Linear Regression

Chapter 3

February 6, 2018

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

The formulation

- Consider linear regression model:

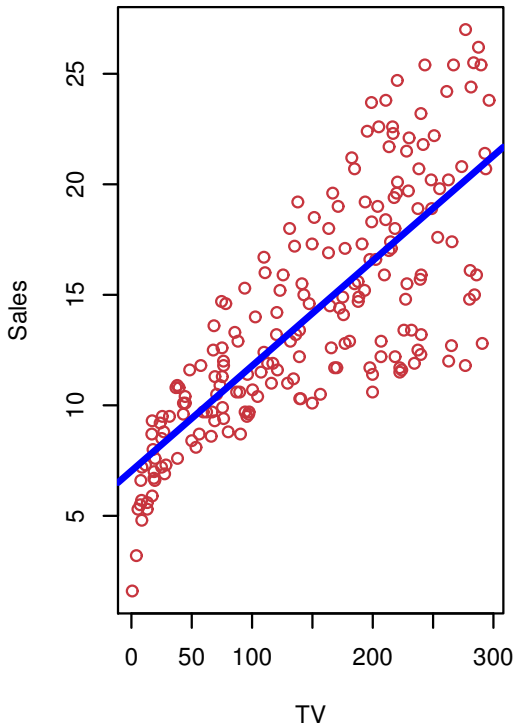
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, ..., n$$

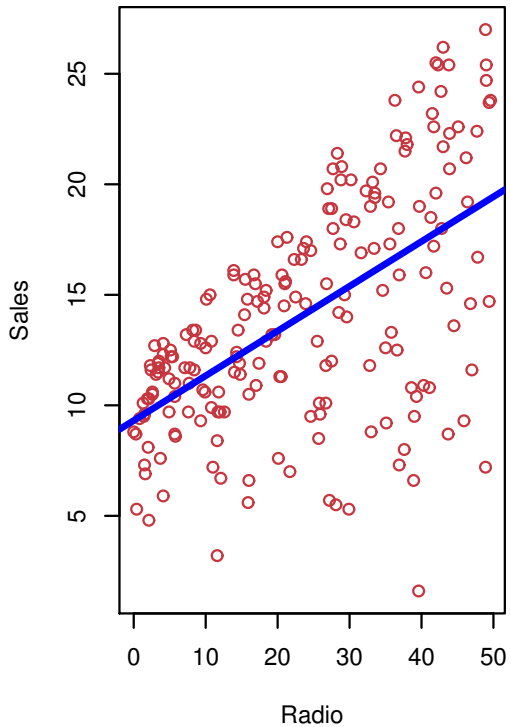
  where y_i, x_i are the i-th observation of the response and covariates. Here x_i is of 1-dimension.
- Responses are sometimes called dependent variables or outputs;
- covariates called independent variables or inputs or regressors.
- obtain the parameter estimation and making prediction of any responses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

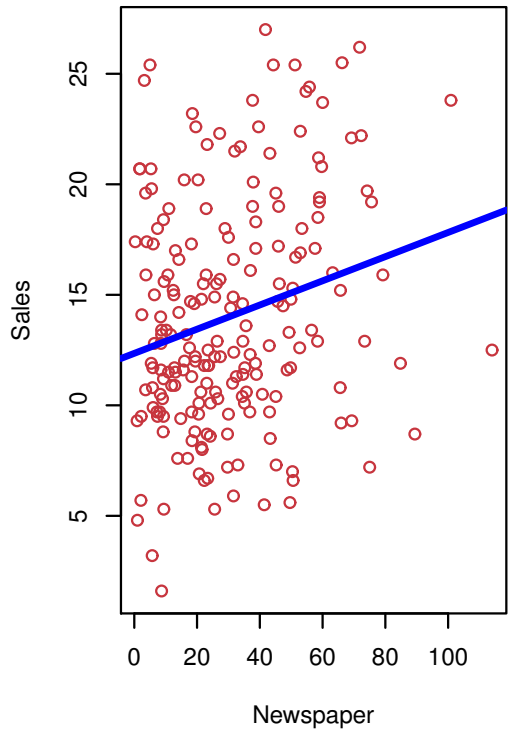
Example: Advertising data

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:

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

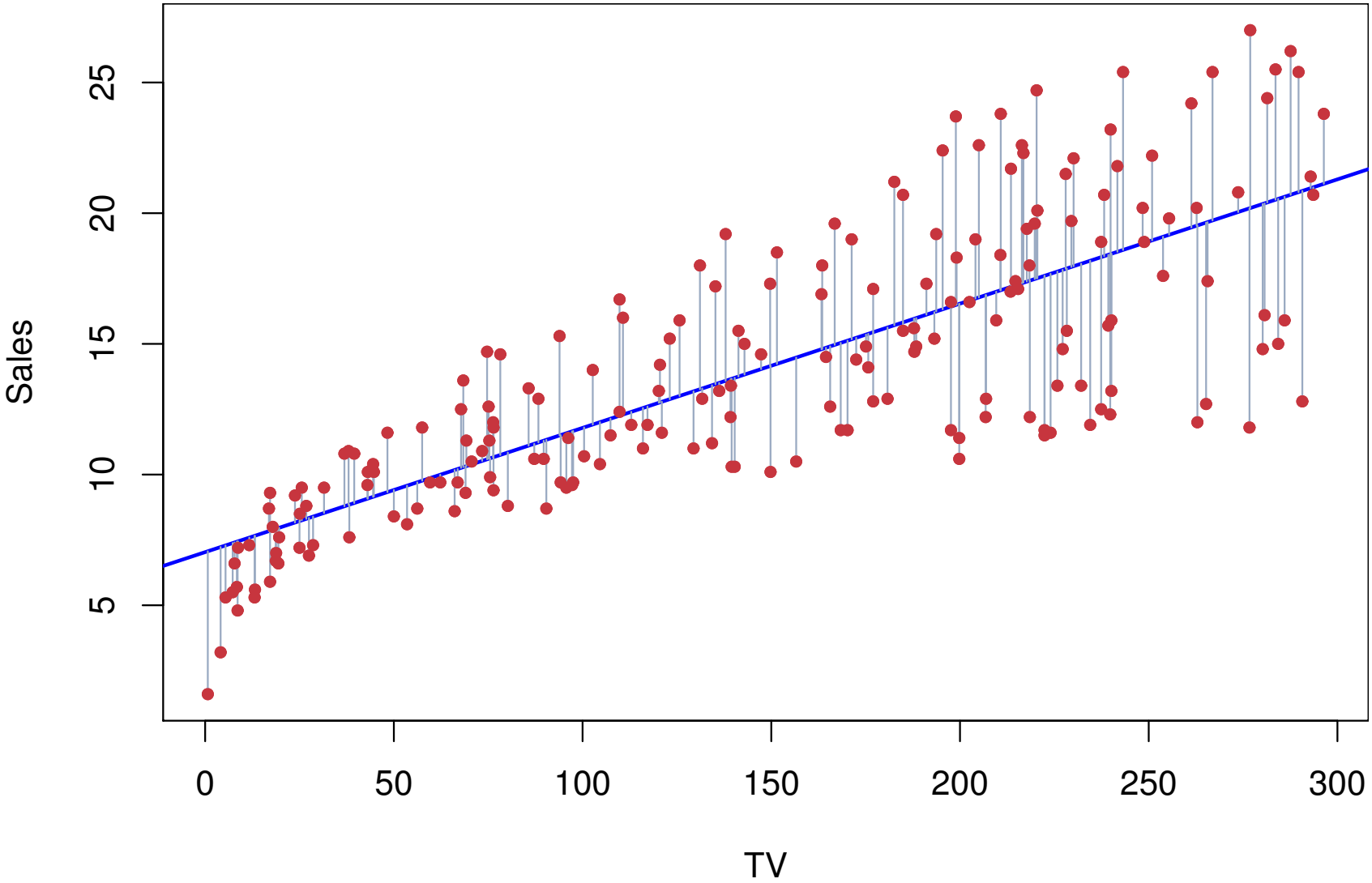
The estimator is

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

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

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

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:

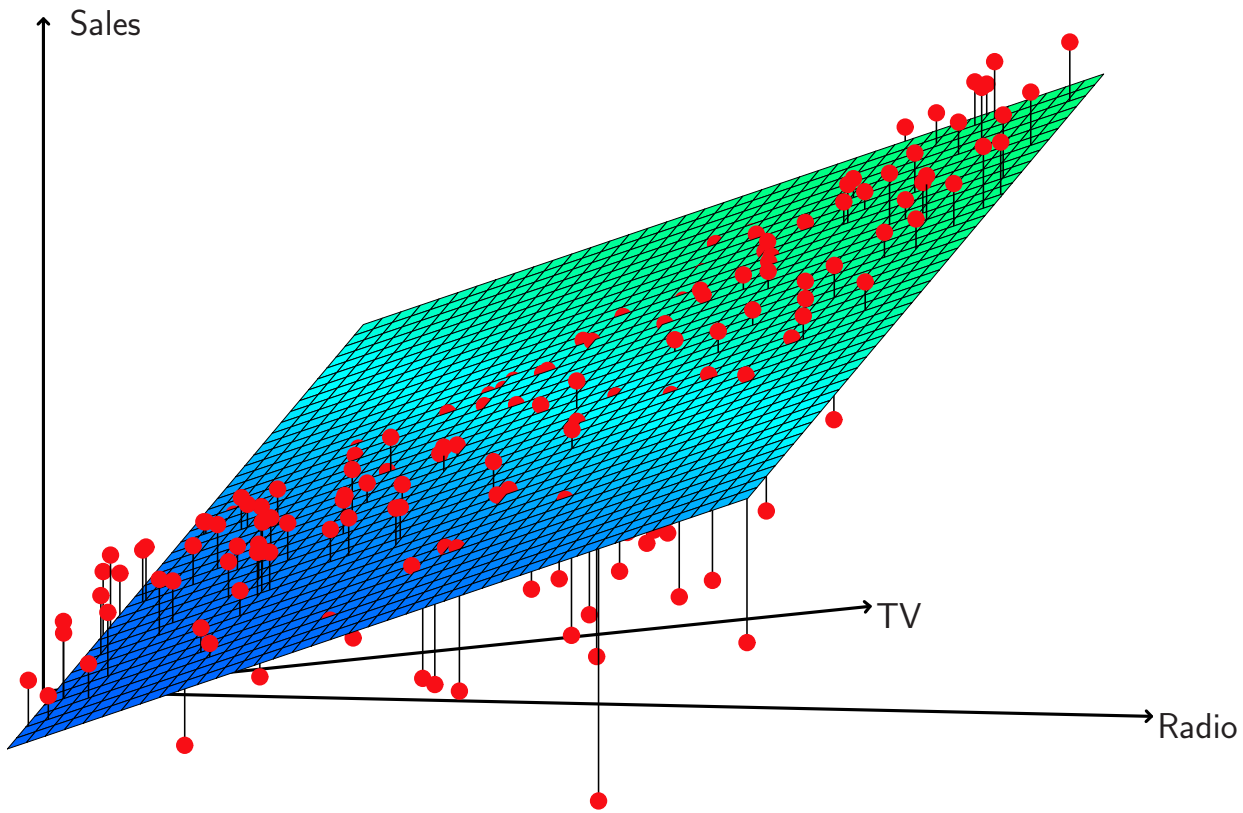
$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3})^2.$$

