Linear Regression

Yuan Yao

Department of Mathematics
Hong Kong University of Science and Technology

Chapter 3

Spring, 2018

About the Course

- Course web: https://yuany-pku.github.io/2018_math4432/
- Related courses:
  - Statistical Learning: Math 5470 by Bingyi JING
  - Unsupervised learning: CSIC5011 2017, Topological and Geometric Data Reduction (http://math.stanford.edu/~yuany/course/2017.fall/)
  - Deep learning: Math 6380o (https://deeplearning-math.github.io/)

Summary of The Bias-Variance Trade-Off

Consider $y_i = f(x_i) + \epsilon_i$, $i = 1, ..., n$ where ϵ_i are mean 0 with variance σ^2. An estimate function from data $\mathcal{D}$ is denoted as $\hat{f}(x; \mathcal{D})$ or simply $\hat{f}(x)$, then the expected predicted error is

$$\begin{aligned}
\text{EPE}(x) &= E((Y - \hat{f}(x))^2 | X = x) \\
&= E((\epsilon + f(x) - \hat{f}(x))^2 | X = x) \\
&= E\{((\epsilon + f(x) - E(\hat{f}(x)) + E(\hat{f}(x)) - \hat{f}(x))^2 | X = x\} \\
&= \sigma^2 + (f(x) - E(\hat{f}_k(x)))^2 + \text{var}(\hat{f}(x)) \\
&= \underbrace{\sigma^2}_{\textbf{Irreducible}} + \underbrace{\text{Bias}^2 + \text{Variance}}_{\textbf{Reducible}}
\end{aligned}$$

- **Bias** refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
- **Variance** refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different $\hat{f}$. But ideally the estimate for f should not vary too much between training sets.
- **Bias and variance trade-off**: The optimal predictive ability is the one that leads to balance between bias and variance.

Bias-variance tradeoff

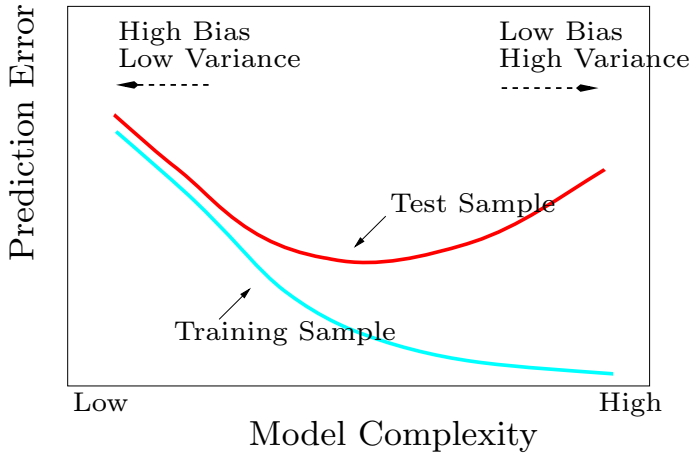


FIGURE 2.11. *Test and training error as a function of model complexity.*

General models/methods

- linear models, generalized linear models.
- K-Nearest-Neighbor (KNN)
- Kernel methods
- local polynomial regression
- regression and smoothing splines
- Tree based methods and boosting
- SVM
- projection pursuit and neural networks.

