Lecture 1
Theory of Linear Models

Chao Song

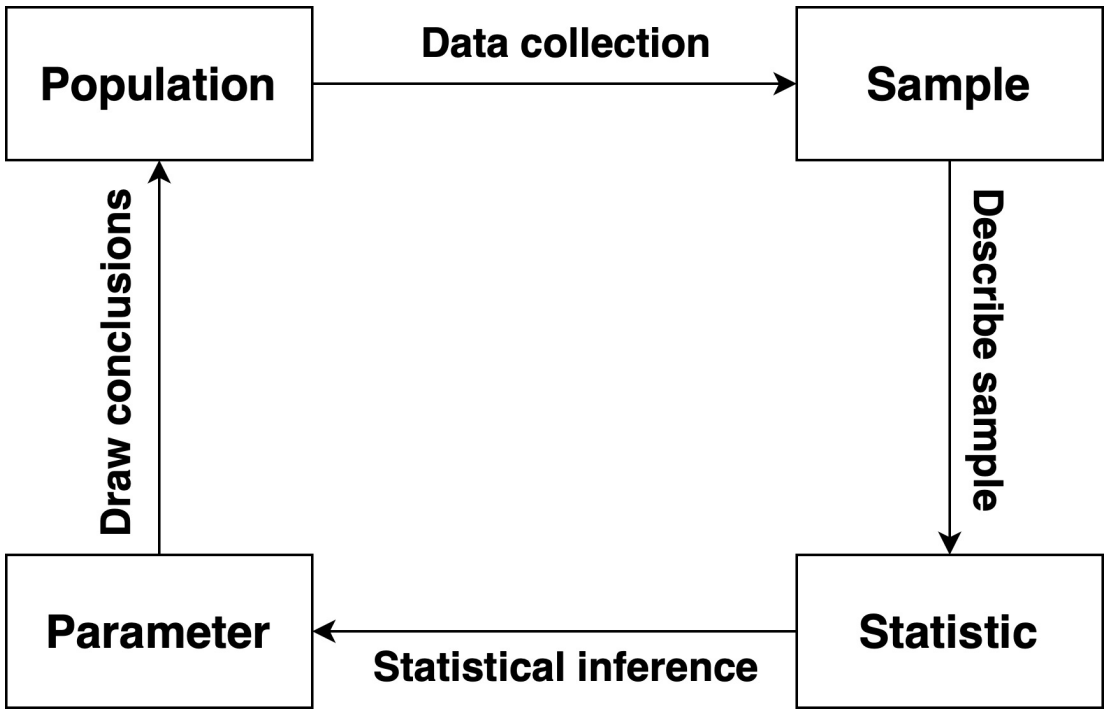
College of Ecology

Lanzhou University

July 15, 2025

Essential tasks of statistics

Parameter estimation and **hypothesis testing** are two essential tasks of statistics. Roughly speaking, parameter estimates tells us what the value of a parameter could be and hypothesis testing aims at reaching a decision about whether or not we reject a hypothesis about the value of the parameter.

