STAT 432: Basics of Statistical Learning

Linear Regression

Shiwei Lan, Ph.D. <shiwei@illinois.edu>

http://shiwei.stat.illinois.edu/stat432.html

February 6, 2019

University of Illinois at Urbana-Champaign

Outline

- Linear Models for Regression
- Bias-variance Trade-off
- Model Selection Criteria
- Model Selection Algorithm

Linear Models for Regression

Regression Models

- Observe a collection of i.i.d. training data

$$\mathcal{D}_n = \{x_i, y_i\}_{i=1}^n$$

where each x_i is a p dimensional vector (prediction variables, covariates, features, inputs), i.e.

$$x_i = (x_{i1}, \ldots, x_{ip})^\mathsf{T}$$

and $y_i \in \mathbb{R}$ is a continuous response (outcome, output).

- We want to estimate $f(X)$ using the training data to describe the relationship between X and Y.

Regression Models

- To clarify some other notations:
- $\mathbf{x}_j$ is an n dimensional vector of the jth feature, i.e.

$$\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^{\mathsf{T}}$$

- The design matrix $\mathbf{X}$ is $n \times p$ dimensional,

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$$

Loss and Risk functions

- To estimate $f(X)$, we need to define a criterion for a good estimator, $\widehat{f}(\cdot)$.
- We define a loss function L that measures the discrepancies between Y and $f(X)$. For regression, a commonly used loss function is the squared error loss:

$$L\big(Y, f(X)\big) = \big(Y - f(X)\big)^2.$$

- Risk is the expected loss over the entire population

$$\mathsf{R}(f) = \mathsf{E}\left[L\big(Y, f(X)\big)\right] = \mathsf{E}\left[\big(Y - f(X)\big)^2\right].$$

Minimizing the Empirical Risk

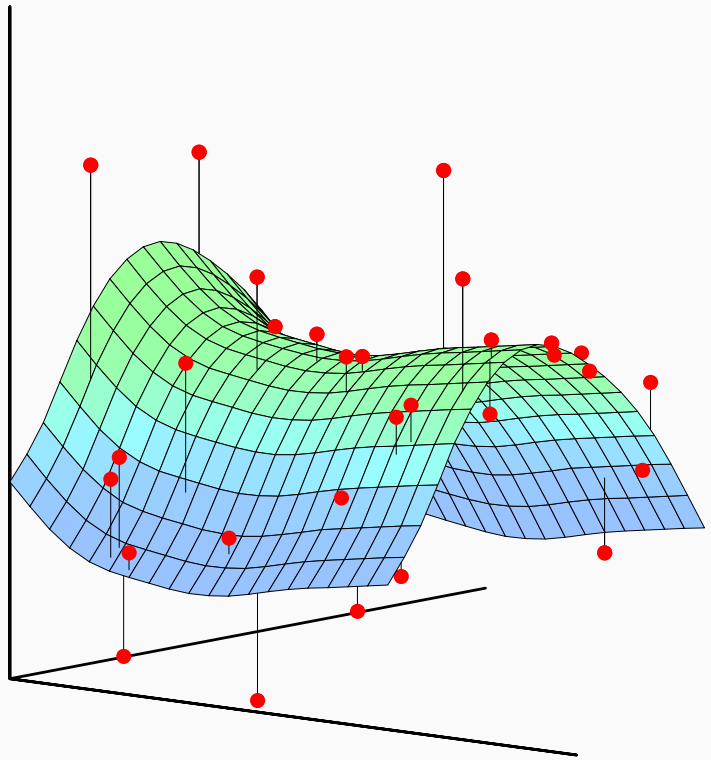
- In practice, we cannot directly calculate the risk, however, with the observed training data $\mathcal{D}_n$, we can calculate the empirical risk, which is simply replacing the expectation with the average over n training samples.

$$\mathsf{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2.$$

- We search for a function $\widehat{f}$ (in a certain space $\mathcal{F}$) to minimize the empirical risk on the training dataset

$$\begin{aligned} \widehat{f} &= \arg\min_{f \in \mathcal{F}} \mathsf{R}_n(f) \\ &= \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2. \end{aligned}$$

Minimizing the Empirical Risk



from ESL textbook

Linear Regression

- A linear regression model describes the dependence between X and Y by

$$\begin{aligned} Y &= X^{\mathsf{T}}\boldsymbol{\beta} + \epsilon \\ &= \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \end{aligned}$$

  where $\mathsf{E}(\epsilon) = 0$, $\mathsf{Var}(\epsilon) = \sigma^2$ and $\epsilon \perp X$.