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

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

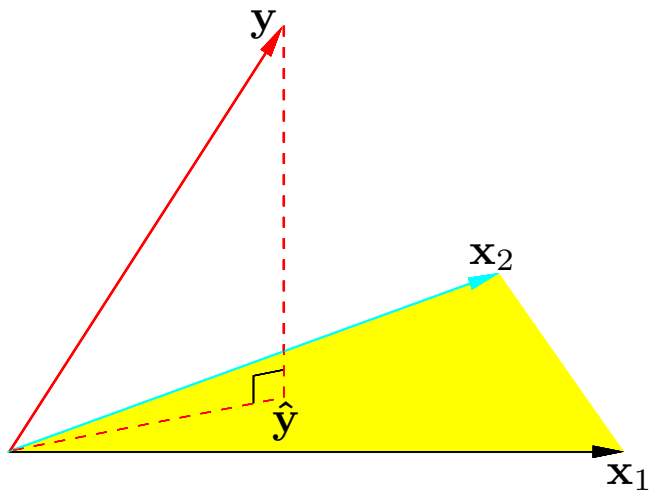


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector* $\mathbf{y}$ *is orthogonally projected onto the hyperplane spanned by the input vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. The projection* $\hat{\mathbf{y}}$ *represents the vector of the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

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

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

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

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

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

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 $\text{MS}_{reg} = \text{SS}_{reg}/p$ and $\text{MS}_{error} = \text{SS}_{error}/(n-p-1)$.
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.

- Further assume that
$$E(\mathbf{y}) \in \mathcal{L}_0$$
where $\mathcal{L}_0$ is some linear subspace of $\mathcal{L}$ of dimension $r < p + 1$.
- Let $\hat{\mathbf{y}}_0$ be the project of $\mathbf{y}$ on to $\mathcal{L}_0$.
- Pythagorean theorem implies
$$\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2$$
- By the same token, $s_0^2 = \|\mathbf{y} - \hat{\mathbf{y}}_0\|^2/(n - r)$ is the unbiased estimator of σ^2 under the hypothesis $E(\mathbf{y}) \in \mathcal{L}_0$.

The F-test

$$F = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/(p+1-r)}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-p-1)} \sim F_{p+1-r,n-p-r}$$

This F-statistic is used to test the hypothesis that $H_0 : E(\mathbf{y}) \in \mathcal{L}_0$, against the alternative H_a : otherwise.
The commonly considered hypothesis, as dealt with in the ANOVA table, $H_0 : \beta_1 = \cdots = \beta_p = 0$ can be formulated as $H_0 : E(\mathbf{y}) \in \mathcal{L}(\mathbf{1})$, where $\mathcal{L}(\mathbf{1})$ represent the linear space of a single vector $\mathbf{1}$.

Variable selection

- We may be concerned with a subset of the p variables are irrelevant with the response.
- Let the subset be denoted as $A = \{i_1, ..., i_r\}$, where $r \leq p$. Then, the null hypothesis is

$$H_0 : \beta_{i_1} = \beta_{i_2} = \cdots = \beta_{i_r} = 0,$$

which again is equivalent to

$$H_0 : E(\mathbf{y}) \in \mathcal{L}(A^c),$$

where $\mathcal{L}(A)$ is the linear space in R^n spanned by $\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_r}$, which is of r dimension.

