Lecture 2: Linear Regression Models

Statistical Learning and Data Mining

Xingye Qiao

Department of Mathematical Sciences
Binghamton University

E-mail: qiao@math.binghamton.edu

Read: ISR Ch. 3 and ESLII Chs. 3.1–3.2

Outline

The next section would be ......

1. Model Continuous Responses

2. Linear Regression

Loss and risk

- 0-1 loss: $\mathbb{1}\{\hat{y} \neq y\}$. Suitable for classification.
- Risk for 0-1 loss: $P(\hat{Y} \neq Y) = \mathrm{E}(\mathbb{1}\{\hat{Y} \neq Y\})$.
- What if the response is continuous?
- Squared error loss: suppose we use statistic Z to estimate parameter μ

$$(Z - \mu)^2$$

- Risk = mean squared error = MSE

$$\mathrm{E}_Z(Z - \mu)^2$$

Bias-Variance Decomposition

The Bias-Variance Decomposition is a phenomenon often seen. For example, suppose we use statistic Z to estimate μ and consider the mean squared error.

$$
\begin{aligned}
&\mathsf{E}(Z-\mu)^2 \\
=&\mathsf{E}(Z-\mathsf{E}Z+\mathsf{E}Z-\mu)^2 \\
=&\mathsf{E}(Z-\mathsf{E}Z)^2+(\mathsf{E}Z-\mu)^2+2\times\mathsf{E}[(Z-\mathsf{E}Z)(\mathsf{E}Z-\mu)] \\
=&\mathsf{E}(Z-\mathsf{E}Z)^2+(\mathsf{E}Z-\mu)^2+2\times[(\mathsf{E}Z-\mathsf{E}Z)(\mathsf{E}Z-\mu)] \\
=&\underbrace{\mathsf{E}(Z-\mathsf{E}Z)^2}_{\text{Variance of } Z}+\underbrace{(\mathsf{E}Z-\mu)^2}_{\text{Bias}^2}
\end{aligned}
$$

The next section would be ......

Framework

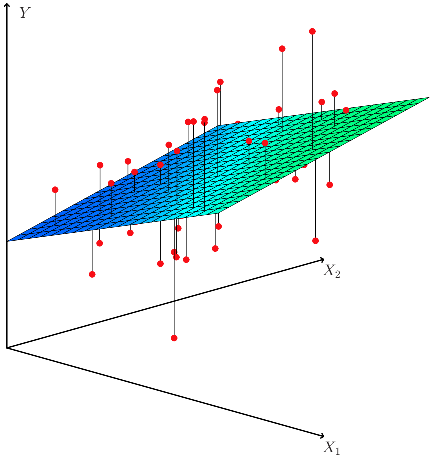
- Suppose we know the true distribution F.
- Then $h_B(x) = \mathrm{E}(Y|X = x)$ minimizes $\mathrm{E}(Y - h(X))^2$ among all measurable function $h(x)$.
- Especially,
  $\mathrm{E}(Y - h(X))^2 = \underbrace{\mathrm{E}(Y - \mathrm{E}(Y|X))^2}_{\text{“conditional variance”}} + (\mathrm{E}(Y|X) - h(X))^2$
- $X^T = (X_1, X_2, \ldots, X_p)$. The linear regression model is

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

- Equivalently,

$$Y = \beta_0 + \sum_{j=1}^{p} X_j\beta_j + \varepsilon$$

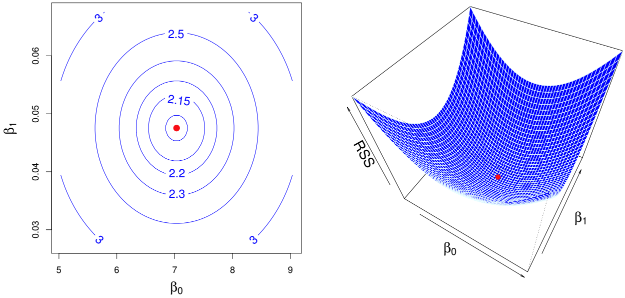
  where $\mathrm{E}\varepsilon = 0$
- Slightly different from previous courses, we adopt the view that X is random too.



