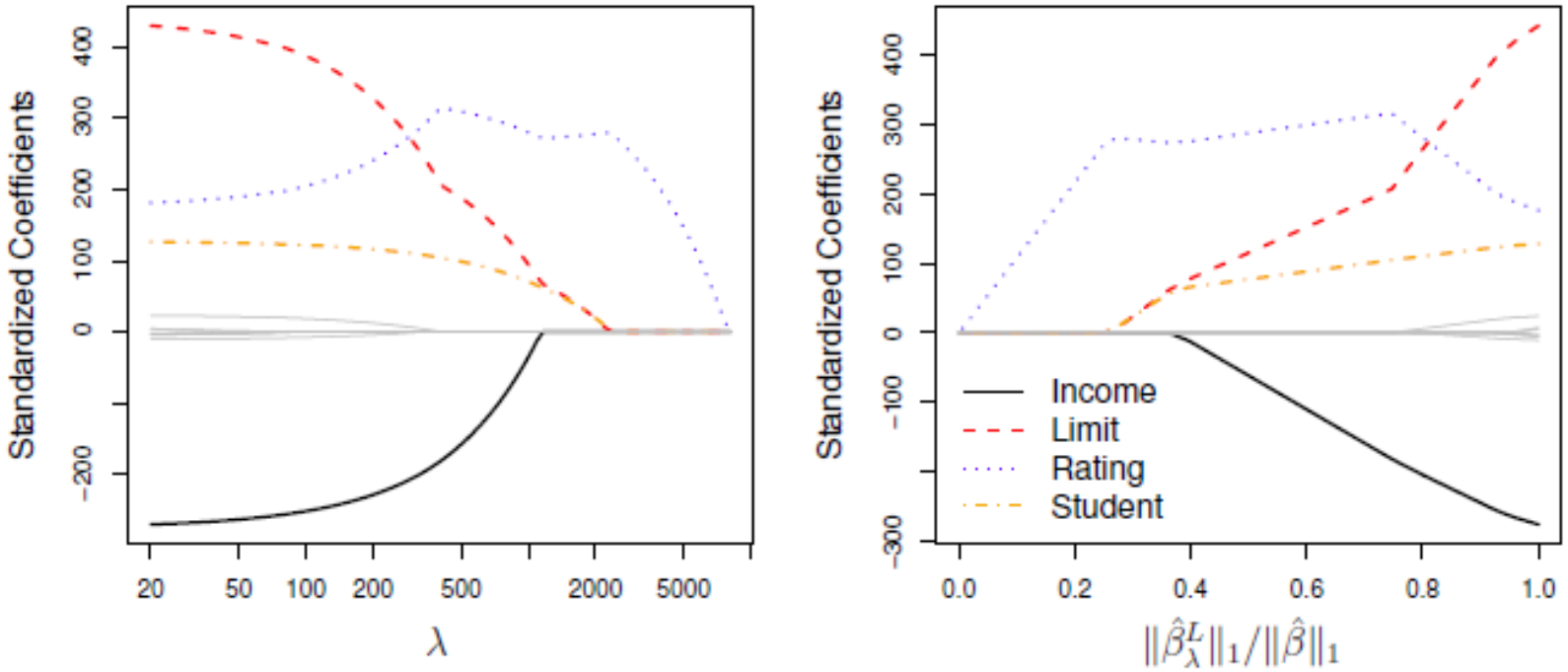
- The lasso coefficients minimize the quantity:

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

- The key difference from ridge regression is that the lasso uses an ℓ_1 penalty instead of an ℓ_2, which has the effect of forcing some of the coefficients to be exactly equal to zero when the tuning parameter λ is sufficiently large.

- Thus, the lasso performs variable/feature selection.

The Lasso (cont.)



- When λ = 0, then the lasso simply gives the OLS fit.
- When λ becomes sufficiently large, the lasso gives the null model in which all coefficient estimates equal zero.

Lasso vs. Ridge Regression

- The lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involved only a subset of predictors.
- The lasso leads to qualitatively similar behavior to ridge regression, in that as λ increases, the variance decreases and the bias increases.
- The lasso can generate more accurate predictions compared to ridge regression.
- Cross-validation can be used in order to determine which approach is better on a particular data set.

Selecting the Tuning Parameter λ

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration is best; thus, we required a method selecting a value for the tuning parameter λ or equivalently, the value of the constraint *s*.

- Select a grid of potential values; use cross-validation to estimate the error rate on test data (for each value of λ) and select the value that gives the smallest error rate.

- Finally, the model is re-fit using all of the variable observations and the selected value of the tuning parameter λ.

QUESTIONS????