The expected prediction error (EPE)

Consider a general model

$$y = f(\mathbf{x}) + \epsilon$$

Given a new input value $\mathbf{x}$, we wish to predict its response using $\hat{y}$, which is obtained from analysis of existing data.
The EPE is

$$\begin{aligned} E\{(y-\hat{y})^2\} &= E\{(\epsilon + f(\mathbf{x}) - \hat{y})^2\} \\ &= E\{(\epsilon + f(\mathbf{x}) - E(\hat{y}) + -\hat{y} + E(\hat{y}))^2\} \\ &= \sigma^2 + (f(\mathbf{x}) - E(\hat{y}))^2 + \text{var}(\hat{y}) \\ &= \text{Irreducible error} + \text{bias}^2 + \text{variance} \end{aligned}$$

Back to linear regression

- Here the covariate $\mathbf{x}$ has its first component being 1, slightly abusing the notatoin.
- Bias $= 0$

$$\begin{aligned} \text{variance} &= \text{var}(\mathbf{x}^T((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\ &= \text{var}(\mathbf{x}^T((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon) \\ &= \sigma^2\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x} \end{aligned}$$

- As a result, the EPE for this particular covairate $\mathbf{x}$ is

$$\sigma^2 + \sigma^2\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}$$

The average EPE

Take $\mathbf{x}$ as the observations $x_1, ..., x_n$ (first component is 1). Then the average EPE over all observations in the data is

$$\begin{aligned}
\sigma^2 + \sigma^2 \frac{1}{n} \sum_{i=1}^{n} x_i^T (\mathbf{X}^T \mathbf{X})^{-1} x_i &= \sigma^2 + \sigma^2 \frac{1}{n} \sum_{i=1}^{n} trace(x_i^T (\mathbf{X}^T \mathbf{X})^{-1} x_i) \\
&= \sigma^2 + \sigma^2 \frac{1}{n} \sum_{i=1}^{n} trace((\mathbf{X}^T \mathbf{X})^{-1} x_i x_i^T) \\
&= \sigma^2 + \sigma^2 \frac{1}{n} trace(\sum_{i=1}^{n} (\mathbf{X}^T \mathbf{X})^{-1} x_i x_i^T) \\
&= \sigma^2 + \sigma^2 \frac{1}{n} trace(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}) \\
&= \sigma^2 + \sigma^2 \frac{1}{n} trace(I_{p+1}) \\
&= \sigma^2 (1 + \frac{p+1}{n})
\end{aligned}$$

The average EPE

The average EPE reflects the prediction accuracy of the model.
Suppose p is large relative to n, and among p inputs, only a few, say q, of them are relevant to the response. Assume q is far smaller than p.
For simplicity, say $p \approx n/2$, if we use all p variables, the average EPE is

$$\sigma^2(1 + \frac{p+1}{n}) \approx \frac{3}{2}\sigma^2$$

If we use those relevant q inputs, the average EPE is

$$\sigma^2(1 + \frac{q+1}{n}) \approx \sigma^2.$$

This implies, using more inputs, although always increase the R-squared, may reduce the prediction accuracy!!!

Superiority of LSE

- Easy computation
- consistency
- efficiency, etc.
- BLUE (best linear unbiased estimator), among estimates of β, that are linear unbiased estimates: $\sum_{i=1}^{n} \mathbf{a}_i y_i$, with $\mathbf{a}_i$ being nonrandom.

- **Theorem** (Gauss-Markov Theorem). Among all linear unbiased estimates, the least squares estimate has the smallest variance, thus smallest mean squared error.

Proof

- Let $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n)$. Then, unbiased estimate of β is $\mathbf{Ay}$ with mean $\mathbf{AX}\beta = \beta$ by the unbiasedness, and variance matrix $\mathbf{AA}^T$.
- Write $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}$. Then, $\mathbf{DX} = 0$.
- Then,

$$\mathbf{AA}^T = (\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{DD}^T \geq (\mathbf{X}^T\mathbf{X})^{-1}.$$

  Here the inequality is for symmetric matrices, i.e., $\mathbf{A} \geq \mathbf{B}$ is defined as $\mathbf{A} - \mathbf{B}$ is nonnegative definite.

a). Error distribution (normality check).

- Non-normality may cause the normality-based inference such as t-test and F-test being inaccurate.
- Use graphics, such histogram, boxplot and qqnorm to visualize the the distribution of the residuals.

Adjusted residuals

- the residuals $\hat{\epsilon}_i$ does not follow the distribution $N(0, \sigma^2)$, even if all model assumptions are correct!
- $$\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\epsilon \sim N(0, \sigma^2(\mathbf{I} - \mathbf{H})).$$

  So, $\hat{\epsilon}_i \sim N(0, (1 - h_{ii})\sigma^2)$.
- Training error is one of the measurement of the quality of fit.
- An (internally) studentized residual is

$$e_i = \frac{\hat{\epsilon}_i}{s\sqrt{1 - h_{ii}}}.$$

- A more appropirate one is the (externally) studentized residual which uses an s from the least squares fitting by deleting the i-th observation.

b). Homoscedasticity versus heteroscedasticity.

- Heteroscedasticity can cause the estimate being not the optimal one
- May be fixed by weighted least squares estimation.
- Use scatter plot of residuals against the fitted values to check the heteroscedasticity (the variance of the errors are not equal).

c). Error dependence.

- The error dependence cause the inference to be incorrect.
- Use autocorrelation of the residuals (ACF) to check the independence assumption of the errors.
- One can also use Durbin-Watson test, which tests whether the first few autocorrelations are 0.

d). Leverage and Cook's D.