Y
X_2
X_1

Solve via RSS

- Denote $\mathbf{X}$ the $n \times (p+1)$ matrix with each row an observation (with 1 at the first position.)
- $RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$
- Gradient equation: $-2\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) = 0$
- Hessian matrix: $2(\mathbf{X}^T\mathbf{X})$
- Assume that $\mathbf{X}$ has full column rank, then $\mathbf{X}^T\mathbf{X} > 0$ and the solution to the gradient equation is indeed the minimizer.
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$
- $\hat{\boldsymbol{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = \mathbf{H}\boldsymbol{y}$
- $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$ (assuming that σ^2 is the conditional variance of Y given X)


3
3
2.5
2.15
2.2
2.3
3
3
0.06
0.05
0.04
0.03
β_1
5
6
7
8
9
β_0
RSS
β_0
β_1

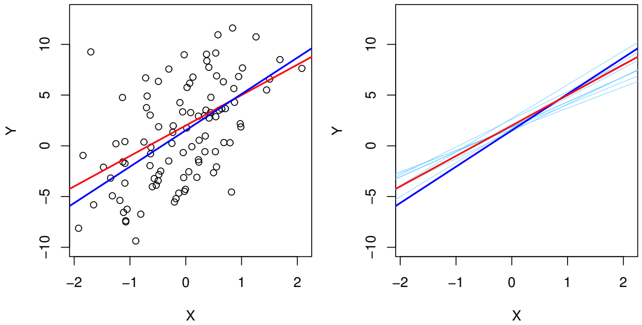
Unbiased estimates and MLE of σ^2.

- σ^2 is unbiasedly estimated by $\frac{1}{N-p-1} RSS(\hat{\beta})$
- When assuming that $Y|X$ is normal distributed, we have additional the MLE of σ^2 (homework).

Statistical Inference

Under the normal assumption:

- $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$
- $RSS/\sigma^2 \sim \chi^2_{N-p-1}$



Statistical Inference

Under the normal assumption:

- $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$
- $RSS/\sigma^2 \sim \chi^2_{N-p-1}$

Can be used to build hypothesis tests and confidence intervals.

- z score: $z_j = \dfrac{\hat{\beta}_j}{\hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}}$ which follows t_{N-p-1} under the null hypothesis that $\beta_j = 0$
- F statistic: $\dfrac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)}$ when two models (0 and 1) are nested, which follows $F_{(p_1-p_0),(N-p_1-1)}$ distribution under the null hypothesis that the smaller model (0) is correct (and the larger model is redundant.)

Gauss-Markov Theorem: OLS is BLUE

- Suppose we would like to estimate any linear combination of the parameter $\boldsymbol{a}^T\boldsymbol{\beta}$
- The least square estimate (based on the OLS $\hat{\boldsymbol{\beta}}$) is $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$
- It is clearly unbiased, and it is linear in $\boldsymbol{y}$
- If we have another linear estimator $\boldsymbol{c}^T\boldsymbol{y}$ that is unbiased for $\boldsymbol{a}^T\boldsymbol{\beta}$, then the Gauss-Markov Theorem states that

$$\mathrm{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) \leq \mathrm{Var}(\boldsymbol{c}^T\boldsymbol{y})$$

In other words, it is the best linear unbiased estimator.

MSE in Linear Regression

Imagine $\boldsymbol{a}$ is x_0, that is, $\boldsymbol{a}^T\boldsymbol{\beta}$ is the mean of a future observation (at x_0). Recall the Bias-Variance decomposition of the MSE of this estimation problem:

$$\mathrm{E}(x_0^T\hat{\boldsymbol{\beta}} - x_0^T\boldsymbol{\beta})^2 = \underbrace{\mathrm{E}(x_0^T\hat{\boldsymbol{\beta}} - E(x_0^T\hat{\boldsymbol{\beta}}))^2}_{\text{Variance of } x_0^T\hat{\boldsymbol{\beta}}} + \underbrace{(E(x_0^T\hat{\boldsymbol{\beta}}) - x_0^T\boldsymbol{\beta})^2}_{\text{Bias}^2}$$

The Gauss-Markov theorem implies that the least squares estimator has the smallest MSE among all linear estimators with no bias. However, there may well exist a biased estimator with **smaller mean squared error.** Such an estimator would trade a little bias for a larger reduction in variance.

