Supervised Learning with Tabular Data

Gov 2018

Naijia Liu

February 2024

1. Regression with Continuous Outcome
   - OLS and Overfitting
   - Ridge
   - LASSO
   - Bias Variance Trade Off
   - Splines

Introduction

- Linear regression models play an important role in social science

Introduction

- Linear regression models play an important role in social science
- However, OLS is known to have some weaknesses
  1. Performs poorly on out-of-sample prediction when there are many features
  2. Difficulty in interpretation as the number of features grow
  3. Assumed linearity in parameters

Introduction

- Linear regression models play an important role in social science
- However, OLS is known to have some weaknesses
  1. Performs poorly on out-of-sample prediction when there are many features
  2. Difficulty in interpretation as the number of features grow
  3. Assumed linearity in parameters
- **Goal**
  - Supervised learning with continuous outcome categories
  - Bias-variance tradeoffs
  - Regularization
  - Flexible models to capture non-linearity

1. Regression with Continuous Outcome
   - OLS and Overfitting
   - Ridge
   - LASSO
   - Bias Variance Trade Off
   - Splines

Linear Regression

The linear regression model assumes that the regression function $\mathbb{E}(\mathbf{Y}|\mathbf{X})$ is linear in the sense that

$$f(\mathbf{X}) = \beta_0 + \sum_{k=1}^{K} X_k \beta_k$$

Linear Regression

The linear regression model assumes that the regression function $\mathbb{E}(\mathbf{Y}|\mathbf{X})$ is linear in the sense that

$$f(\mathbf{X}) = \beta_0 + \sum_{k=1}^{K} X_k \beta_k$$

It minimizes the residual sum of squares.

$$\begin{aligned}\mathrm{RSS}(\beta) &= \sum_{i=1}^{N} (\underbrace{Y_i - f(x_i)}_{cost})^2 \\ &= \sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{k=1}^{K} x_{ik} \beta_k \right)^2\end{aligned}$$

That is,

$$\hat{\beta} = \arg\min_{\beta} \left[(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \right]$$

Bias Variance Trade off

Let's consider fitting a higher order (linear) model on a given set of data:

$$y = \sum_{m=0}^{M} \beta_m x^m$$

Bias Variance Trade off

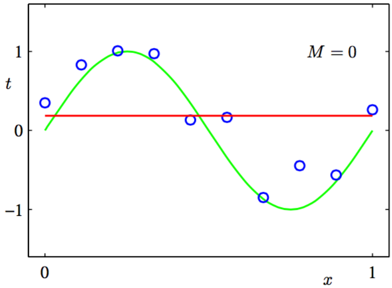
Let's consider fitting a higher order (linear) model on a given set of data:

$$y = \sum_{m=0}^{M} \beta_m x^m$$

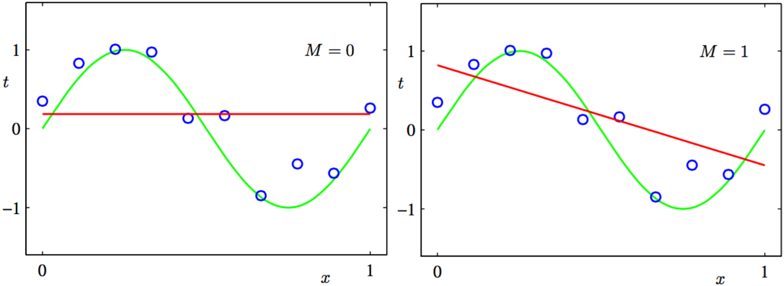
For example, if $M = 4$ we have:

$$Y = \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta 4 X^4 + \epsilon$$

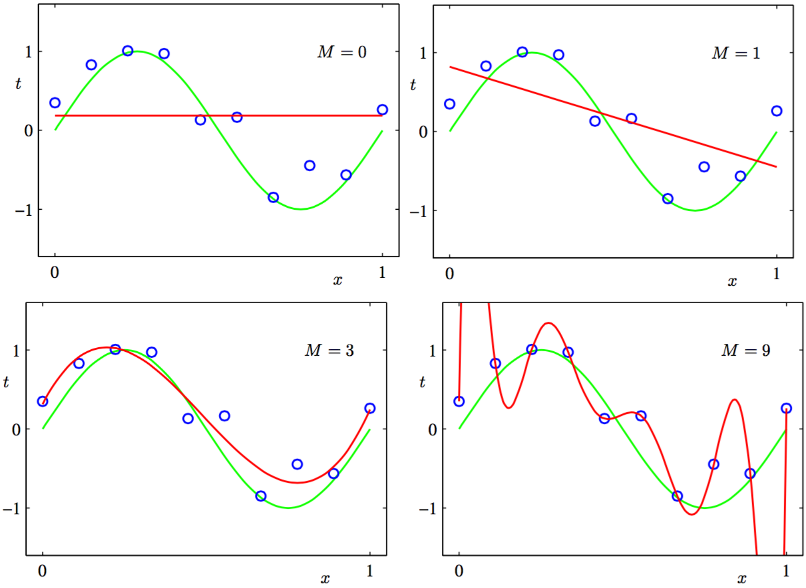
The Bias-Variance Trade-off : $y = \sum_{m=0}^{M} \beta_m x^m$



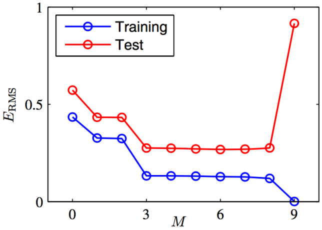
The Bias-Variance Trade-off : $y = \sum_{m=0}^{M} \beta_m x^m$



The Bias-Variance Trade-off : $y = \sum_{m=0}^{M} \beta_m x^m$

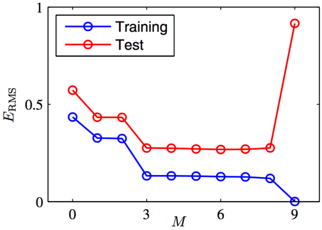


Problem of Over-fitting



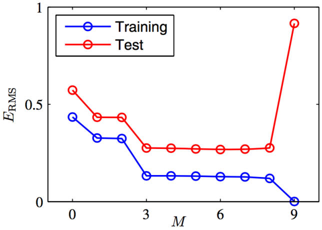
- If you have a large number of variables with a relatively small training set, you might suffer from over-fittin

Problem of Over-fitting



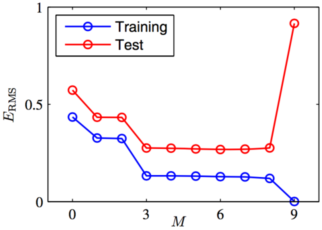
- If you have a large number of variables with a relatively small training set, you might suffer from over-fittin
- By trying to fit the training set too well, we might be fitting to noise

Problem of Over-fitting



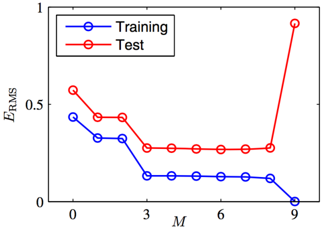
- If you have a large number of variables with a relatively small training set, you might suffer from over-fittin
- By trying to fit the training set too well, we might be fitting to noise

Problem of Over-fitting



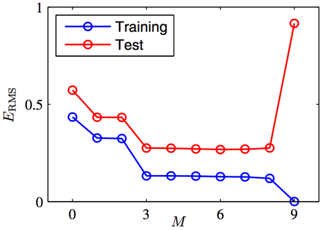
- If you have a large number of variables with a relatively small training set, you might suffer from over-fittin
- By trying to fit the training set too well, we might be fitting to noise $\rightarrow$ actually perform worse in the test set.
- Flexible models are very good at "explaining" outliers

Problem of Over-fitting



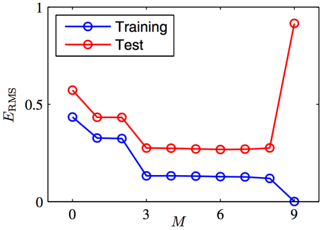
- If you have a large number of variables with a relatively small training set, you might suffer from over-fittin
- By trying to fit the training set too well, we might be fitting to noise
  $\rightarrow$ actually perform worse in the test set.
- Flexible models are very good at "explaining" outliers
- We want to penalize models that are too flexible (preference for simpler theories)

Problem of Over-fitting



- If you have a large number of variables with a relatively small training set, you might suffer from over-fittin
- By trying to fit the training set too well, we might be fitting to noise $\rightarrow$ actually perform worse in the test set.
- Flexible models are very good at "explaining" outliers
- We want to penalize models that are too flexible (preference for simpler theories)

Problem of Over-fitting



- If you have a large number of variables with a relatively small training set, you might suffer from over-fittin
- By trying to fit the training set too well, we might be fitting to noise

  $\rightarrow$ actually perform worse in the test set.
- Flexible models are very good at “explaining” outliers
- We want to penalize models that are too flexible (preference for simpler theories) while allowing for model flexibility if the data demands it

In-sample MSE vs. Out-of-sample MSE?