- Given the training data $\mathcal{D}_n$, we express the regression model in the matrix form

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{e}_{n\times 1}$$

  where $\mathbf{X}_{n\times p}$ is called the design matrix with each row representing one subject.

- Intercept can be included by setting the first column of $\mathbf{X}$ to be 1.

Linear Regression

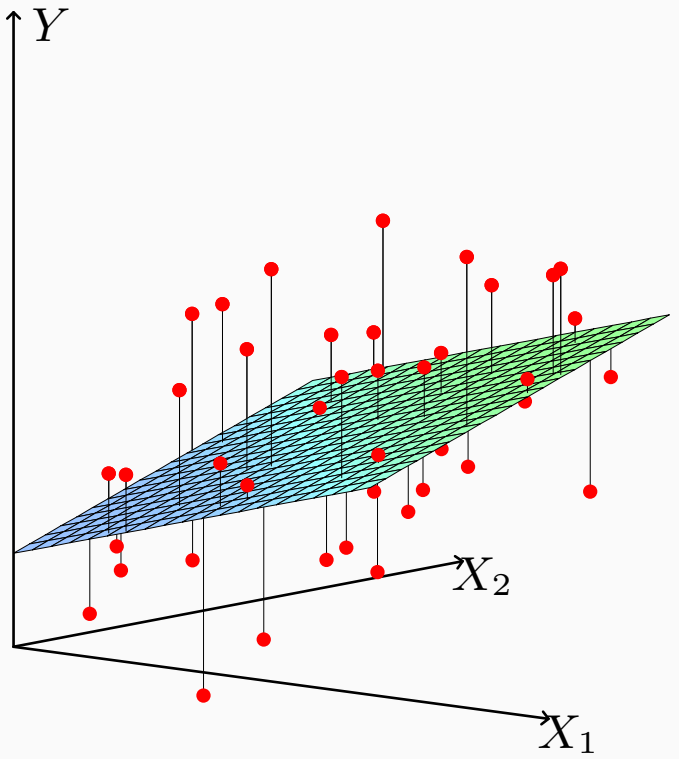
- Now, estimating f comes down to estimating $\boldsymbol{\beta}$.
- Based on our previous definition of the empirical risk, we solve for $\boldsymbol{\beta}$ that minimizes the residual sum of squares (RSS)

$$\begin{aligned}\text{RSS} &= \sum_{i=1}^{n} \Big(y_i - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p \Big)^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ &= \big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\big)^{\mathsf{T}} \big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\big)\end{aligned}$$

- The ordinary least squares estimator (OLS) is

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\big)^{\mathsf{T}} \big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\big)$$

Linear Regression



from ESL textbook

Estimating β

- To estimate β, we set the derivative equal to 0

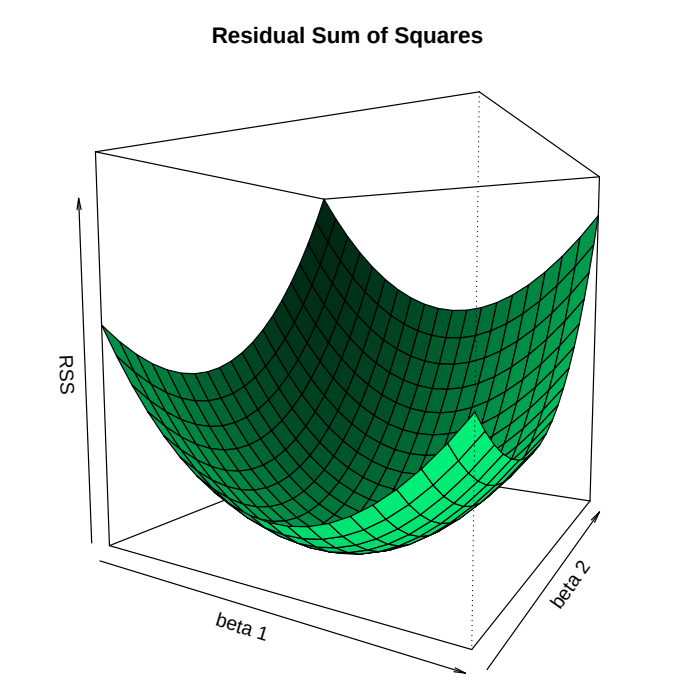
$$\frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$
$$\implies \quad \mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}$$

  which is commonly known as the normal equation.
- $\mathbf{X}$ full rank $\iff \mathbf{X}^\mathsf{T}\mathbf{X}$ invertible
- We then have, if $\mathbf{X}^\mathsf{T}\mathbf{X}$ is invertible,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