- Recall the hat matrix

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

  Let $h_{ij} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_j$ be the (i, j) elements of $\mathbf{H}$.
- The leverage of the i-th observation is just the i-th diagonal element of $\mathbf{H}$, denoted as h_{ii}.
- A high leverage implies that observation is quite influential. Note that the average of h_{ii} is $(p + 1)/n$.
- So, if h_{ii} is greater than $2(p + 1)/n$, twice of the average, is generally considered large.

- The Cook'D is often used measure how important an observation is. Cook's D is defined

$$D_i = \frac{\sum_{k=1}^{n} (\hat{y}_k - \hat{y}_k^{(-i)})^2}{(p+1)s^2}$$

where $\hat{y}_k$ is the k-th fitted value; and $\hat{y}_k^{(-i)}$ is the k-th fitted value by deleting the i-th observation.

- If D_i is large, it implies once i-th observation is not available, the prediction would be much different, thus reflecting the importance of this observation.
- In general, the observations with large D_i, such as larger than a quarter of the sample size, may be considered influential.

e). Multicollinearity.

- The multicollinarity can cause the parameter estimation to be very unstable.
- Suppose two inputs are strongly correlated, their separate effect on regression is difficult to identify from the regression.
- When data change slightly, the two regression coefficients can differ greatly, though their joint effect may stay little changed.
- Use variance inflation factor (VIF) to measure one input's correlation with the others.

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

- The largest value of VIF, ∞, means this input is perfectly linearly related with the other inputs.
- The smallest value of VIF, 1, means this input is uncorrelated with the other inputs.
- In general, variable selection methods may be used to reduce the number of highly correlated variables.

- Variable selection, or more generally, model selection, is an important tool in minimizing prediction error.
- There are substantial research development regarding methods of model selection.
- The aim is to minimize generalization error or prediction error.
- The naive approach is to exhaust all models. However, with the curse of dimensionality, this is quickly prohibitive when the number of variables increase.
- More sophisticated methods such as cross validation or regularization methods, such as LASSO (Tibshirani 1996).

Caution

- More inputs do not imply better prediction, particularly if the inputs in the model are irrelevant with the response.
- Moreover, more inputs also imply more danger of overfit, resulting in small training error but large test error.
- Here we introduce more basic and simple methods.

a). Adjusted R-squared.

- The R-squared is the percentage of the total variation in response due to the inputs. The R-squared is commonly used as a measurement of how good the linear fit is.
- However, a model with larger R-squared is not necessarily better than another model with smaller R-squared!
- If model A has all the inputs of model B, then model A's R-squared will always be greater than or as large as that of model B.
- If model A's additional inputs are entirely uncorrelated with the response, model A contain more noise than model B. As a result, the prediction based on model A would inevitably be poorer or no better.

- b). Recall that the R-squared is defined as:

$$R^2 = \frac{\text{SS}_{reg}}{\text{SS}_{total}} = 1 - \frac{\text{SS}_{error}}{\text{SS}_{total}}$$

where $\text{SS}_{error} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$ is often called residual sum of squares (RSS).

- The adjusted R-squared, taking into account of the degrees of freedom, is defined as

$$\begin{aligned} \text{adjusted } R^2 &= 1 - \frac{\text{MS}_{error}}{\text{MS}_{total}} \\ &= 1 - \frac{\text{SS}_{error}/(n-p-1)}{\text{SS}_{total}/(n-1)} \\ &= 1 - \frac{s^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)} \end{aligned}$$

  With more inputs, the R^2 always increase, but the adjusted R^2 could decrease since more inputs is penalized by the smaller degree of freedom of the residuals.

- The adjusted R-squared is preferred over the R-squared in evaluating models.

b). Mallows' C_p.

Recall that our linear model (2.1) has p covariates, and $s^2 = \text{SS}_{error}/(n-p-1)$ is the unbiased estimator of σ^2.
Assume now more covariates are available. Suppose we use only p of the K covariates with $K \geq p$.
The statistic of Mallow's C_p is defined as

$$\frac{\text{SS}_{error}(p)}{s_K^2} - 2(p+1) - n.$$

where SS_{error} is the residual sum of squares for the linear model with p inputs and s_K^2 is the unbiased estimator of σ^2 based on K inputs.
The smaller Mallows' C_p is, the better the model is.
The following AIC is more often used, despite that Mallws' C_p and AIC usually give the same best model.

c). AIC.

AIC stands for Akaike information criterion, which is defined as

$$\text{AIC} = \log(s^2) + 2(1+p)/n,$$

for a linear model with p inputs, where $s^2 = \text{SS}_{error}/(n-p-1)$. AIC aims at maximizing the predictive likelihood. The model with the smallest AIC is preferred.