By construction, OLS will do well for in-sample MSE

- When $n >> k$, it will probably do well in a stable environment (i.e., observations all from the same data-generating process and effects are strong)
- When $n << k$, out-of-sample MSE might be really bad, because nothing prevents flexible models from chasing outliers (finding spurious effects)

In-sample MSE vs. Out-of-sample MSE?

By construction, OLS will do well for in-sample MSE

- When $n >> k$, it will probably do well in a stable environment (i.e., observations all from the same data-generating process and effects are strong)
- When $n << k$, out-of-sample MSE might be really bad, because nothing prevents flexible models from chasing outliers (finding spurious effects)

Two reasons why we might not be satisfied with the least squares estimates

1. **Bias-variance tradeoff:** The least squares estimates often have lower bias with larger variance $\rightarrow$ poor prediction

In-sample MSE vs. Out-of-sample MSE?

By construction, OLS will do well for in-sample MSE

- When $n >> k$, it will probably do well in a stable environment (i.e., observations all from the same data-generating process and effects are strong)
- When $n << k$, out-of-sample MSE might be really bad, because nothing prevents flexible models from chasing outliers (finding spurious effects)

Two reasons why we might not be satisfied with the least squares estimates

1. **Bias-variance tradeoff:** The least squares estimates often have lower bias with larger variance $\rightarrow$ poor prediction
2. **Interpretation:** We often include a long list of independent variables (a kitchen sink regression) $\rightarrow$ unparsimonious, difficult to interpret

Ridge Regression

- The objective:

$$\min_{\beta} \left[(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{k=1}^{K} \beta_k^2 \right]$$

Ridge Regression

- The objective:

$$\min_{\beta}\left[(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\sum_{k=1}^{K}\beta_k^2\right]$$

- Re-expressing the problem

$$\text{PRSS}(\lambda)=(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\beta^T\beta$$

Ridge Regression

- The objective:

$$\min_{\beta}\left[(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\sum_{k=1}^{K}\beta_k^2\right]$$

- Re-expressing the problem

$$\text{PRSS}(\lambda)=(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\beta^T\beta$$

$$\frac{\partial}{\partial\beta}\text{PRSS}(\lambda)=\frac{\partial}{\partial\beta}\left\{(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\beta^T\beta\right\}$$

Ridge Regression

- The objective:

$$\min_{\beta}\left[(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\sum_{k=1}^{K}\beta_k^2\right]$$

- Re-expressing the problem

$$\text{PRSS}(\lambda)=(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\beta^T\beta$$

$$\frac{\partial}{\partial\beta}\text{PRSS}(\lambda)=\frac{\partial}{\partial\beta}\left\{(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\beta^T\beta\right\}$$

To minimize the above equation, we solve for zero .

Continued

$$
\begin{aligned}
\frac{\partial}{\partial \beta}\left\{(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\beta^T\beta\right\} &= 0\\
\frac{\partial}{\partial \beta}\left\{\mathbf{y}^T\mathbf{y}+\mathbf{X}\beta^T\mathbf{X}\beta-2(\mathbf{X}\beta^T)\mathbf{y}+\lambda\beta^T\beta\right\} &= 0\\
2(\mathbf{X}^T\mathbf{X})\beta-2(\mathbf{X}^T\mathbf{y})+2\lambda\beta &= 0\\
(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})\beta &= (\mathbf{X}^T\mathbf{y})
\end{aligned}
$$

Continued

$$\frac{\partial}{\partial \beta}\left\{(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\beta^T\beta\right\}=0$$

$$\frac{\partial}{\partial \beta}\left\{\mathbf{y}^T\mathbf{y}+\mathbf{X}\beta^T\mathbf{X}\beta-2(\mathbf{X}\beta^T)\mathbf{y}+\lambda\beta^T\beta\right\}=0$$

$$2(\mathbf{X}^T\mathbf{X})\beta-2(\mathbf{X}^T\mathbf{y})+2\lambda\beta=0$$

$$(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})\beta=(\mathbf{X}^T\mathbf{y})$$

$$\hat{\beta}^{\text{ridge}}=(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Intuition on Ridge

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- When $\lambda = 0$, it is OLS.

Intuition on Ridge

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- When $\lambda = 0$, it is OLS.
- We can invert even when $(\mathbf{X}^T\mathbf{X})$ is singular!

Intuition on Ridge

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- When $\lambda = 0$, it is OLS.
- We can invert even when $(\mathbf{X}^T\mathbf{X})$ is singular!
- When $\mathbf{X}$ is orthonormal (i.e., $\mathbf{X}^T\mathbf{X} = \mathbf{I}$), the ridge estimates uniformly shrink all OLS coefficients by a factor of $\frac{1}{1+\lambda}$

Intuition on Ridge

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- When $\lambda = 0$, it is OLS.
- We can invert even when $(\mathbf{X}^T\mathbf{X})$ is singular!
- When $\mathbf{X}$ is orthonormal (i.e., $\mathbf{X}^T\mathbf{X} = \mathbf{I}$), the ridge estimates uniformly shrink all OLS coefficients by a factor of $\frac{1}{1+\lambda}$
- The objective function or Ridge regression minimizes both RSS and $\sum \beta^2$, at the same time.

How does RSS look for OLS?

Let's take an example of two dimensions, with no constant.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Here we treat β as the changing variables, become we want to compare RSS with different OLS models.

How does RSS look for OLS?

Let's take an example of two dimensions, with no constant.

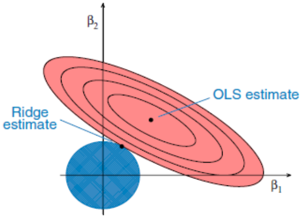
$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Here we treat β as the changing variables, become we want to compare RSS with different OLS models.

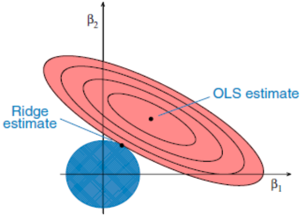
$$\begin{aligned}
\text{PRSS}_{\text{OLS}} &= \sum_{1}^{N} (y_i - \beta_1 x_{1i} - \beta_2 x_{2i})^2 \\
&= \sum_{1}^{N} a\beta_1^2 + b\beta_2^2 + c\beta_1\beta_2 + \text{constant}
\end{aligned}$$

This is how ellipse looks like on a 2 dimension coordinates!

Ridge Plot

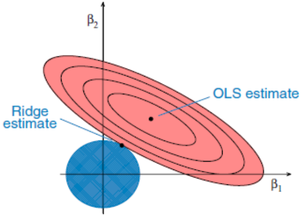


Ridge Plot



- The ellipses correspond to the contours of residual sum of squares (RSS): the inner ellipse has smaller RSS, and RSS is minimized at ordinal least square (OLS) estimates.