A Convex Problem

- There are many different ways to view a linear regression.
- One way is to view it as a convex optimization problem, which helps understand Lasso and Ridge.
- When $\mathbf{X}^\mathsf{T}\mathbf{X}$ is invertible, the RSS is a strictly convex function of $\boldsymbol{\beta}$



Hat Matrix

- The fitted values (i.e., prediction at the n observed data points) are

$$\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \doteq \mathbf{H}_{n\times n}\mathbf{y}$$

- The "hat matrix"

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

  is a project matrix that projects onto the column space of $\mathbf{X}$.
  - symmetric: $\mathbf{H}^\mathsf{T} = \mathbf{H}$
  - idempotent: $\mathbf{HH} = \mathbf{H}$

Residuals

- The residual $\mathbf{r}$ is defined as

$$\begin{aligned}\widehat{\mathbf{e}} = \mathbf{r}_{n\times 1} &= \mathbf{y} - \widehat{\mathbf{y}} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

- $\mathbf{r}$ can be used to estimate the error variance

$$\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 = \frac{\text{RSS}}{n-p}$$

Vector Space Interpretation

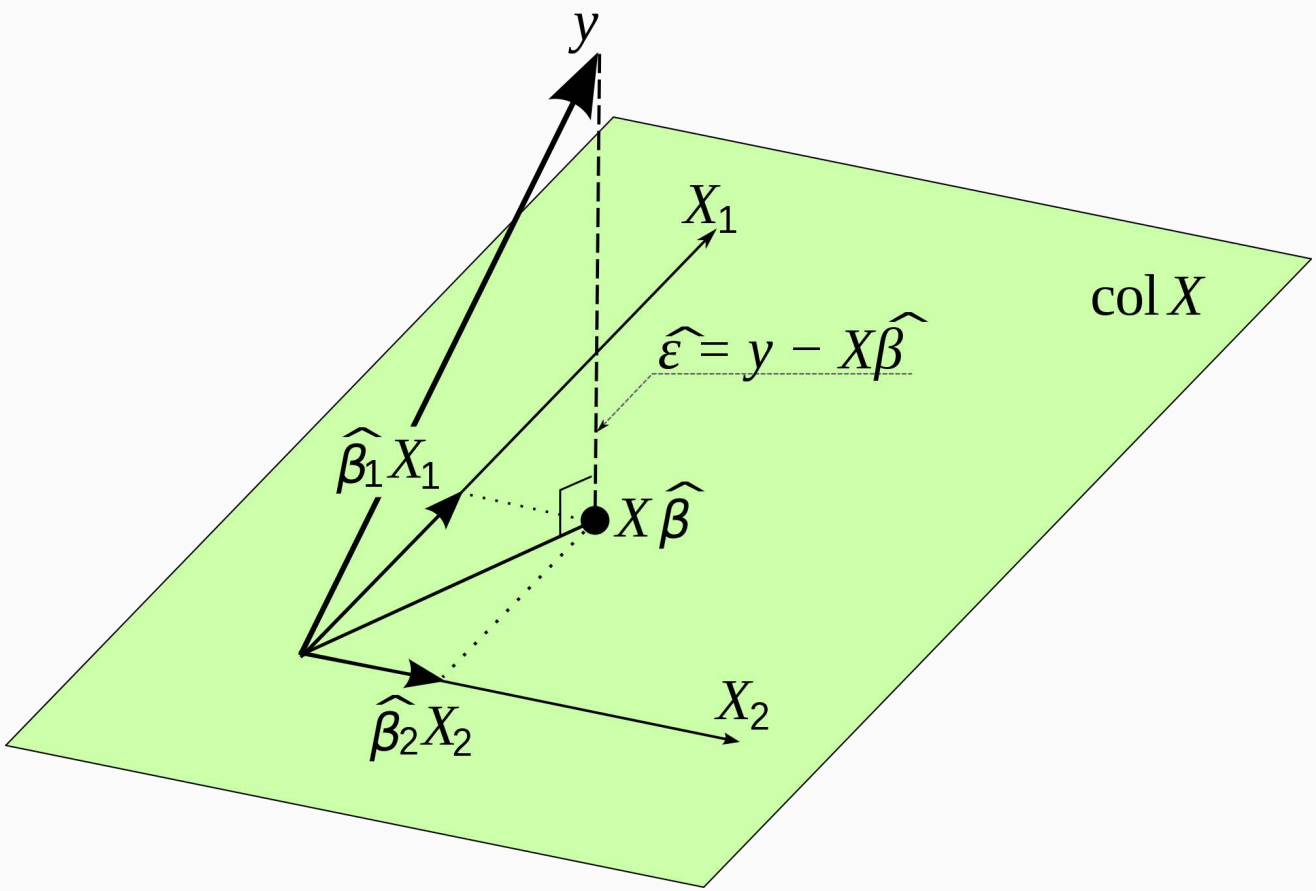


Figure from Wikipedia

Vector Space Interpretation

- The essence of LS is to decompose the data vector $\mathbf{y}$ into two orthogonal vectors

$$\begin{aligned}\mathbf{y} &= \mathbf{H}\mathbf{y} + (\mathbf{I} - \mathbf{H})\mathbf{y} \\ &= \widehat{\mathbf{y}} + \mathbf{r}\end{aligned}$$

- Note that since $\mathbf{H}$ is a projection matrix, $\mathbf{r}$ is orthogonal to each column of $\mathbf{X}$, i.e.,

$$\mathbf{X}^{\mathsf{T}}\mathbf{r} = \mathbf{0}_{p\times 1}.$$

Properties of $\widehat{\beta}$

- If the samples are indeed generated from a linear model

$$Y = X^{\mathsf{T}}\boldsymbol{\beta} + \epsilon,$$

  where the errors ϵ_i are i.i.d., independent of X, with $\mathsf{E}(\epsilon_i) = 0$ and $\mathsf{Var}(\epsilon_i) = \sigma^2$.
- Then $\widehat{\boldsymbol{\beta}}$ is unbiased: $\mathsf{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
- Variance-covariance

$$\begin{aligned}