Prediction vs. Interpretability

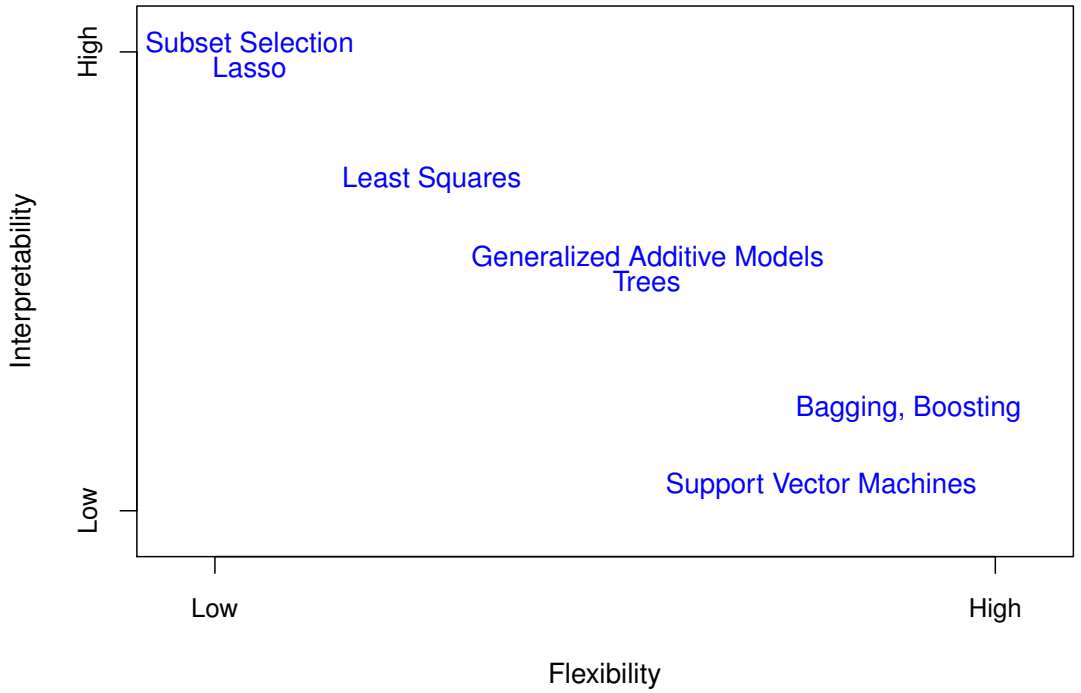


Figure: "Occam's Razor" or "Simplicity is beauty" or "Everything else the same, bring the simplest model". Vote for parsimonious models or bet on sparsity

1. 3.1. Simple linear regression
2. 3.2 Multiple linear regression
3. 3.3. The least squares estimation
4. 3.4. The statistical properties of the least squares estimates.
5. 3.5 The variance decomposition and analysis of variance (ANOVA).
6. 3.6. More about prediction
7. 3.7. The optimality of the least squares estimation.
8. 3.8. Assessing the regression model.
9. 3.9. Variable selection.
10. 3.10. A comparison with KNN through a simulated example

Outline

1. 3.1. Simple linear regression
2. 3.2 Multiple linear regression
3. 3.3. The least squares estimation
4. 3.4. The statistical properties of the least squares estimates.
5. 3.5 The variance decomposition and analysis of variance (ANOVA).
6. 3.6. More about prediction
7. 3.7. The optimality of the least squares estimation.
8. 3.8. Assessing the regression model.
9. 3.9. Variable selection.
10. 3.10. A comparison with KNN through a simulated example

About linear regression model

- Fundamental statistical models. (supervised learning)
- Covering one-sample, two-sample, multiple sample problems.
- Most illustrative on various important issues: fitting, prediciton, model checking, ...
- In-depth understanding of linear model helps learning further topics.
- This chapter is slightly more advanced than Chapter 3 of ISLR

The formulation

- Consider linear regression model:

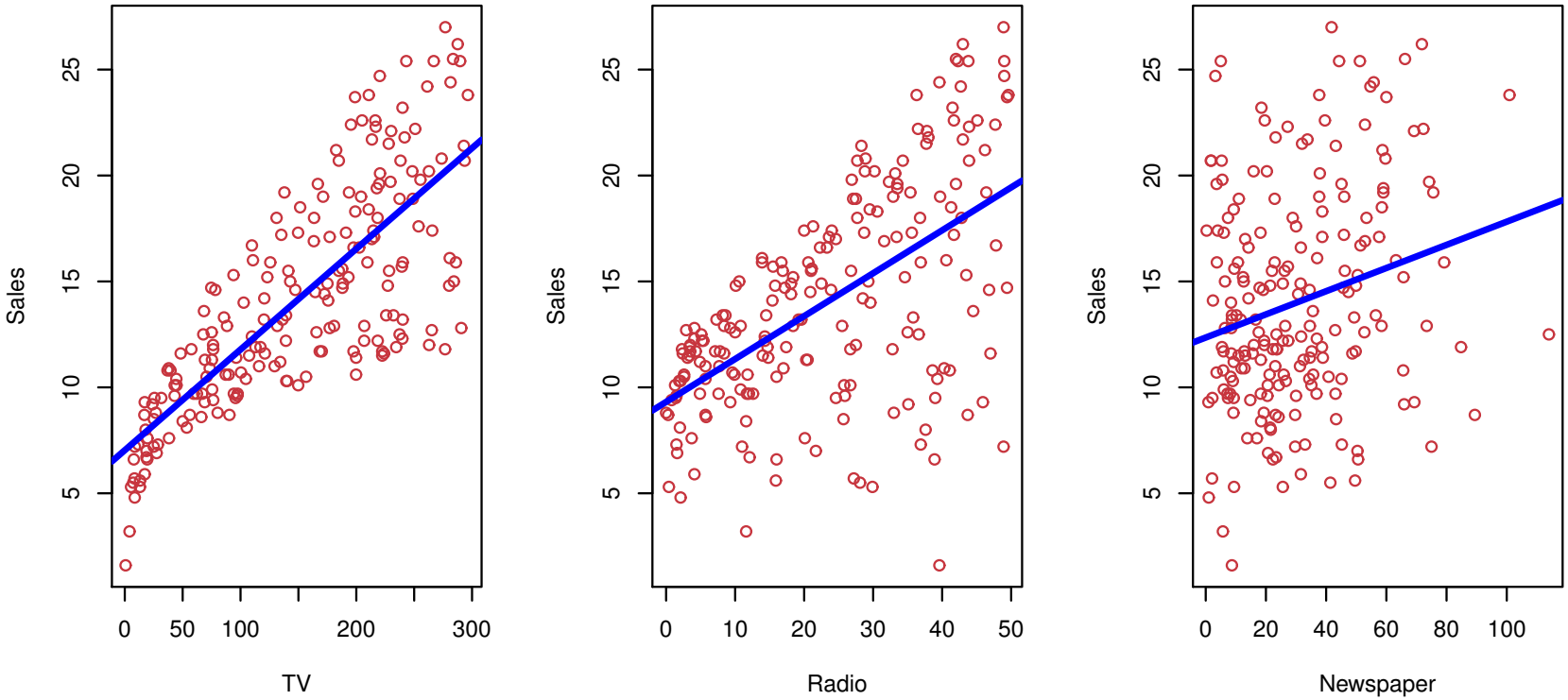
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, ..., n$$

  where y_i, x_i are the i-th observation of the response and covariates. Here x_i is of 1-dimension.
- Responses are sometimes called dependent variables or outputs;
- covariates called independent variables or inputs or features or predictors or regressors.
- obtain the parameter estimation and making prediction of any responses on given covariates.

Example: Advertising data

The data contains 200 observations.
Sample size: $n = 200$.
Sales: y_i, $i = 1..., n$.
TV (bugdets): $x_{i1}, i = 1, ..., n$.
Radio (budgets): $x_{i2}, i = 1, ..., n$.
Newspaper (budgets): $x_{i3}, i = 1, ..., n$.

Example: Advertising data



Simple linear regression

For the time being, we only consider one covariate: TV.
The linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i, \qquad i = 1, ..., n$$

Estimating the coefficient by the least squares

Minimizing the sum of squares of error:

$$\sum_{i=1}^{n}(y_i = \beta_0 + \beta_1 x_{i1})^2.$$

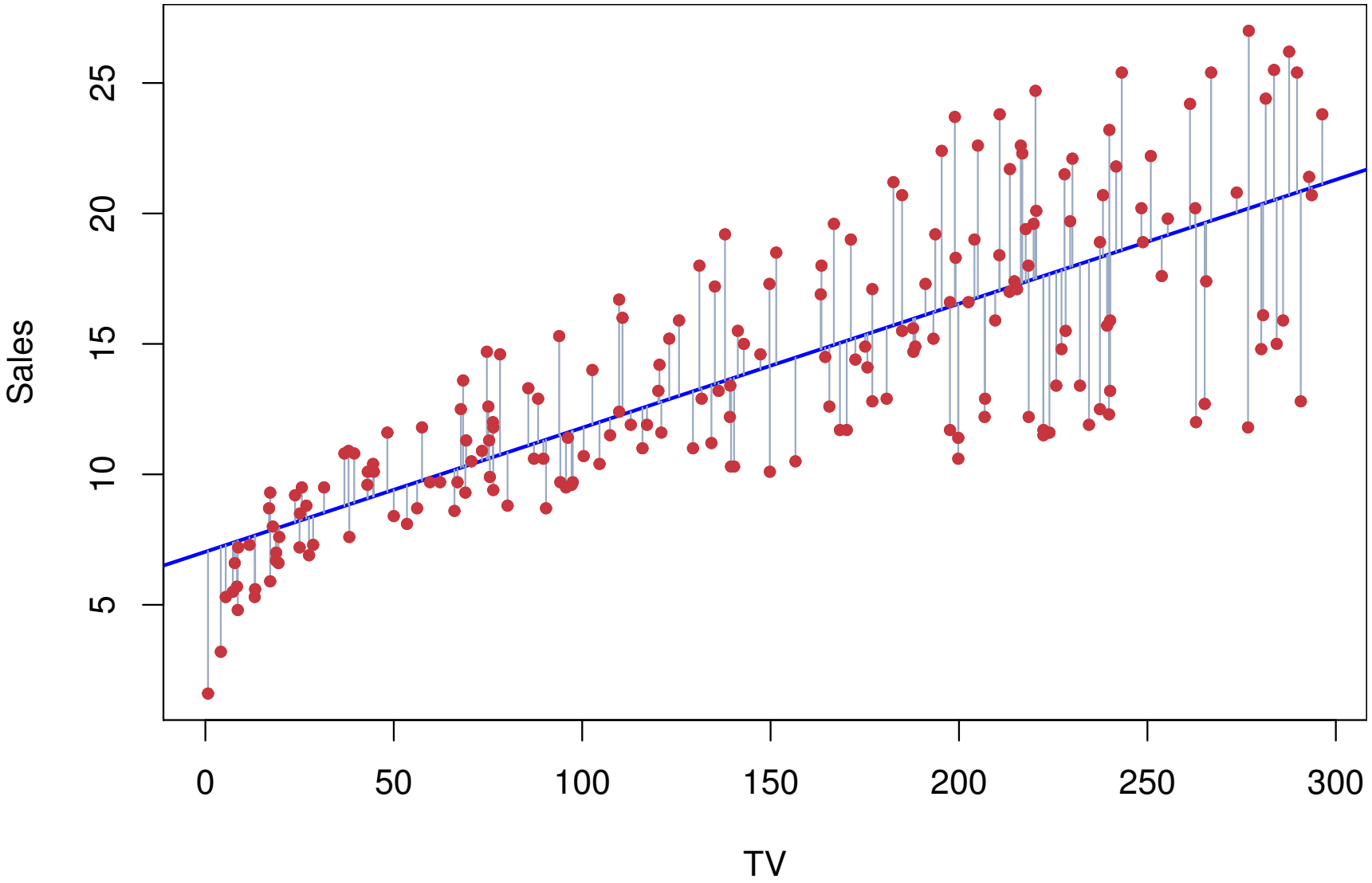
The estimator is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Here $x_i = x_{i1}$.

Illustrating least squares



Inference

$$\frac{\hat{\beta}_1 - \beta_1}{s\sqrt{1/\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_0 - \beta_0}{s\sqrt{1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2}$$

where

$$s^2 = \mathrm{RSS}/(n-2)$$

is an unbiased estimator of the variance of the error, and, setting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ as the so-called fitted value,

$$\mathrm{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

are so-called residual sum of squares.

Details would be provided in multiple linear regression.

Result of the estimation

TABLE 3.1. (from ISLR) The advertising data: coefficients of the LSE for the regression on number of units sold on TV advertising budget. An increase of $1000 in the TV advertising budget would cause an increase of sales of about 50 units.

	Coefficient	Std.error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Outline

1. 3.1. Simple linear regression
2. 3.2 Multiple linear regression
3. 3.3. The least squares estimation
4. 3.4. The statistical properties of the least squares estimates.
5. 3.5 The variance decomposition and analysis of variance (ANOVA).
6. 3.6. More about prediction
7. 3.7. The optimality of the least squares estimation.
8. 3.8. Assessing the regression model.
9. 3.9. Variable selection.
10. 3.10. A comparison with KNN through a simulated example

Linear models formulation

- Consider linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, ..., n \tag{2.1}$$

  where $y_i, x_i = (x_{i1}, ..., x_{ip})$ are the i-th observation of the response and covariates.
- Responses are sometimes called dependent variables or outputs;
- covariates called independent variables or inputs or regressors.
- obtain the parameter estimation and making prediction of any responses on given covariates.