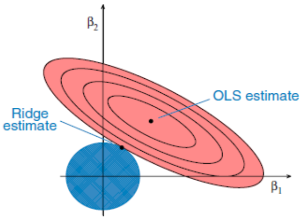
Ridge Plot



- The ellipses correspond to the contours of residual sum of squares (RSS): the inner ellipse has smaller RSS, and RSS is minimized at ordinal least square (OLS) estimates.
- For $k = 2$, the constraint in ridge regression corresponds to a circle, with radius as C

$$\sum_{j=1}^{p} = \beta_j^2 < C$$

Ridge Plot

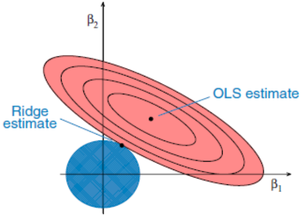


- The ellipses correspond to the contours of residual sum of squares (RSS): the inner ellipse has smaller RSS, and RSS is minimized at ordinal least square (OLS) estimates.
- For $k = 2$, the constraint in ridge regression corresponds to a circle, with radius as C

$$\sum_{j=1}^{p} = \beta_j^2 < C$$

- We are trying to minimize the ellipse size and circle simultanously in the ridge regression.

Ridge Plot

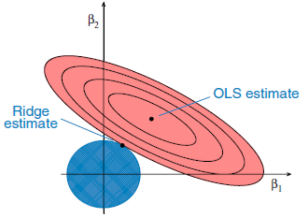


- The ellipses correspond to the contours of residual sum of squares (RSS): the inner ellipse has smaller RSS, and RSS is minimized at ordinal least square (OLS) estimates.
- For $k = 2$, the constraint in ridge regression corresponds to a circle, with radius as C

$$\sum_{j=1}^{p} = \beta_j^2 < C$$

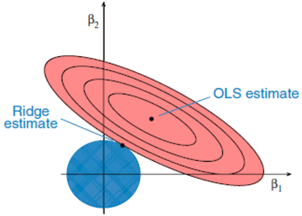
- We are trying to minimize the ellipse size and circle simultanously in the ridge regression.
- The ridge estimate is given by the point at which the ellipse and the circle touch.

Continued



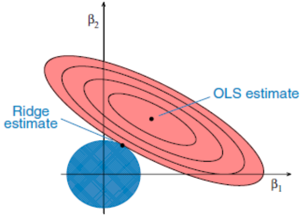
- There is a trade-off between the penalty term and RSS.

Continued



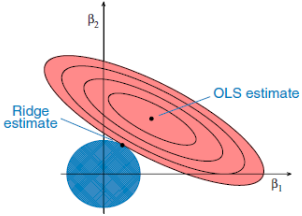
- There is a trade-off between the penalty term and RSS.
- Maybe a large β would give you a better residual sum of squares but then it will push the penalty term higher.

Continued



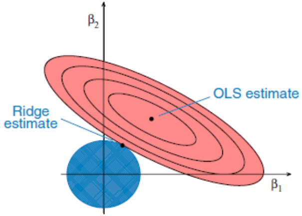
- There is a correspondence between $\frac{1}{\lambda}$ and C.

Continued



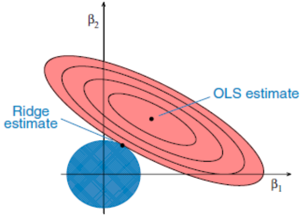
- There is a correspondence between $\frac{1}{\lambda}$ and C.
- The larger the λ is, the more you prefer the β_j's close to zero.

Continued



- There is a correspondence between $\frac{1}{\lambda}$ and C.
- The larger the λ is, the more you prefer the β_j's close to zero.
- In the extreme case when $\lambda = 0$, then you would simply be doing a normal linear regression.

Continued



- There is a correspondence between $\frac{1}{\lambda}$ and C.
- The larger the λ is, the more you prefer the β_j's close to zero.
- In the extreme case when $\lambda = 0$, then you would simply be doing a normal linear regression.
- And the other extreme as λ approaches infinity, you set all the β's to zero.

Ridge Regression: Another Perspective

Suppose you first transform each $\mathbf{x}_i$ into a more complicated $\mathbf{z}_i$, so that your $\mathbf{X}$ matrix is now a (wider) $\mathbf{Z}$ matrix. You now want to solve for the ridge coefficients β

Ridge Regression: Another Perspective

$$\text{PRSS}(\lambda) = \sum_{i=1}^{N} (\beta^\top \mathbf{z}_i - y_i)^2 + \lambda \beta^\top \beta$$

Ridge Regression: Another Perspective

$$\text{PRSS}(\lambda) = \sum_{i=1}^{N} (\beta^\top \mathbf{z}_i - y_i)^2 + \lambda \beta^\top \beta$$

Taking the derivative and setting it to zero:

$$0 = \sum_{i=1}^{N} 2(\beta^\top \mathbf{z}_i - y_i)\mathbf{z}_i + 2\lambda\beta$$

$$\beta = \sum_{i=1}^{N} \frac{-1}{\lambda}(\beta^\top \mathbf{z}_i - y_i)\mathbf{z}_i$$

Ridge Regression: Another Perspective

$$\text{PRSS}(\lambda) = \sum_{i=1}^{N} (\beta^\top \mathbf{z}_i - y_i)^2 + \lambda \beta^\top \beta$$

Taking the derivative and setting it to zero:

$$0 = \sum_{i=1}^{N} 2(\beta^\top \mathbf{z}_i - y_i)\mathbf{z}_i + 2\lambda\beta$$

$$\beta = \sum_{i=1}^{N} \frac{-1}{\lambda} (\beta^\top \mathbf{z}_i - y_i)\mathbf{z}_i$$

β can be rewritten in terms of the residuals and $\mathbf{Z}$!

Ridge Regression: Another Perspective

$$\text{PRSS}(\lambda) = \sum_{i=1}^{N} (\beta^\top \mathbf{z}_i - y_i)^2 + \lambda \beta^\top \beta$$

Taking the derivative and setting it to zero:

$$0 = \sum_{i=1}^{N} 2(\beta^\top \mathbf{z}_i - y_i)\mathbf{z}_i + 2\lambda\beta$$

$$\beta = \sum_{i=1}^{N} \frac{-1}{\lambda} (\beta^\top \mathbf{z}_i - y_i)\mathbf{z}_i$$

β can be rewritten in terms of the residuals and $\mathbf{Z}$!

Let's call the (scaled) residuals $\gamma_i = \frac{-1}{\lambda}(\beta^\top \mathbf{z}_i - y_i)$.

Ridge Regression: Another Perspective

$$\text{PRSS}(\lambda) = \sum_{i=1}^{N} (\beta^\top \mathbf{z}_i - y_i)^2 + \lambda \beta^\top \beta$$

Taking the derivative and setting it to zero:

$$0 = \sum_{i=1}^{N} 2(\beta^\top \mathbf{z}_i - y_i)\mathbf{z}_i + 2\lambda\beta$$

$$\beta = \sum_{i=1}^{N} \frac{-1}{\lambda}(\beta^\top \mathbf{z}_i - y_i)\mathbf{z}_i$$

β can be rewritten in terms of the residuals and $\mathbf{Z}$!

Let's call the (scaled) residuals $\gamma_i = \frac{-1}{\lambda}(\beta^\top \mathbf{z}_i - y_i)$. Then $\beta = \mathbf{Z}^\top \gamma$.