\mathsf{Var}(\widehat{\boldsymbol{\beta}}) &= \mathsf{Var}\big((\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}\big) \\
&= \mathsf{Var}\big((\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{X}\beta + \boldsymbol{\epsilon})\big) \\
&= \mathsf{Var}\big((\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{\epsilon})\big) \\
&= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{I}\sigma^2 \\
&= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\sigma^2
\end{aligned}$$

Properties of $\widehat{\beta}$

- By the Gauss-Markov Theorem, $\widehat{\beta}$ is the best linear unbiased estimator (BLUE)
- If the errors are generated from a Gaussian distribution, then $\widehat{\beta}$ is also the minimum variance unbiased estimator (MVUE)
- However, based on our understanding of the bias-variance trade-off, we could sacrifice the unbiasedness to trade for a large reduction in variance. Then the overall prediction error may perform better.

Bias-Variance Trade-Off in Linear Regression

Dealing with large p

- In many applications nowadays, we have many explanatory variables, i.e., p is large or even $p \gg n$.
  - There are more than 20,000 human protein-coding genes
  - About 10 million single nucleotide polymorphisms (SNPs)
  - Number of subjects, n, is usually in hundreds or thousands
- In some applications, the key question is to identify a subset of X variables that are most relevant to Y
- Let's examine the training and testing errors from a linear model

Training and Testing Data

- Training data $\mathcal{D}_n = \{x_i, y_i\}_{i=1}^n$
- Suppose $\{x_i, y_i^*\}_{i=1}^n$ is an independent (imaginary) testing dataset collected at the same location x_i's (aka, in-sample prediction
- Assume that the data are generated from

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$
$$\mathbf{y}^* = \boldsymbol{\mu} + \mathbf{e}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}^*$$

  where both $\mathbf{y}$ and $\mathbf{y}^*$ are $n \times 1$ response vectors, $\mathbf{e}$ and $\mathbf{e}^*$ are i.i.d. error terms with mean 0 and variance σ^2.
- The true model is indeed linear!
- Goal: What is the best model that predicts $\mathbf{y}^*$?