The AIC criterion is try to maximize the expected predictive likelihood. In general, it can be roughly derived in the following. Let θ be a parameter of d dimension. $\hat{\theta}$ is the maximum likelihood estimator of θ based on observations $y_1, ..., y_n$. Let θ_0 be the true (unknown) value of θ, and $\mathcal{I}(\theta_0)$ be the Fisher information.

the (expected) predictive log-likelihood

$$
\begin{aligned}
& E(\log f(Y|\theta))|_{\theta=\hat{\theta}} \\
\approx\ & E(\log f(Y|\theta))|_{\theta=\theta_0} - \frac{1}{2}(\hat{\theta}-\theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta}-\theta_0) \\
\approx\ & \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\hat{\theta}) - \frac{1}{2}(\hat{\theta}-\theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta}-\theta_0) \\
& -\frac{1}{2}(\hat{\theta}-\theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta}-\theta_0) \\
\approx\ & \frac{1}{n}(\text{maximum log likelihood}) - (\hat{\theta}-\theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta}-\theta_0) \\
\approx\ & \frac{1}{n}(\text{maximum log likelihood}) - d/n
\end{aligned}
$$

The approximations are due to the Taylor expansion.

- Then, maximizing the above predictive likelihood is the same as minimize

$$-2(\text{maximum log likelihood}) + 2d$$

where, the first term is called deviance. In the case of linear regression with normal errors, the deviance is the same as $\log(s^2)$.

d). BIC.

- BIC stands for Schwarz's Bayesian information criterion, which is defined as

$$\text{BIC} = \log(s^2) + (1 + p)\log(n)/n,$$

for a linear model with p inputs. Again, the model with the smallest BIC is perferred. The derivation of BIC results from Bayesian statistics and has Bayesian interpretation. It is seen that BIC is formally similar to AIC. The BIC penalizes more heavily the models with more number of inputs.

Simulated Example of KNN

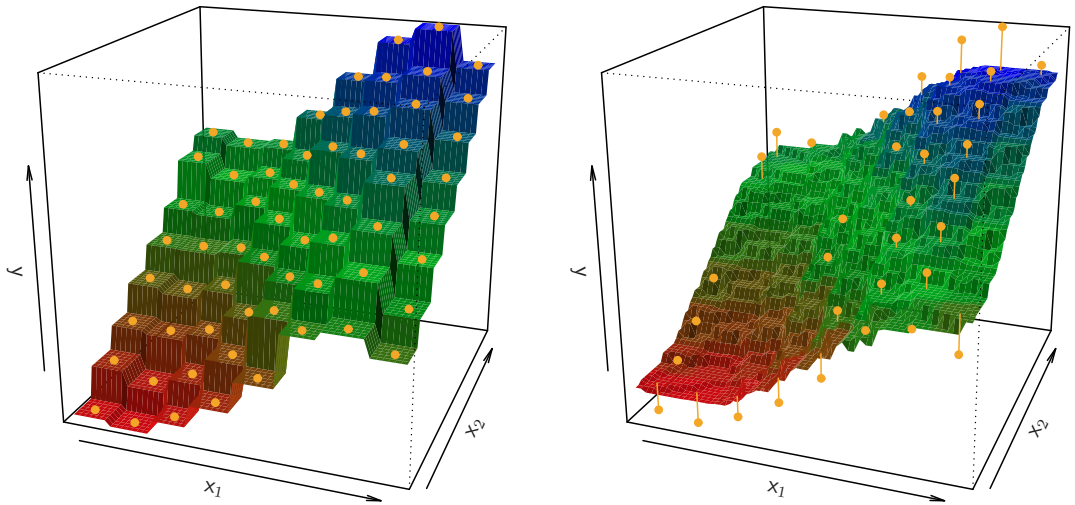


Figure: 3.16. Plots of $\hat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots). Left: K = 1 results in a rough step function fit. Right: K = 9 produces a much smoother fit.