Ridge Regression: Another Perspective

Now let's go back to the objective function and rewrite it.

$$
\begin{aligned}
\text{PRSS}(\lambda) &= (\mathbf{Z}\beta - \mathbf{y})^\top(\mathbf{Z}\beta - \mathbf{y}) + \lambda\beta^\top\beta \\
&= (\beta^\top\mathbf{Z}^\top - \mathbf{y}^\top)(\mathbf{Z}\beta - \mathbf{y}) + \lambda\beta^\top\beta \\
&= \beta^\top\mathbf{Z}^\top\mathbf{Z}\beta - 2\beta^\top\mathbf{Z}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y} + \lambda\beta^\top\beta \\
&= \left(\gamma^\top\mathbf{Z}\right)\mathbf{Z}^\top\mathbf{Z}\left(\mathbf{Z}^\top\gamma\right) - 2\left(\gamma^\top\mathbf{Z}\right)\mathbf{Z}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y} + \lambda\left(\gamma^\top\mathbf{Z}\right)\left(\mathbf{Z}^\top\gamma\right)
\end{aligned}
$$

Continued

What is $\mathbf{Z}\mathbf{Z}^\top$?

Continued

What is $\mathbf{Z}\mathbf{Z}^\top$?

$$\mathbf{Z}\mathbf{Z}^\top = \left[\mathbf{z}_i^\top \mathbf{z}_j\right] \; = \; [\kappa(\mathbf{x}_i, \mathbf{x}_j)] = \mathbf{K}$$

We don't actually need to create $\mathbf{Z}$ to solve this problem!

Continued

What is $\mathbf{Z}\mathbf{Z}^\top$?

$$\mathbf{Z}\mathbf{Z}^\top = \left[\mathbf{z}_i^\top \mathbf{z}_j\right] \; = \; [\kappa(\mathbf{x}_i, \mathbf{x}_j)] = \mathbf{K}$$

We don't actually need to create $\mathbf{Z}$ to solve this problem!

Taking the derivative w.r.t. γ and setting it to zero:

$$0 = 2\mathbf{K}\mathbf{K}\gamma - 2\mathbf{K}\mathbf{y} + 2\lambda\mathbf{K}\gamma$$
$$\gamma = (\mathbf{K} + \lambda\mathbf{I})^{-1}\,\mathbf{y}$$

and we can plug this back in to find the coefficients of interest, β

1. Regression with Continuous Outcome
   - OLS and Overfitting
   - Ridge
   - LASSO
   - Bias Variance Trade Off
   - Splines

LASSO : Least Absolute Shrinkage and Selection Operator

- Limitations of OLS
  1. Prediction Accuracy: large variance (with low bias)
  2. Interpretation: Large number of predictors (ridge regression shrinks, but does not set any coefficients to zero)

LASSO : Least Absolute Shrinkage and Selection Operator

- Limitations of OLS
  1. Prediction Accuracy: large variance (with low bias)
  2. Interpretation: Large number of predictors (ridge regression shrinks, but does not set any coefficients to zero)
- Lasso
  - The objective:

$$\min_{\beta} \left[\frac{1}{2}(y - X\beta)^T (y - X\beta) + \lambda \sum_{k=1}^{K} \mid \beta_k \mid \right] \tag{1}$$

- The first term in this objective function is the residual sum of a squares.
- The second term has two components: the tuning parameter λ, indexed by sample size, and the penalty term $|\widetilde{\beta}|$.

Lasso with a single covariate

- Take as observed data an outcome Y_i for $i \in \{1, 2, \ldots, N\}$, and a single observed covariate, X_i with associated parameter β^o. We assume the data are generated as

$$Y_i = X_i\beta^o + \epsilon_i$$

Lasso with a single covariate

- Take as observed data an outcome Y_i for $i \in \{1, 2, \ldots, N\}$, and a single observed covariate, X_i with associated parameter β^o. We assume the data are generated as

$$Y_i = X_i \beta^o + \epsilon_i$$

- For simplicity: we scale $\frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{N}\sum_{i=1}^{N} Y_i = 0$ and $\sum_{i=1}^{N} X_i^2 = N - 1$, so X_i has sample standard deviation one.

Lasso with a single covariate

- Take as observed data an outcome Y_i for $i \in \{1, 2, \ldots, N\}$, and a single observed covariate, X_i with associated parameter β^o. We assume the data are generated as

$$Y_i = X_i \beta^o + \epsilon_i$$

- For simplicity: we scale $\frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{N}\sum_{i=1}^{N} Y_i = 0$ and $\sum_{i=1}^{N} X_i^2 = N - 1$, so X_i has sample standard deviation one.
- We assume the error is mean-zero, equivariant, and that all fourth moments of $[Y_i, X_i]$ exist.

Continued

- Under the setup, we will denote the least squares estimate as

$$\widehat{\beta}^{LS} = \frac{\sum_{i=1}^{N} Y_i X_i}{\sum_{i=1}^{N} X_i^2} = \frac{\sum_{i=1}^{N} Y_i X_i}{N-1}$$

Lasso with a single covariate

- Let's consider LASSO in the case with a single covariate.

$$\widehat{\beta}^L = \arg\min_{\widetilde{\beta}} \frac{1}{2} \sum_{i=1}^{N} (Y_i - X_i \widetilde{\beta})^2 + \lambda |\widetilde{\beta}|$$

Lasso with a single covariate

- Let's consider LASSO in the case with a single covariate.

$$\widehat{\beta}^L = \arg\min_{\widetilde{\beta}} \frac{1}{2} \sum_{i=1}^{N} (Y_i - X_i \widetilde{\beta})^2 + \lambda |\widetilde{\beta}|$$

- We take the partial with regard to β, because we are looking for the best β.

$$\frac{\partial}{\partial \beta} = (Y - X\beta)(-X) + ??? = 0$$

Lasso with a single covariate

- Let's consider LASSO in the case with a single covariate.

$$\widehat{\beta}^L = \arg\min_{\widetilde{\beta}} \frac{1}{2} \sum_{i=1}^{N} (Y_i - X_i \widetilde{\beta})^2 + \lambda |\widetilde{\beta}|$$

- We take the partial with regard to β, because we are looking for the best β.

$$\frac{\partial}{\partial \beta} = (Y - X\beta)(-X) + ??? = 0$$

- For simplicity, we will say $\lambda \geq 0$.

Continued

- Let's now consider a certain two scenarios:

Continued

- Let's now consider a certain two scenarios:
  - If $\beta > 0$:

$$
\begin{aligned}
(Y - X\beta)(-X) + \lambda &= 0 \\
-XY + X^2\beta + \lambda &= 0 \\
\beta &= \frac{XY - \lambda}{X^2} \\
&= \frac{XY - \lambda}{N - 1} && \text{because we scaled} X^2 \\
&= \beta^{LS} - \frac{\lambda}{N - 1}
\end{aligned}
$$

Continued

- Let's now consider a certain two scenarios:
  - If $\beta > 0$:

$$\begin{aligned}
(Y - X\beta)(-X) + \lambda &= 0 \\
-XY + X^2\beta + \lambda &= 0 \\
\beta &= \frac{XY - \lambda}{X^2} \\
&= \frac{XY - \lambda}{N - 1} && \text{because we scaled} X^2 \\
&= \beta^{LS} - \frac{\lambda}{N - 1}
\end{aligned}$$

- Since we assumed $\beta > 0$, and we know $\frac{\lambda}{N-1} > 0$, it would be weird if $\beta^{LS} < \frac{\lambda}{N-1}$

Continued

- Let's now consider a certain two scenarios:
  - If $\beta > 0$:

$$
\begin{aligned}
(Y - X\beta)(-X) + \lambda &= 0 \\
-XY + X^2\beta + \lambda &= 0 \\
\beta &= \frac{XY - \lambda}{X^2} \\
&= \frac{XY - \lambda}{N - 1} && \text{because we scaled} X^2 \\
&= \beta^{LS} - \frac{\lambda}{N - 1}
\end{aligned}
$$

- Since we assumed $\beta > 0$, and we know $\frac{\lambda}{N-1} > 0$, it would be weird if $\beta^{LS} < \frac{\lambda}{N-1}$
- When that happens, we will shrink β to zero instead.