Example: Advertising data

Now, we consider three covariates: TV, radio and newspapers.
The number of covarites $p = 3$.
The linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \qquad i = 1, ..., n$$

Estimating the coefficient by the least squares

Minimizing the sum of squares of error:

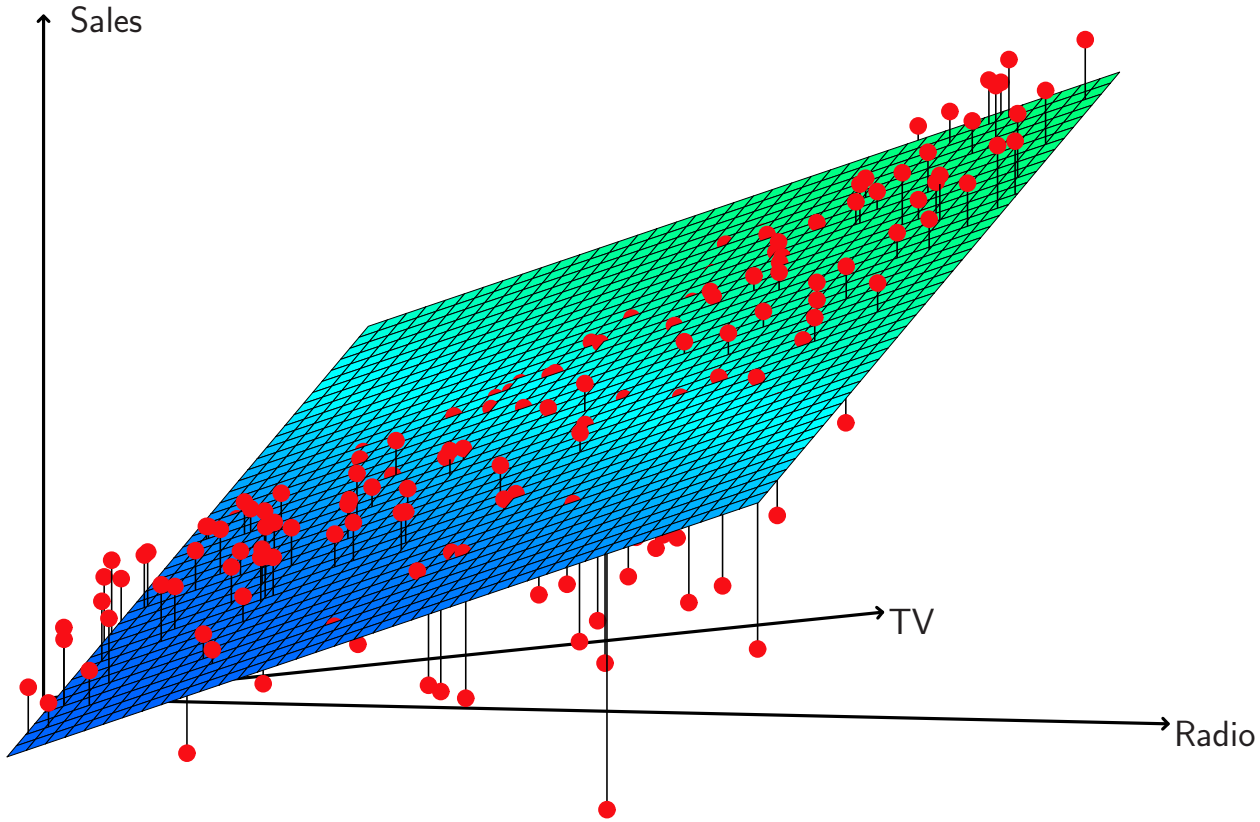
$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3})^2.$$

which is

$$\sum_{i=1}^{n}(y_i - \beta^T x_i)^2$$

The expression of the LSE of β, as a vector, has a simple matrix expression, even though the estimator of the individual $\hat{\beta}_i$ is not eqaully simple.