In spirit, this is similar to the fit vs. complexity tradeoff we discussed in Lecture 1.

Bias-Variance Tradeoff and Model Complexity

Complex models have less bias; simpler models have less variance.

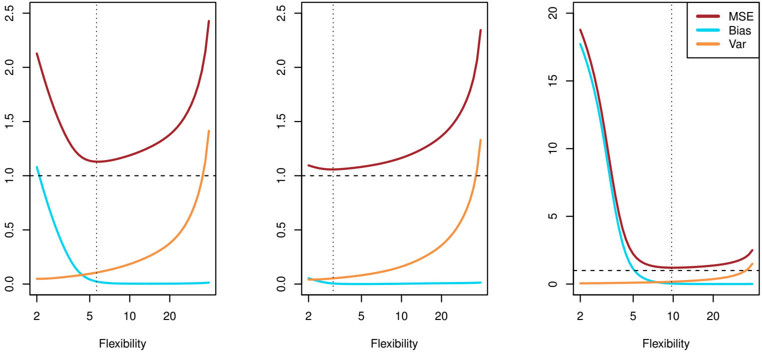
- Model too simple $\Rightarrow$ does NOT fit the data well. A biased solution.
- Model too complex $\Rightarrow$ small changes to the data changes predictor a lot. A high-variance solution.

Expected prediction error and MSE

- Often we talk about prediction errors instead of error of estimation of the mean.
- The Expected Prediction Error in predicting Y_0 at x_0 using $x_0^T\hat{\boldsymbol{\beta}}$ is

$$
\begin{aligned}
\mathrm{E}(Y_0 - x_0^T\hat{\boldsymbol{\beta}})^2 &= \mathrm{E}(Y_0 - \mathrm{E}(Y_0|X) + \mathrm{E}(Y_0|X) - x_0^T\hat{\boldsymbol{\beta}})^2 \\
&= \mathrm{E}(Y_0 - \mathrm{E}(Y_0|X))^2 + \mathrm{E}(\mathrm{E}(Y_0|X) - x_0^T\hat{\boldsymbol{\beta}})^2 \\
&= \mathrm{E}(Y_0 - x_0^T\boldsymbol{\beta})^2 + \mathrm{E}(x_0^T\boldsymbol{\beta} - x_0^T\hat{\boldsymbol{\beta}})^2 \\
&= \mathrm{E}(\varepsilon_0^2) + \mathrm{E}(x_0^T\boldsymbol{\beta} - x_0^T\hat{\boldsymbol{\beta}})^2 \\
&= \sigma^2 + \underbrace{\mathrm{E}(x_0^T\hat{\boldsymbol{\beta}} - E(x_0^T\hat{\boldsymbol{\beta}}))^2}_{\text{Variance of } x_0^T\hat{\boldsymbol{\beta}}} + \underbrace{(E(x_0^T\hat{\boldsymbol{\beta}}) - x_0^T\boldsymbol{\beta})^2}_{\text{Bias}^2}
\end{aligned}
$$

- Hence, the prediction error can be decomposed to the systematic noise (which we cannot control), the variance, and the bias.



0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Flexibility
0
5
10
15
20
MSE
Bias
Var

when $p > n$ or when $\mathbf{X}$ is not full rank

- In either case, $(\mathbf{X}^T\mathbf{X})$ is not invertible.
- Note $-2\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) = 0$ is

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\boldsymbol{y}$$

- The gradient equation may have infinitely many solutions.
- When $p > n$, the equation $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{y}$ has infinitely many solutions. That is, the residuals are zeros. So is the training error.
- In these case, the OLS estimate of $\boldsymbol{\beta}$ (or like), is not meaningful.

Considerations covered in standard linear regression course

- Qualitative Predictors
- Beyond the Additive Assumption: interaction term
- Non-linear Relationships
- Diagnostics (Non-linearity, correlation of error term, Non-constant variance, outliers, High-leverage points, Collinearity)

Lecture 2 R code

- Verify the OLS formula.
- Zero training error when $p > n$.

Summary

- Framework of Linear Regression
- Bias-Variance Decomposition (Tradeoff)
- Gauss-Markov theorem
- High-dimensional data