Simulated Example of KNN

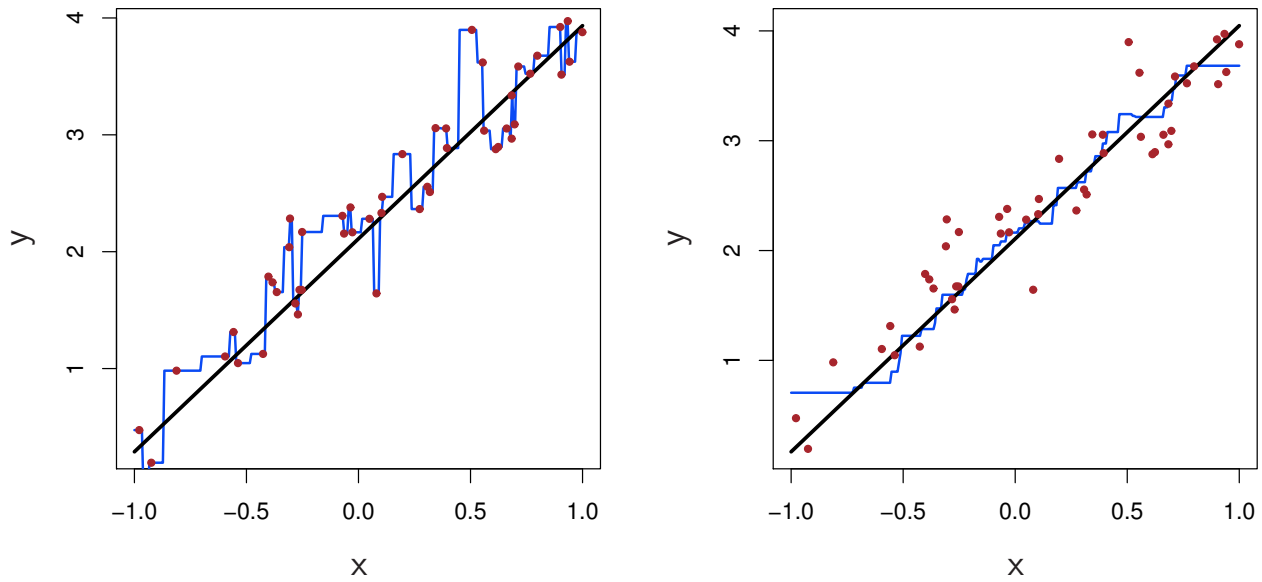
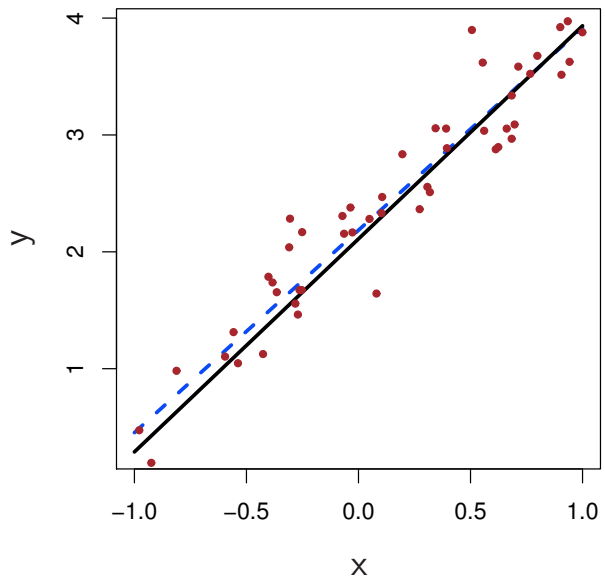


Figure: 3.17. Plots of $\hat{f}(X)$ using KNN regression on a one-dimensional data set with 100 observations. The true relationship is given by the black solid line. Left: The blue curve corresponds to K = 1 and interpolates (i.e. passes directly through) the training data. Right: The blue curve corresponds to K = 9, and represents a smoother fit.

Simulated Example of KNN



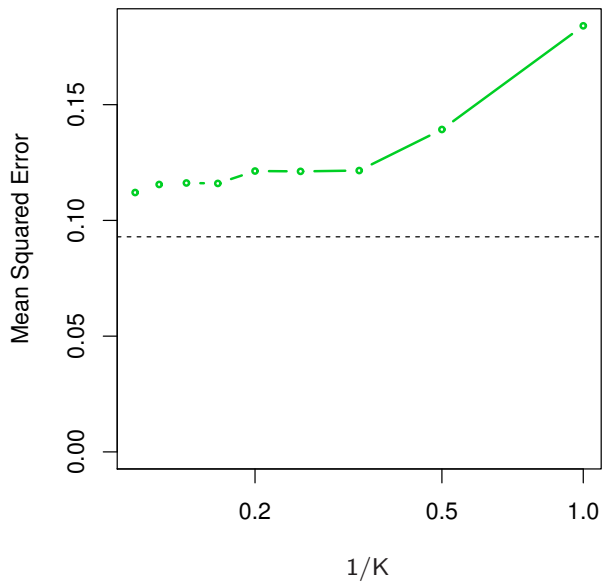
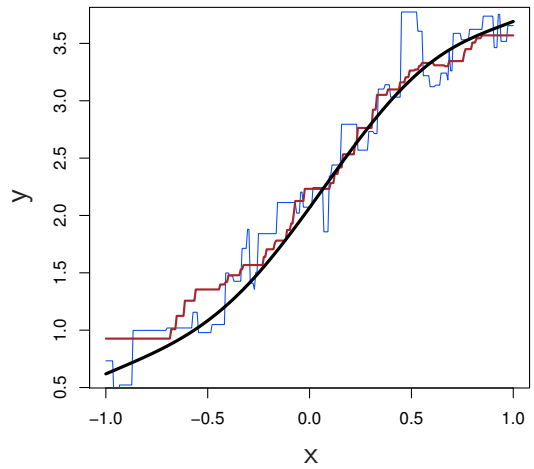
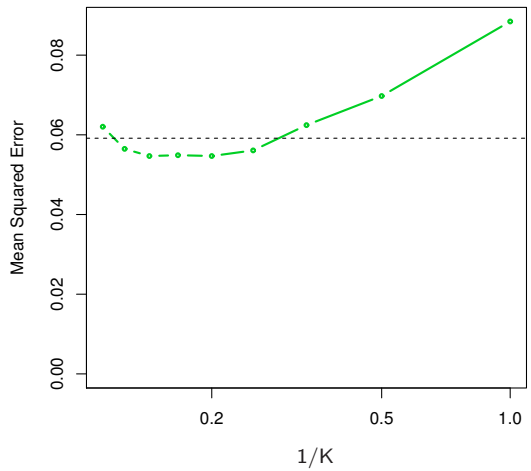
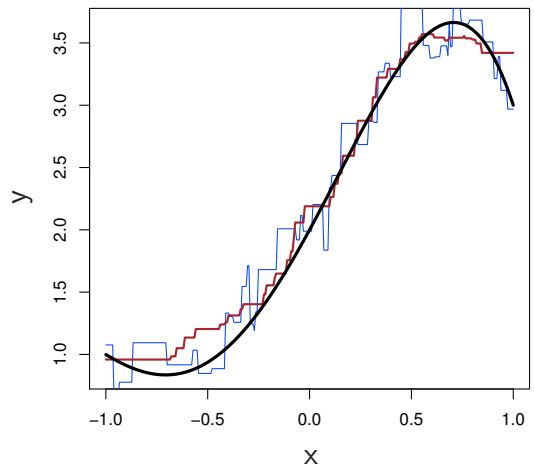