Illustrating the least squares



Result of the estimation

TABLE 3.9. (from ISLR) The advertising data: coefficients of the LSE for the regression on number of units sold on TV, radio and newspaper advertising budgets.

	Coefficient	Std.error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Outline

1. 3.1. Simple linear regression
2. 3.2 Multiple linear regression
3. 3.3. The least squares estimation
4. 3.4. The statistical properties of the least squares estimates.
5. 3.5 The variance decomposition and analysis of variance (ANOVA).
6. 3.6. More about prediction
7. 3.7. The optimality of the least squares estimation.
8. 3.8. Assessing the regression model.
9. 3.9. Variable selection.
10. 3.10. A comparison with KNN through a simulated example

Notations

With slight abuse of notation, in this chapter, we use

$$\begin{aligned}\mathbf{X} &= \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{1} \vdots \mathbf{x}_1 \vdots \mathbf{x}_2 \vdots \cdots \vdots \mathbf{x}_p \end{pmatrix}.\end{aligned}$$

Here a column of ones, $\mathbf{1}$, is added, which corresponds to the intercept β_0. Then $\mathbf{X}$ is a n by $p+1$ matrix.

Recall that

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \mathbf{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

The least squares criterion

The least squares criterion is try to minimize the sum of squares:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

Using matrix algebra, the above sum of squares is

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta).$$

The LSE, fitted values and residuals

By some linear algebra calcuation, the least squares estimator of β is then

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

Then

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

is called the fitted values; viewed as the predicted values of the reponses based on the linear model.

Terminology and notation

$$\mathbf{y} - \hat{\mathbf{y}}$$

are called residuals, which is denoted as $\hat{\epsilon}$. The sum of squares of these residuals

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2.$$

- The zero-correlation of two variables from multivariate normal random variable implies their independence.
- Suppose $\mathbf{z} = (z_1, ..., z_n)^T$, and z_i are iid standard normal random variables.
- Let $\mathbf{z}_1 = \mathbf{A}\mathbf{z}$ and $\mathbf{z}_2 = \mathbf{B}\mathbf{z}$ with $\mathbf{A}$ and $\mathbf{B}$ are two nonrandom matrices.
- Then

$$\text{cov}(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{A}\mathbf{B}^T = 0$$

  implies the independence between $\mathbf{z}_1$ and $\mathbf{z}_2$.
- We also call $\mathbf{z}_1$ and $\mathbf{z}_2$ orthogonal.

Orthogonality

- The residual $\hat{\epsilon}$ is orthogonal to all columns of $\mathbf{X}$, i.e, all $\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p$. This can be seen by

$$\begin{aligned} & \mathbf{X}^T \hat{\epsilon} = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} \hat{\beta} \\ = \quad & \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = 0. \end{aligned}$$

- The residual vector $\hat{\epsilon}$ is orthogonal to the hyperplane formed by vectors $\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p$ in n dimensional real space.

A proof of the LSE

$$
\begin{aligned}
& \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 \\
= \; & \|\mathbf{y} - \mathbf{X}\hat{\beta} - \mathbf{X}(\mathbf{b} - \hat{\beta})\|^2 \\
= \; & \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 + \|\mathbf{X}(\mathbf{b} - \hat{\beta})\|^2 \qquad \text{by orthogonality} \\
\geq \; & \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2
\end{aligned}
$$

- The fitted value $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$, which, also as a vector in n dimensional real space, is a linear combination of the vectors $\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p$, with the $p+1$ linear combination coefficients being the components of $\hat{\beta}$.
- The fitted values are orthogonal to the residuals, i.e., $\hat{\mathbf{y}}$ is orthogonal to $\mathbf{y} - \hat{\mathbf{y}}$ or

$$\hat{\mathbf{y}}^T(\mathbf{y} - \hat{\mathbf{y}}) = 0.$$

This implies

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2.$$

The projection matrix

- Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.
- This n by n matrix is called projection matrix or hat matrix.
- It has the property that, for any vector, $\mathbf{b}$ in n dimensional real space $\mathbf{Hb}$ projects $\mathbf{b}$ onto the linear space formed by the columns of $\mathbf{X}$.
- $\mathbf{Hb}$ is in this linear space formed by the columns of $\mathbf{X}$.
- And $\mathbf{b} - \mathbf{Hb}$ is orthogonal to this space.