Continued

- Similarly if $\beta \leq 0$:

$$
\begin{aligned}
(Y - X\beta)(-X) - \lambda &= 0 \\
-XY + X^2\beta - \lambda &= 0 \\
\beta &= \frac{XY + \lambda}{X^2} \\
&= \frac{XY + \lambda}{N - 1} \qquad \text{because we scaled } X^2 \\
&= \beta^{LS} + \frac{\lambda}{N - 1}
\end{aligned}
$$

Continued

- Similarly if $\beta \leq 0$:

$$
\begin{aligned}
(Y - X\beta)(-X) - \lambda &= 0 \\
-XY + X^2\beta - \lambda &= 0 \\
\beta &= \frac{XY + \lambda}{X^2} \\
&= \frac{XY + \lambda}{N - 1} && \text{because we scaled } X^2 \\
&= \beta^{LS} + \frac{\lambda}{N - 1}
\end{aligned}
$$

- Since we assumed $\beta \leq 0$, and we know $\frac{\lambda}{N-1} > 0$, it would be weird if $\beta^{LS} + \frac{\lambda}{N-1} > 0$

Continued

- Similarly if $\beta \leq 0$:

$$
\begin{aligned}
(Y - X\beta)(-X) - \lambda &= 0 \\
-XY + X^2\beta - \lambda &= 0 \\
\beta &= \frac{XY + \lambda}{X^2} \\
&= \frac{XY + \lambda}{N - 1} && \text{because we scaled } X^2 \\
&= \beta^{LS} + \frac{\lambda}{N - 1}
\end{aligned}
$$

- Since we assumed $\beta \leq 0$, and we know $\frac{\lambda}{N-1} > 0$, it would be weird if $\beta^{LS} + \frac{\lambda}{N-1} > 0$
- When that happens, we will shrink β to zero instead.

Combine the Two

- Denote the sign of the least squares estimate as $\widehat{s}^L = \text{sign}(\widehat{\beta}) \in \{-1, 1\}$.

Combine the Two

- Denote the sign of the least squares estimate as $\widehat{s}^L = \text{sign}(\widehat{\beta}) \in \{-1, 1\}$.

  With one parameter, the LASSO estimate is (Tibshirani 1996, sec 2.2)

$$\widehat{\beta}^L = \left(\widehat{\beta}^{LS} - \widehat{s}^L \frac{\lambda}{N-1}\right) \mathbf{1}\left(|\widehat{\beta}^{LS}| > \frac{\lambda}{N-1}\right)$$

Combine the Two

- Denote the sign of the least squares estimate as $\widehat{s}^L = \text{sign}(\widehat{\beta}) \in \{-1, 1\}$.

  With one parameter, the LASSO estimate is (Tibshirani 1996, sec 2.2)

$$\widehat{\beta}^L = \left(\widehat{\beta}^{LS} - \widehat{s}^L \frac{\lambda}{N-1}\right) \mathbf{1}\left(|\widehat{\beta}^{LS}| > \frac{\lambda}{N-1}\right)$$

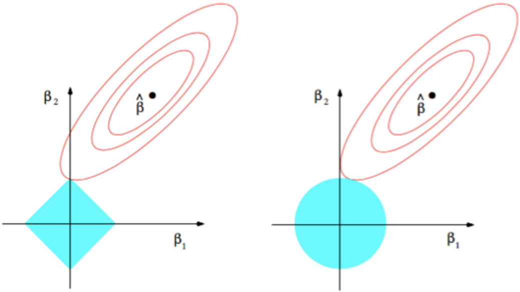
- For those variables with a relatively small OLS coefficient, we shrink them to zero.
- Rest of the variables, we shrink the size.

Lasso Plot

Similarly, Lasso plot consists of a square, because we minimize the RSS and absolute value of coefficients.

Question: Try drawing in R:

$$|\beta_1| + |\beta_2| \leq C$$



When should we use LASSO?

- Advantages

When should we use LASSO?

- Advantages
  - LASSO works well for prediction when the true model is "sparse" (i.e., only a few variables really matter)

When should we use LASSO?

- Advantages
  - LASSO works well for prediction when the true model is "sparse" (i.e., only a few variables really matter)
  - Post-LASSO will give you asymptotically valid confidence interval

When should we use LASSO?

- Advantages
  - LASSO works well for prediction when the true model is "sparse" (i.e., only a few variables really matter)
  - Post-LASSO will give you asymptotically valid confidence interval
  - LASSO is designed for models that start with many parameters ("wide" data)

When should we use LASSO?

- Advantages
  - LASSO works well for prediction when the true model is "sparse" (i.e., only a few variables really matter)
  - Post-LASSO will give you asymptotically valid confidence interval
  - LASSO is designed for models that start with many parameters ("wide" data)
  - Prediction accuracy

When should we use LASSO?

- Advantages
  - LASSO works well for prediction when the true model is "sparse" (i.e., only a few variables really matter)
  - Post-LASSO will give you asymptotically valid confidence interval
  - LASSO is designed for models that start with many parameters ("wide" data)
  - Prediction accuracy
  - Interpretation with sparsity

- Disadvantages
  - LASSO won't work when there are a lot of variables that actually matter (ridge works better in that case)
  - With high collinearity, the LASSO arbitrarily selects only one among highly correlated variables (fine if goal is prediction)
  - You will get a completely different coefficient estimate "chosen" by LASSO with a slightly different sample, but predictions will be similar
  - This is why you need to be really careful about interpreting coefficients (remember that LASSO aims to optimally predict out-of-sample)

Statistical Inference with LASSO

We often care about confidence intervals for $\hat{\beta}$

Statistical Inference with LASSO

We often care about confidence intervals for $\hat{\beta}$

1. Post-Lasso (Belloni and Chernozhukov (2013))

Statistical Inference with LASSO

We often care about confidence intervals for $\hat{\beta}$

1. Post-Lasso (Belloni and Chernozhukov (2013))
    - Two-step estimation that will give a consistent estimate under some conditions ("approximate sparsity" assumption: the truth is simple)

Statistical Inference with LASSO

We often care about confidence intervals for $\hat{\beta}$

1. Post-Lasso (Belloni and Chernozhukov (2013))
    - Two-step estimation that will give a consistent estimate under some conditions ("approximate sparsity" assumption: the truth is simple)
    - After "hard thresholding" with LASSO, take the surviving features and run OLS. (Uses LASSO to choose variables, but OLS to get the right effects instead of shrinking them

Statistical Inference with LASSO

We often care about confidence intervals for $\hat{\beta}$

1. Post-Lasso (Belloni and Chernozhukov (2013))
   - Two-step estimation that will give a consistent estimate under some conditions ("approximate sparsity" assumption: the truth is simple)
   - After "hard thresholding" with LASSO, take the surviving features and run OLS. (Uses LASSO to choose variables, but OLS to get the right effects instead of shrinking them
2. Covariance test: Lockhart, Taylor, Tibshirani (2014)

Statistical Inference with LASSO

We often care about confidence intervals for $\hat{\beta}$

1. Post-Lasso (Belloni and Chernozhukov (2013))
   - Two-step estimation that will give a consistent estimate under some conditions (“approximate sparsity” assumption: the truth is simple)
   - After “hard thresholding” with LASSO, take the surviving features and run OLS. (Uses LASSO to choose variables, but OLS to get the right effects instead of shrinking them
2. Covariance test: Lockhart, Taylor, Tibshirani (2014)
   - p-value for each variable as it is added to lasso model

Oracle Inequality (Optional)

- The single-parameter least squares estimator achieves predictive error

$$E\left\{\frac{1}{N}\sum_{i=1}\left(X_i(\widehat{\beta}^{LS}-\beta^o)\right)^2\right\}=\sigma^2/N.$$

  A LASSO estimator satisfies the Oracle Inequality if it achieves a prediction rate similar to that of the OLS estimator, were the true model known in advance.

- Oracle Inequality in the Single-Parameter Case
  An estimator satisfies the Oracle Inequality if

$$\frac{1}{N}\sum_{i=1}\left(X_i\left(\widehat{\beta}^L(\lambda_N)-\beta^o\right)\right)^2\leq C\frac{\sigma^2}{N}$$

  with a high probability for some constant $C>0$.