Testing Error

$$
\begin{aligned}
\mathsf{E}[\text{Test Err}] &= \mathsf{E}\|\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 \\
&= \mathsf{E}\|(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}) + (\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\widehat{\boldsymbol{\beta}})\|^2 \\
&= \mathsf{E}\|\mathbf{e}^*\|^2 + \mathsf{E}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 \\
&= n\sigma^2 + \mathsf{E}\big[\mathsf{Trace}\big((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big)\big] \\
&= n\sigma^2 + \mathsf{Trace}\big(\mathbf{X}^\mathsf{T}\mathbf{X}\mathsf{Cov}(\widehat{\boldsymbol{\beta}})\big) \\
&= n\sigma^2 + p\sigma^2
\end{aligned}
$$

- We used the properties:
  - $\mathsf{Trace}(ABC) = \mathsf{Trace}(CBA)$
  - $\mathsf{E}(\mathsf{Trace}(A)) = \mathsf{Trace}(\mathsf{E}(A))$

Training Error

$$
\begin{aligned}
\mathsf{E}[\text{Train Err}] &= \mathsf{E}\|\mathbf{y} - \widehat{\mathbf{y}}\|^2 = \mathsf{E}\|(\mathbf{I} - \mathbf{H})\mathbf{y}\|^2 \\
&= \mathsf{E}\|(\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{e})\|^2 \\
&= \mathsf{E}\|(\mathbf{I} - \mathbf{H})\mathbf{e}\|^2 \\
&= \mathsf{E}\left[\mathsf{Trace}\left(\mathbf{e}^\mathsf{T}(\mathbf{I} - \mathbf{H})^\mathsf{T}(\mathbf{I} - \mathbf{H})\mathbf{e}\right)\right] \\
&= \mathsf{Trace}\left((\mathbf{I} - \mathbf{H})^\mathsf{T}(\mathbf{I} - \mathbf{H})\mathsf{Cov}(\mathbf{e})\right) \\
&= (n - p)\sigma^2
\end{aligned}
$$

- We used the property:
  - $\mathbf{HX} = \mathbf{X}$

Training vs. Testing error

- Summary:
  - testing error: $n\sigma^2 + p\sigma^2$
  - training error: $(n-p)\sigma^2$
- The expected testing error increase with p and the expected training error decreases with p.
- When p gets large, this is a big trouble. Consider the case $p = n$, this is equivalent to 1NN.
- Can we just select a few number of variables to reduce p?
- What could be the consequences?

Variable Selection

- Variable/model selection may improve
  - Prediction accuracy
  - Interpretability
- However, this may also increase bias (we did not discuss them in the previous derivation) because we are taking the risk of removing some important variables.
- Overall, this is a difficult task.
  - No natural ordering of importance for the variables
  - The role of a variable needs be measured conditioning on others, high correlation causes trouble
  - It is essential to check all possible combinations, however, this may be computationally expansive

Model Selection Criteria

Motivation

- If we compare the two errors:
  - testing error: $n\sigma^2 + p\sigma^2$
  - training error: $(n-p)\sigma^2$

  we have:

$$\text{testing error} = \text{training error} + 2p\sigma^2$$

- Training error (RSS) is always computable, and we can estimate σ^2 using $\widehat{\sigma}^2$.
- Hence, how about searching for a model that minimizes

$$\text{RSS} + 2\widehat{\sigma}^2_{\text{full}} \cdot p$$

- $\widehat{\sigma}^2_{\text{full}}$ can be estimated using the full model, with all variables.
- The method is called Mallows' C_p (Mallows 1973)

Model Selection Criteria

- Model selection is usually done in the following way
  1. Give each fitted model a score (goodness-of-fit)
  2. Design an algorithm to find the model with the best score
- The score of a fitted model usually takes the the form

$$\text{goodness-of-fit} + \text{model-complexity}$$

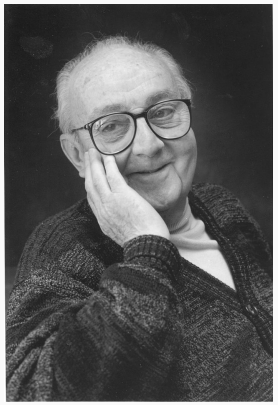
- The first term will decrease as the model gets more complicated (recall 1NN, or linear model with $p = n$)
- The second term increases with the number of predictors used, which prefers “smaller” models

Model Selection Criteria

- Popular choices of scores:
  - Mallows' C_p (Mallows 1973): RSS + $2\widehat{\sigma}^2_{\text{full}} \cdot p$
  - AIC (Akaike 1970): -2 Log-likelihood $+ 2 \cdot p$
  - BIC (Schwarz, 1978): -2 Log-likelihood $+ \log n \cdot p$
- AIC is motivated from the Kullback–Leibler divergence; BIC is motivated from Bayesian posterior.
- C_p performs similarly to AIC.
- When n is large, adding one predictor costs a lot more in BIC than AIC (or C_p). So AIC tends to pick a larger model than BIC.