The least squares projection

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 3

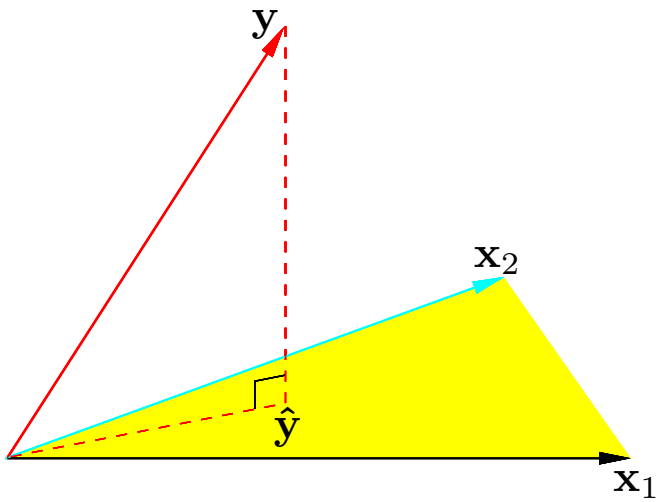


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

symmetric and idempotent

- projection matrix $\mathbf{H}$ is symmetric and idempotent; i.e., $\mathbf{H}^2 = \mathbf{H}$.
- eigenvalues are either 1 or 0.
- All eigenvectors associated with eigenvalue 1 form a space, say $\mathcal{L}_1$;
- Those with eigenvalue 0 form the orthogonal space, $\mathcal{L}_0$, of $\mathcal{L}_1$.
- Then $\mathbf{H}$ is the projection onto space $\mathcal{L}_1$ and $\mathbf{I} - \mathbf{H}$ is the projection onto $\mathcal{L}_0$, where $\mathbf{I}$ is the n by n identity matrix.

Matrix decomposition

- Suppose, for convenience $n \geq p$, any matrix n by p matrix A can always be decomposed into

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

where $\mathbf{U}$ is $n \times p$ orthogonal matrix, $\mathbf{D}$ is a $p \times p$ diagonal matrix and $\mathbf{V}$ is $p \times p$ orthogonal matrix. In particular

$$\mathbf{X} = \mathbf{U}\mathbf{R},$$

where $\mathbf{R} = \mathbf{D}\mathbf{V}$.

- If $\mathbf{A}$ and $\mathbf{B}^T$ are two matrices of same dimension, then

$$trace(\mathbf{A}\mathbf{B}) = trace(\mathbf{B}\mathbf{A}).$$

Outline

1. 3.1. Simple linear regression
2. 3.2 Multiple linear regression
3. 3.3. The least squares estimation
4. 3.4. The statistical properties of the least squares estimates.
5. 3.5 The variance decomposition and analysis of variance (ANOVA).
6. 3.6. More about prediction
7. 3.7. The optimality of the least squares estimation.
8. 3.8. Assessing the regression model.
9. 3.9. Variable selection.
10. 3.10. A comparison with KNN through a simulated example

Model assumptions

- The linear regression model general assumes the error ϵ_i has zero conditional mean and constant conditional variance σ^2, and the covariates x_i are non-random;
- Independence across the observations
- A more restrictive (but common) assumption: the errors follow normal distribution, i.e, $N(0, \sigma^2)$.

Statistical properties of LSE

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1});$$

$$\text{RSS} = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \sim \sigma^2 \chi^2_{n-p-1}$$

$\hat{\beta}$ and RSS are independent

$s^2 = \text{RSS}/(n-p-1)$ unbiased estimate of σ^2

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \sim t_{n-p-1}$$

$$\frac{(\hat{\beta}-\beta)^T(\mathbf{X}^T\mathbf{X})(\hat{\beta}-\beta)/p}{s^2} \sim F_{p+1,n-p-1}$$

where $c_{00}, c_{11}, ..., c_{pp}$ are the diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$.