- Specifically, an estimator satisfies the Oracle Inequality if we can bound the loss function, with high probability, at a rate going to zero at $1/N$. We next show that if λ_N grows as $\sqrt{N}$, it will satisfiy the Oracle Inequality.
- Oracle Inequaltiy for the LASSO in the Single Parameter Case:

  For $\lambda_N = \sigma \times t \times \sqrt{(N-1)}$, the single-parameter LASSO estimator satistfies the Oracle Inequality

$$\frac{1}{N}\left\{\sum_{i=1}\left(X_i(\widehat{\beta}^L(\lambda_N)-\beta^o)\right)^2+\lambda_N\left|\widehat{\beta}^L(\lambda_N)-\beta^o\right|\right\}\leq\frac{20\sigma^2t^2}{N-1}$$

  with probability at least $1-2\exp\left\{-t^2/2\right\}$.
- Proof as the bonus question for Pset I

Variance and Bias Trade-off

Variance and Bias Trade-off

- Let's take a closer look mean squared error, to mathematically capture the trade off.

Variance and Bias Trade-off

- Let's take a closer look mean squared error, to mathematically capture the trade off.

$$
\begin{aligned}
\text{MSE} &= \mathbf{E}\left((\hat{\beta}-\beta)^2\right) \\
&= \mathbf{E}\left((\underbrace{\hat{\beta}-\mathbf{E}(\hat{\beta})}_{A} + \underbrace{\mathbf{E}(\hat{\beta})-\beta}_{B})^2\right) \\
&= \mathbf{E}\left(A^2 + B^2 + 2AB\right) \\
&= \underbrace{\mathbf{E}\left((\hat{\beta}-\mathbf{E}(\hat{\beta}))^2\right)}_{\text{variance}} + \mathbf{E}\underbrace{\left(\mathbf{E}(\hat{\beta})-\beta)\right)^2}_{\text{bias}^2} + 2\mathbf{E}\left((\hat{\beta}-\mathbf{E}(\hat{\beta}))(\mathbf{E}(\hat{\beta})-\beta)\right)
\end{aligned}
$$

Variance and Bias

- Let's first take a look at the cross term.

Variance and Bias

- Let's first take a look at the cross term.

Variance and Bias

- Let's first take a look at the cross term.

$$
\begin{aligned}
&\mathbf{E}\left((\hat{\beta} - \mathbf{E}(\hat{\beta}))(\mathbf{E}(\hat{\beta}) - \beta)\right) \\
&= \mathbf{E}\left(\hat{\beta}\mathbf{E}(\hat{\beta}) - \mathbf{E}(\hat{\beta})\mathbf{E}(\hat{\beta}) - \hat{\beta}\beta + \mathbf{E}(\hat{\beta})\beta\right) \\
&= \mathbf{E}(\hat{\beta})\mathbf{E}(\hat{\beta}) - \mathbf{E}(\hat{\beta})\mathbf{E}(\hat{\beta}) - \beta\mathbf{E}(\hat{\beta}) + \beta\mathbf{E}(\hat{\beta}) \\
&= 0
\end{aligned}
$$

Variance and Bias

$$
\begin{aligned}
\text{MSE} &= \underbrace{\mathbf{E}\left((\hat{\beta} - \mathbf{E}(\hat{\beta}))^2\right)}_{\text{variance}} + \underbrace{\mathbf{E}\left(\mathbf{E}(\hat{\beta}) - \beta)\right)^2}_{\text{bias}^2} \\
&= \underbrace{\mathbf{E}\left((\hat{\beta} - \mathbf{E}(\hat{\beta}))^2\right)}_{\text{variance}} + \underbrace{\left(\mathbf{E}(\hat{\beta}) - \beta\right)^2}_{\text{bias}^2}
\end{aligned}
$$

Piecewise Polynomials

- The conditional expectation function $E[Y|X = x]$ is often (usually?) nonlinear

Piecewise Polynomials

- The conditional expectation function $E[Y|X = x]$ is often (usually?) nonlinear
- One option: LASSO and ridge could be used to capture nonlinearities with predefined basis functions (e.g., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_1 \cdot X_2 + \cdots + \beta_k X_1^k + \epsilon$)

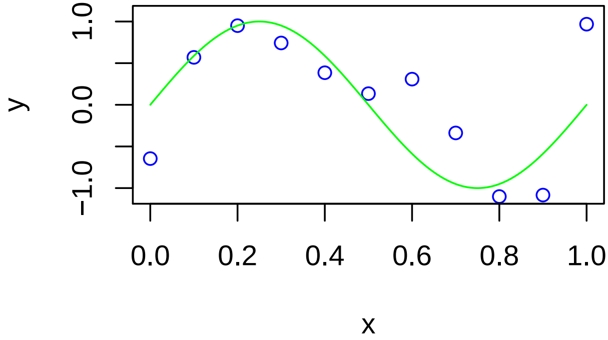
Piecewise Polynomials

- The conditional expectation function $E[Y|X = x]$ is often (usually?) nonlinear
- One option: LASSO and ridge could be used to capture nonlinearities with predefined basis functions (e.g., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_1 \cdot X_2 + \cdots + \beta_k X_1^k + \epsilon$)
- Here, we consider basic splines

Piecewise Polynomials

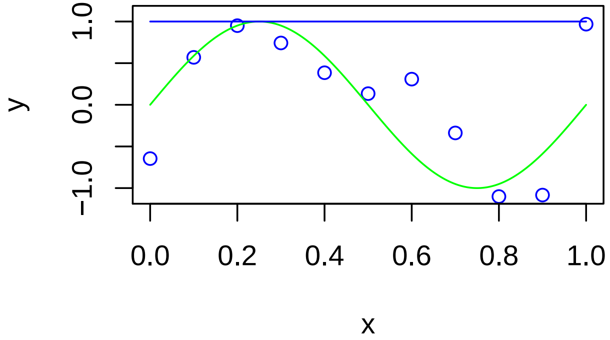
- The conditional expectation function $E[Y|X = x]$ is often (usually?) nonlinear
- One option: LASSO and ridge could be used to capture nonlinearities with predefined basis functions (e.g., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_1 \cdot X_2 + \cdots + \beta_k X_1^k + \epsilon$)
- Here, we consider basic splines
- Other options for capturing nonlinearity include weighted moving average and generalizations (LOESS)

Piecewise Polynomials



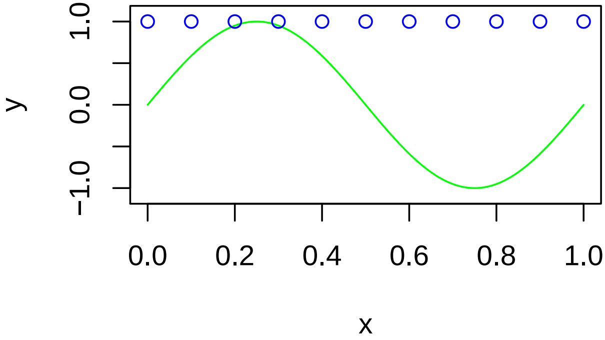
True DGP: $Y_i = \sin(2\pi X_i) + \varepsilon$

Piecewise Polynomials



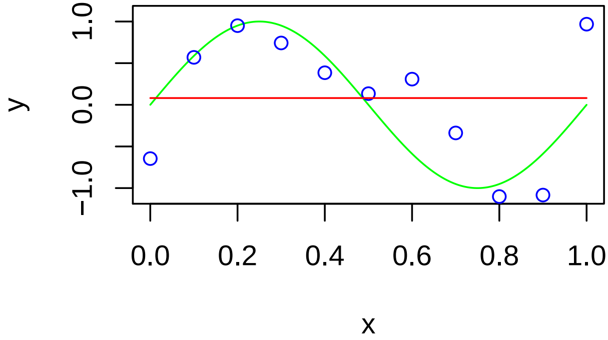
Consider the *constant* basis function $f_1(x) = 1$

Piecewise Polynomials



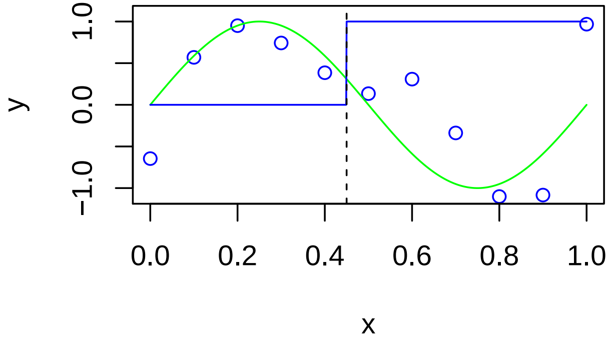
For observation i, this creates a feature $f_1(X_i)$. Prediction performance is not ideal.