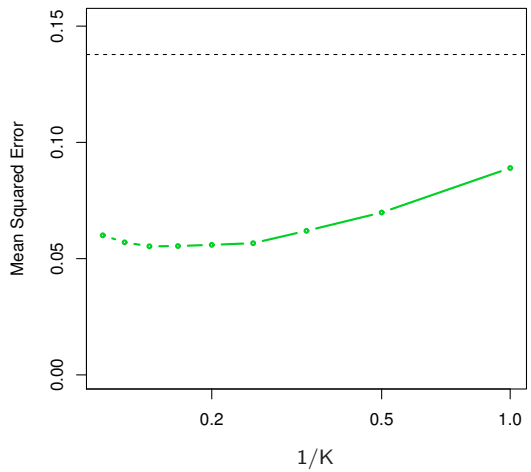
FIGURE 3.18.The same data set shown in Figure 3.17 is investigated further. Left: The blue dashed line is the least squares fit to the data. Since $f(X)$ is in fact linear (displayed as the black line), the least squares regression line provides a very good estimate of f(X). Right: The dashed horizontal line represents the least squares test set MSE, while the green solid line corresponds to the MSE for KNN as a function of $1/K$ (on the log scale). Linear regression achieves a lower test MSE than does KNN regression, since $f(X)$ is in fact linear. For KNN regression, the best results occur with a very large value of K, corresponding to a small value of $1/K$.

Simulated Example of KNN









Simulated Example of KNN

FIGURE 3.19. Top Left: In a setting with a slightly non-linear relationship between X and Y (solid black line), the KNN fits with $K = 1$ (blue) and $K = 9$ (red) are displayed. Top Right: For the slightly non-linear data, the test set MSE for least squares regression (horizontal black) and KNN with various values of $1/K$ (green) are displayed. Bottom Left and Bottom Right: As in the top panel, but with a strongly non-linear relationship between X and Y.

Simulated Example of KNN

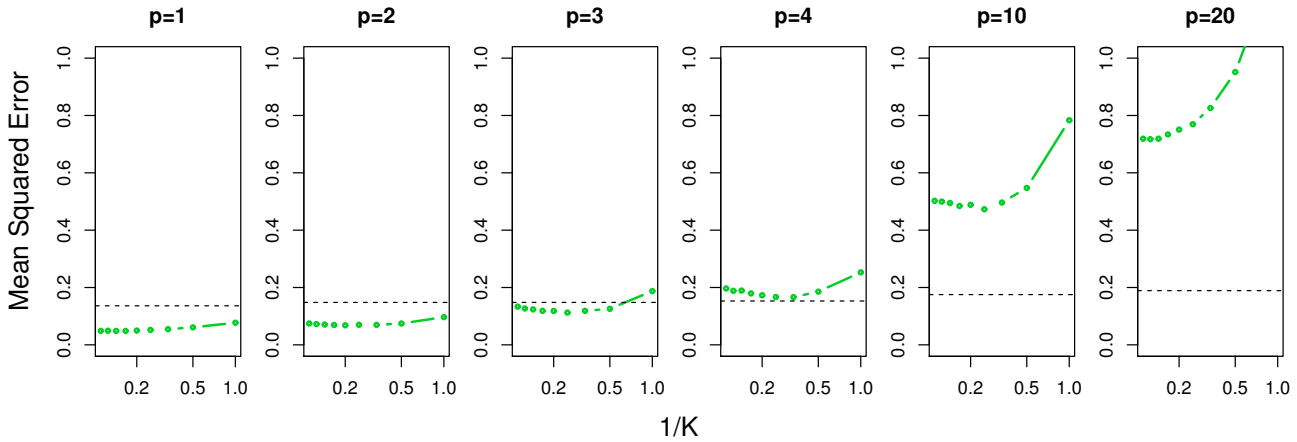


Figure: 3.20.Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non linear in the first variable, as in the lower panel in Figure 3.19, and does not depend on the additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNNs performance degrades much more quickly as p increases.

Homework

- ISLR Chapter 3: 1; 2; 5; 8.

End of Chapter 3