Bias-Variance Trade-Off

- Recall our previous analysis of the training and testing errors with $\mathbf{y}$ and $\mathbf{y}^*$, no bias term was involved.
- This is because we assume that the true model is linear, and we always include all the necessary variables.
- What will happen if linear model is wrong? or we eliminated some true variables?
- "All models are wrong, but some are useful."



George E. P. Box, (1919 - 2013)

Bias-Variance Trade-Off

- Now, lets assume that the model is not necessarily a linear model, i.e.,

$$\mathbf{y} = f(\mathbf{X}) + \mathbf{e} = \boldsymbol{\mu} + \mathbf{e}$$
$$\mathbf{y}^* = f(\mathbf{X}) + \mathbf{e} = \boldsymbol{\mu} + \mathbf{e}^*$$

- But we don't have $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. However, we still perform a linear regression.
- Note that $\boldsymbol{\mu}$ is a vector of n elements, the best linear model is essentially projecting this mean vector onto the column space defined by $\mathbf{X}$. Hence, the best linear model to describe this $\mathbf{H}\boldsymbol{\mu}$ — projecting the mean vector onto the column space of $\mathbf{X}$.
- This will introduce bias as long as $\mathbf{H}\boldsymbol{\mu} \neq \boldsymbol{\mu}$.

Justification of Mallows' C_p

$$
\begin{aligned}
\text{E}[\text{Test Err}] &= \text{E}\|\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\mathbf{y}^* - \mathbf{H}\mathbf{y}\|^2 \\
&= \text{E}\|(\mathbf{y}^* - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\mu}) + (\mathbf{H}\boldsymbol{\mu} - \mathbf{H}\mathbf{y})\|^2 \\
&= \text{E}\|\mathbf{y}^* - \boldsymbol{\mu}\|^2 + \text{E}\|\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\mu}\|^2 + \text{E}\|\mathbf{H}\boldsymbol{\mu} - \mathbf{H}\mathbf{y}\|^2 \\
&= \text{E}\|\mathbf{e}^*\|^2 + \text{E}\|\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\mu}\|^2 + \text{E}\|\mathbf{H}\mathbf{e}\|^2 \\
&= n\sigma^2 + \text{Bias}^2 + p\sigma^2
\end{aligned}
$$

$$
\begin{aligned}
\text{E}[\text{Train Err}] &= \text{E}\|\mathbf{y} - \widehat{\mathbf{y}}\|^2 = \text{E}\|(\mathbf{I} - \mathbf{H})\boldsymbol{\mu} + (\mathbf{I} - \mathbf{H})\mathbf{e}\|^2 \\
&= \text{E}\|(\mathbf{I} - \mathbf{H})\boldsymbol{\mu}\|^2 + \text{E}\|(\mathbf{I} - \mathbf{H})\mathbf{e}\|^2 \\
&= \text{Bias}^2 + (n - p)\sigma^2
\end{aligned}
$$

Hence, we still have Test Err $=$ Train Err $+ 2\sigma^2 p$.

Model Selection Algorithm

Basic Idea

- Basic idea:
  - Pick a penalty for model complexity (Mallows' C_p, AIC or BIC)
  - Try models with different variables
  - For each model, calculate the sum of goodness-of-fit and the penalty for model complexity
  - Compare all candidates, and pick the best one
- Note: When comparing two models with the same number of variables, only the goodness-of-fit measure matters.
- Commonly used algorithms: best subset selection; backward/forward selection.

Best Subset Selection

- Best subset selection is a level-wise search algorithm, which returns the global optimal solution for a given model size.
- Only feasible for p not very large (< 50)
- Algorithm:
  1). For each $k = 1, \ldots, p$, check 2^k possible combinations, and find the model with smallest RSS
    - The penalty term is the same for models with the same size
  2). To choose the best k, use model selection criteria

Best Subset Selection

- Note: if RSS$(X_1, X_2) <$ RSS(X_3, X_4, X_5, X_6) then we do not need to visit any size 2 or 3 sub-models of (X_3, X_4, X_5, X_6), which can be leaped over.
- Implemented in R contributed package leaps, using the leaps and bounds algorithm (Furnival and Wilson, 1974)

Stepwise Regression

- Greedy algorithms: fast, but only return a local optimal solution (which might be good enough in practice).
  - Backward: start with the full model and sequentially delete predictors until the score does not improve.
  - Forward: start with the null model and sequentially add predictors until the score does not improve.
  - Stepwise: consider both deleting and adding one predictor at each stage.

Applications

- See the supplementary R file
- Example 1: the diabetes data analysis
- Read more in ISL Ch 6.1. Check ESL Video.