Piecewise Polynomials



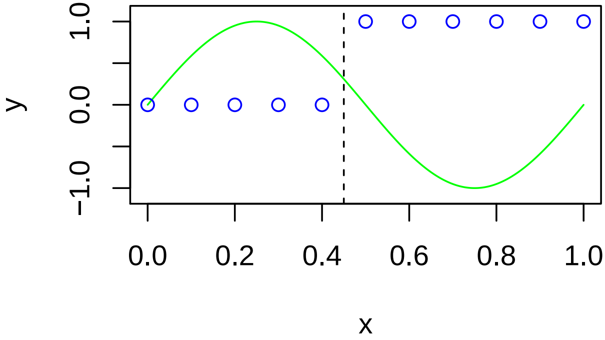
If one were to run a linear regression with $f(x)$ and y, you will get the red line. OLS is confused because $f(x) = 1$ always.

Piecewise Polynomials



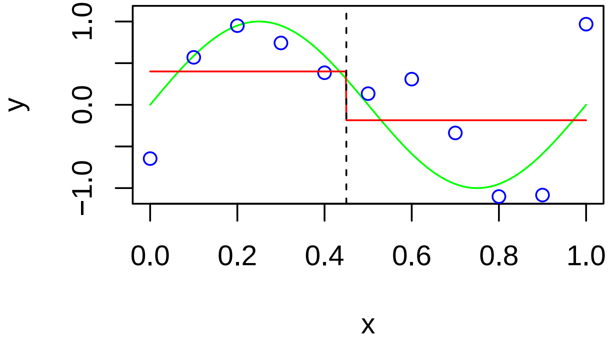
We could add a second *piecewise constant* basis function $f_2(x) = 1(x > \xi)$, with a discontinuity at some *knot*, ξ

Piecewise Polynomials



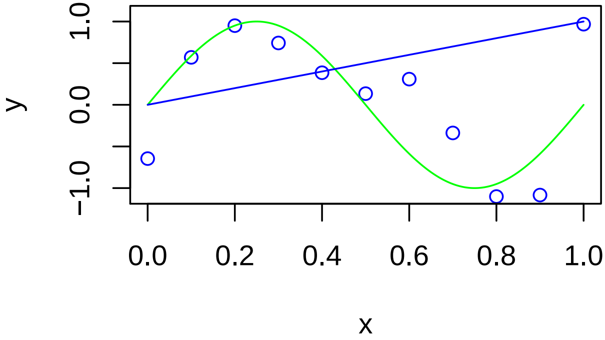
This would produce a second feature in the data matrix, prediction performance is slightly better.

Piecewise Polynomials



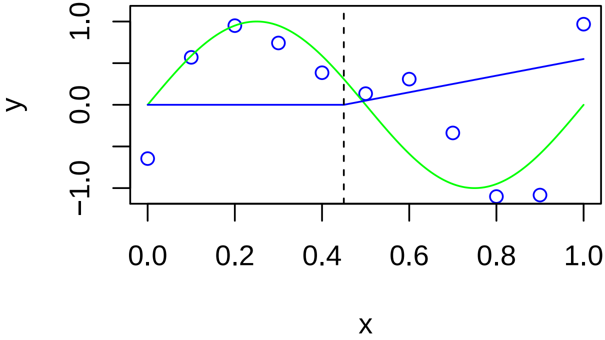
With this expanded basis set, a richer set of approximating functions could be constructed from $\beta_1 f_1(x) + \beta_2 f_2(x)$. The one that minimizes MSE is plotted in red here.

Piecewise Polynomials



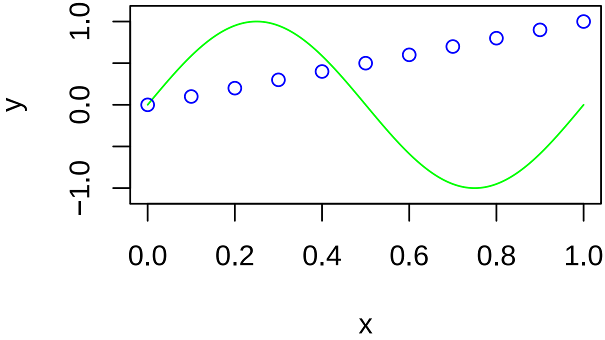
Higher order basis functions can be added, e.g. the linear function $g_1(x) = x$...

Piecewise Polynomials



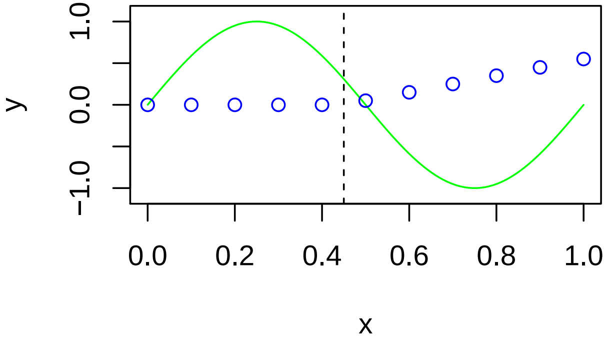
... or the continuous and piecewise linear $g_2(x) = (x - \xi) \cdot 1(x > \xi)$

Piecewise Polynomials



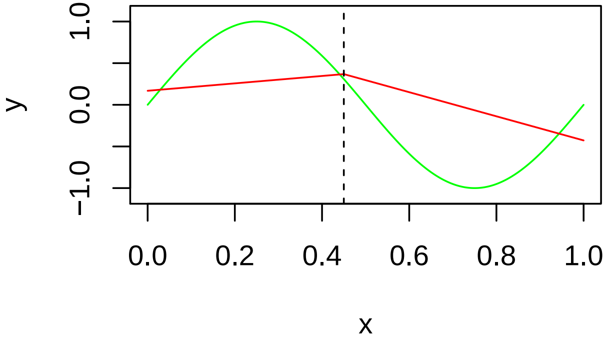
OLS will predict the points like this.

Piecewise Polynomials



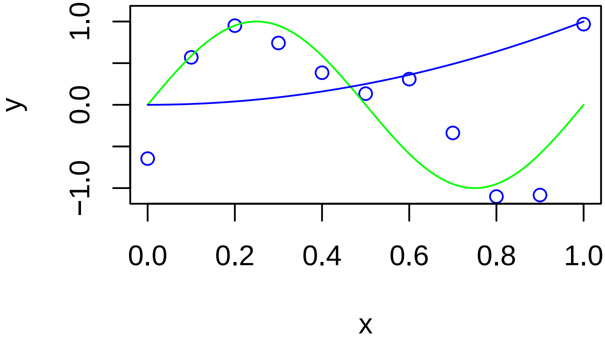
Piecewise function will predict the points like this.