Understanding

$$\begin{aligned} & \mathrm{cov}(\hat{\beta}, \hat{\epsilon}) \\ = \quad & \mathrm{cov}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon,\ (\mathbf{I}-\mathbf{H})\epsilon) \\ = \quad & \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{var}(\epsilon)(\mathbf{I}-\mathbf{H}) \\ = \quad & 0 \end{aligned}$$

because $\mathbf{H}$ is idempotent.

Confidence intervals

For example,

$$\hat{\beta}_j \pm t_{n-p-1}(\alpha/2) s \sqrt{c_{jj}}$$

is a confidence interval for β_j at confidence level $1-\alpha$. Here $t_{n-p-1}(\alpha/2)$ is the $1-\alpha/2$ percentile of the t-distribution with degree of freedom $n-p-1$.

Confidence intervals

For a given value of input $\mathbf{x}$ which is a $p+1$ vector (the first component is constant 1), its mean response is $\beta^T \mathbf{x}$. The confidence interval for this mean response is

$$\hat{\beta}^T \mathbf{x} \pm t_{n-p-1}(\alpha/2) s \sqrt{\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}$$

The confidence interval for β_j is a special case of the above formula by taking $\mathbf{x}$ as a vector that all zero except the $(j+1)$ entry corresponding β_j. (Because of β_0, β_j is at the $j+1$th position of $\hat{\beta}$.)

Prediction interval

- To predict the actual response y, rather than its mean, we would use the same point estimator $\hat{\beta}^T \mathbf{x}$, but the accuracy is much decreased as more uncertainty in the randomness of the actual response from the error is involved.
- The confidence interval, often called prediction interval, for y is

$$\hat{\beta}^T \mathbf{x} \pm t_{n-p-1}(\alpha/2) s \sqrt{1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}.$$

Outline

1. 3.1. Simple linear regression
2. 3.2 Multiple linear regression
3. 3.3. The least squares estimation
4. 3.4. The statistical properties of the least squares estimates.
5. 3.5 The variance decomposition and analysis of variance (ANOVA).
6. 3.6. More about prediction
7. 3.7. The optimality of the least squares estimation.
8. 3.8. Assessing the regression model.
9. 3.9. Variable selection.
10. 3.10. A comparison with KNN through a simulated example

Variance decomposition

Recall that

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2.$$

The common variance decomposition takes a similar form, but leaving out sample mean,

$$\|\mathbf{y} - \bar{y}\|^2 = \|\hat{\mathbf{y}} - \bar{y}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2;$$

which is often written as

$$\mathrm{SS}_{total} = \mathrm{SS}_{reg} + \mathrm{SS}_{error}.$$

Understanding

- SS_{total}, the total sum of squares, measures the total variation in response.
- SS_{reg}, the sum of squares due to regression or, more precisely, due to the inputs, measures variation in response explained by that of the inputs.
- SS_{error}, the sum of squares due to error, measures the size of randomness due to error or noise.

The ANOVA table

Source of Variation	SumOfSquares	Degree of Freedom	Mean Squared	F-statistic
Regression	SS_{reg}	p	MS_{reg}	MS_{reg}/MS_{error}
Error	SS_{error}	$n-p-1$	MS_{error}	
Total	SS_{total}	$n-1$		

where $\mathrm{MS}_{reg} = \mathrm{SS}_{reg}/p$ and $\mathrm{MS}_{error} = \mathrm{SS}_{error}/(n-p-1)$.
And the F-statistic follows $F_{p,n-p-1}$ distribution under the hypothesis that $\beta_1 = \beta_2 = ...\beta_p = 0$, i.e., all inputs are unrelated with the output.
The p-value is the probability for the distribution $F_{p+1,n-p-1}$ taking value greater than the value of the F-statistic.

General variance decomposition

- The above ANOVA is a special case of a general variance decomposition.
- Let $\mathcal{L}$ be the linear space spanned by $\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p$, all columns of $\mathbf{X}$.
- The linear model assumption:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

can be written as $E(\mathbf{y}) = \mathbf{X}\beta$, or

$$E(\mathbf{y}) \in \mathcal{L}.$$

- The fitted values $\hat{\mathbf{y}}$ is projection of $\mathbf{y}$ onto $\mathcal{L}$.
- $s^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n - p - 1)$ is the unbiased estimator of σ^2.