Piecewise Polynomials



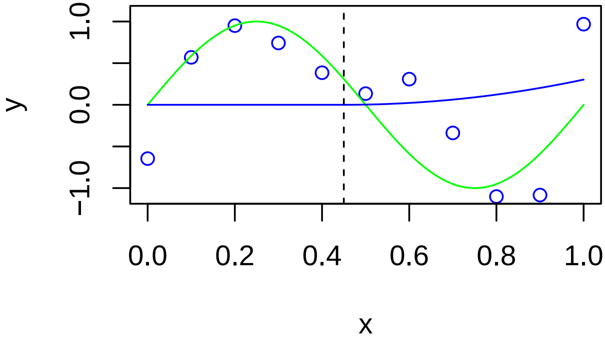
From $f(x) = 1$, $g_1(x) = x$, and $g_2(x) = (x - \xi) \cdot 1(x > \xi)$, many approximating functions of the form $\alpha f(x) + \beta_1 g_1(x) + \beta_2 g_2(x)$ can be constructed for the true conditional expectation—all of which are continuous, but have discontinuous first derivatives.

Piecewise Polynomials



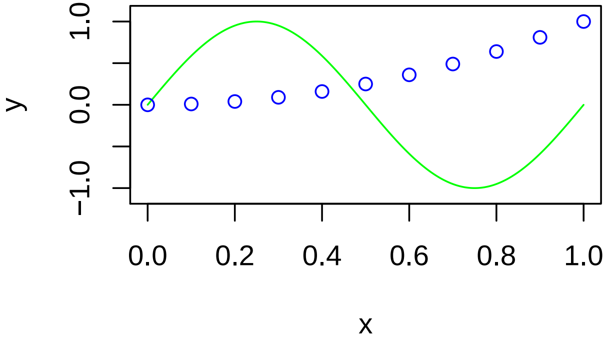
Further basis functions $h_1(x) = x^2$ and $h_2(x) = (x - \xi)^2 \cdot 1(x > \xi)$ and the corresponding features

Piecewise Polynomials



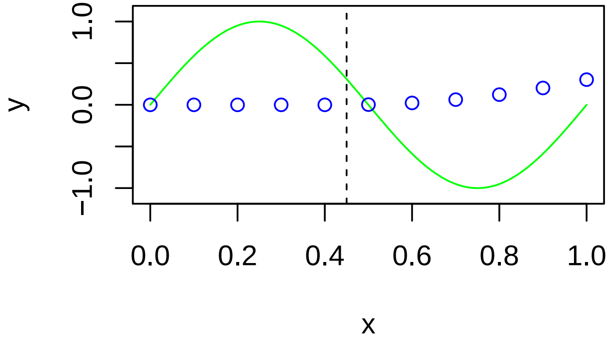
Further basis functions $h_1(x) = x^2$ and $h_2(x) = (x - \xi)^2 \cdot 1(x > \xi)$ and the corresponding features

Piecewise Polynomials



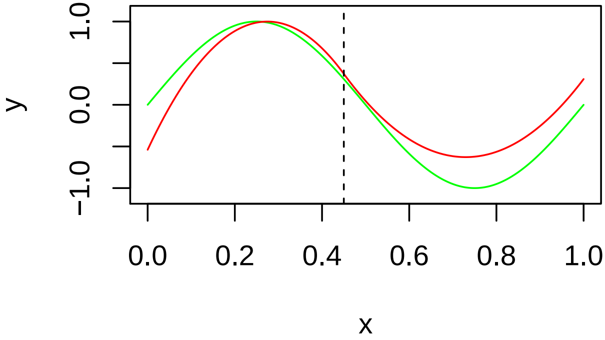
Further basis functions $h_1(x) = x^2$ and $h_2(x) = (x - \xi)^2 \cdot 1(x > \xi)$ and the corresponding features

Piecewise Polynomials



Further basis functions $h_1(x) = x^2$ and $h_2(x) = (x - \xi)^2 \cdot 1(x > \xi)$ and the corresponding features

Piecewise Polynomials



An function of the form $\alpha f(x) + \beta g(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x)$. Observe that it is both continuous and has a continuous first derivative.

Cubic Splines

- A cubic spline extends this idea to create functions that have continuous first and second derivatives

Cubic Splines

- A cubic spline extends this idea to create functions that have continuous first and second derivatives
  - The basis set consists of $f_0(x) = x^0$, $f_1(x) = x^1$, $f_2(x) = x^2$, and piecewise cubic terms $f_3(x) = x^3$, $f_k(x) = (x - \xi_k)^3 1(x > \xi_k)$, $\cdots$

Cubic Splines

- A cubic spline extends this idea to create functions that have continuous first and second derivatives
  - The basis set consists of $f_0(x) = x^0$, $f_1(x) = x^1$, $f_2(x) = x^2$, and piecewise cubic terms $f_3(x) = x^3$, $f_k(x) = (x - \xi_k)^3 1(x > \xi_k)$, $\cdots$
  - Many knots can be used: spaced equally, at quantiles of X, etc.

Cubic Splines

- A cubic spline extends this idea to create functions that have continuous first and second derivatives
  - The basis set consists of $f_0(x) = x^0$, $f_1(x) = x^1$, $f_2(x) = x^2$, and piecewise cubic terms $f_3(x) = x^3$, $f_k(x) = (x - \xi_k)^3 1(x > \xi_k)$, $\cdots$
  - Many knots can be used: spaced equally, at quantiles of X, etc.
- *Natural* cubic splines force the outermost regions to be linear

Cubic Splines

- A cubic spline extends this idea to create functions that have continuous first and second derivatives
  - The basis set consists of $f_0(x) = x^0$, $f_1(x) = x^1$, $f_2(x) = x^2$, and piecewise cubic terms $f_3(x) = x^3$, $f_k(x) = (x - \xi_k)^3 1(x > \xi_k)$, $\cdots$
  - Many knots can be used: spaced equally, at quantiles of X, etc.
- *Natural* cubic splines force the outermost regions to be linear

Cubic Splines

- A cubic spline extends this idea to create functions that have continuous first and second derivatives
  - The basis set consists of $f_0(x) = x^0$, $f_1(x) = x^1$, $f_2(x) = x^2$, and piecewise cubic terms $f_3(x) = x^3$, $f_k(x) = (x - \xi_k)^3 1(x > \xi_k)$, $\cdots$
  - Many knots can be used: spaced equally, at quantiles of X, etc.
- *Natural* cubic splines force the outermost regions to be linear (reduces overfitting near the boundary, where there are no additional knots to constrain)

Cubic Smoothing Splines

- In the extreme, a natural cubic spline can be used with one knot at every value of X. Call this $g(x) = \sum_k \beta_k f_k(x)$.

Cubic Smoothing Splines

- In the extreme, a natural cubic spline can be used with one knot at every value of X. Call this $g(x) = \sum_k \beta_k f_k(x)$.
- To address overfitting when estimating β_k, a penalty is applied to functions with large second derivatives (high curvature)

$$\sum_{i=1}^{N} (Y_i - g(X_i))^2 + \lambda_k \int g''(z)^2 dz$$

where z sweeps over all possible values that X_i can take on

Cubic Smoothing Splines

- In the extreme, a natural cubic spline can be used with one knot at every value of X. Call this $g(x) = \sum_k \beta_k f_k(x)$.
- To address overfitting when estimating β_k, a penalty is applied to functions with large second derivatives (high curvature)

$$\sum_{i=1}^{N} (Y_i - g(X_i))^2 + \lambda_k \int g''(z)^2 dz$$

where z sweeps over all possible values that X_i can take on

Cubic Smoothing Splines

- In the extreme, a natural cubic spline can be used with one knot at every value of X. Call this $g(x) = \sum_k \beta_k f_k(x)$.
- To address overfitting when estimating β_k, a penalty is applied to functions with large second derivatives (high curvature)

$$\sum_{i=1}^{N} (Y_i - g(X_i))^2 + \lambda_k \int g''(z)^2 dz$$

where z sweeps over all possible values that X_i can take on (what happens to this integral outside the outermost knots?)