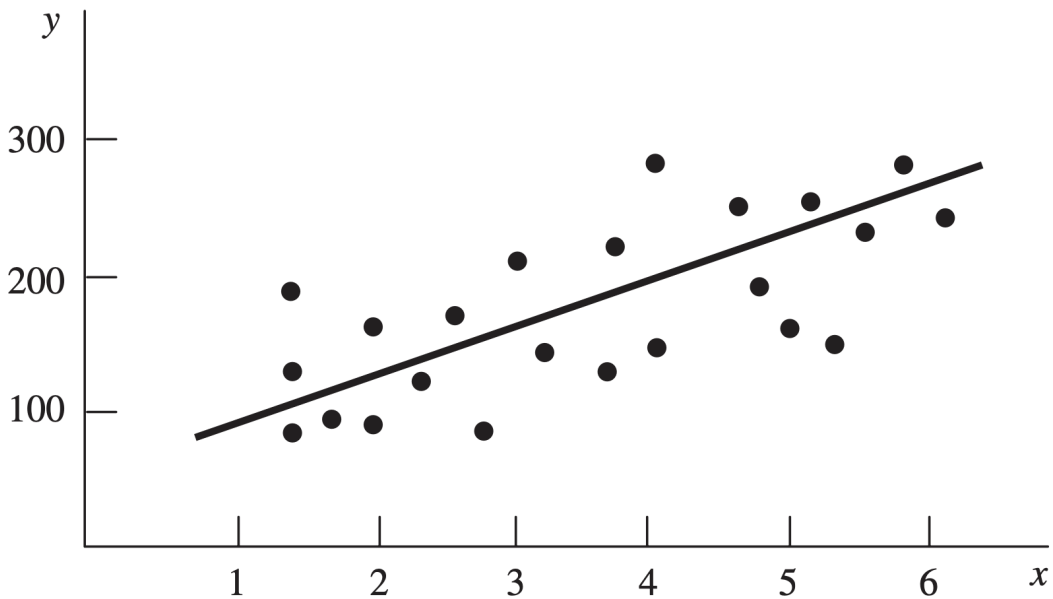


Deterministic and probabilistic model

While deterministic models are common in physics and mathematics, it is rarely applicable in ecology.

- Contexts that influence the dependent variable may vary;
- We usually cannot measure things without error.

Often, we encounter data that are noisy. The average of Y seems to change with X but a deterministic relationship cannot exactly fit the data.



Deterministic and probabilistic model

In these scenarios, statisticians use probabilistic models. For example, we may represent the data in the previous figure by the model

$$E(Y) = \beta_0 + \beta_1 X$$

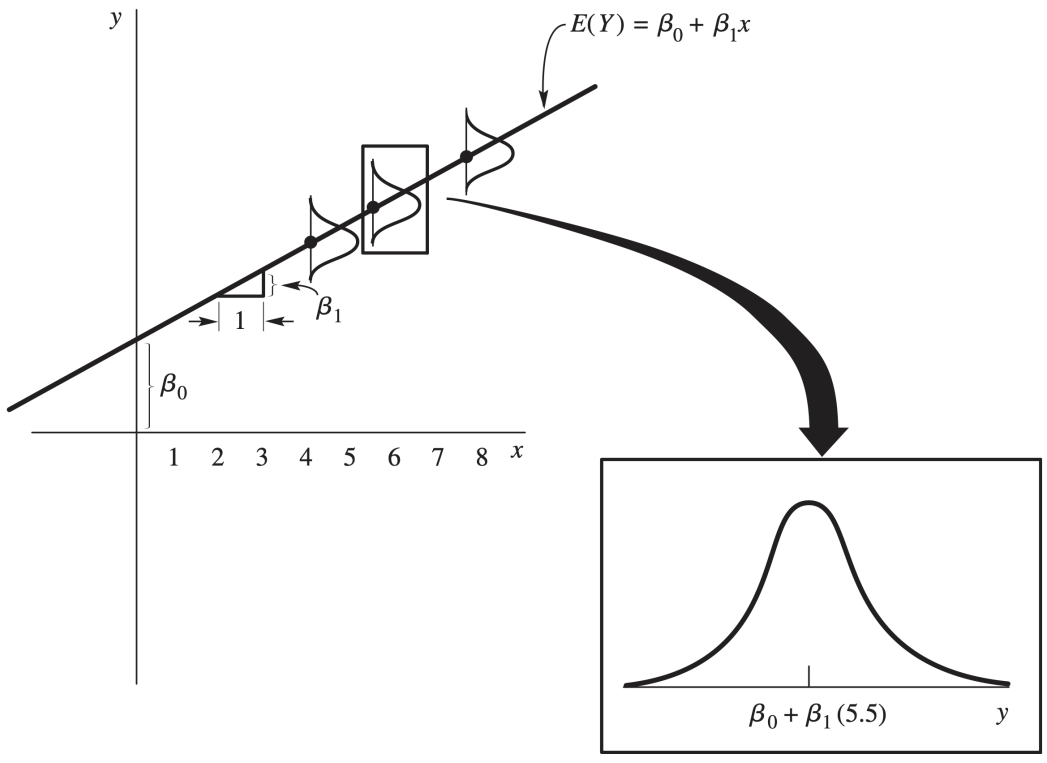
Each observation deviates from the mean by an unknown random error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where ε is a unknown random error. We often further assume that it possess a specified probability distribution with mean 0.

Deterministic and probabilistic model

In the model $Y = \beta_0 + \beta_1 X + \varepsilon$, we assume that there is a population of possible values of Y for a particular value of X. The distribution has a mean that is predicted by the deterministic part of the model, i.e., $\beta_0 + \beta_1 X$. The observation deviates from the mean by the random component ε.



Linear models

Definition: A linear model relating a random response Y to a set of independent variables $X_1, X_2, \ldots, X_k$ is of the form

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

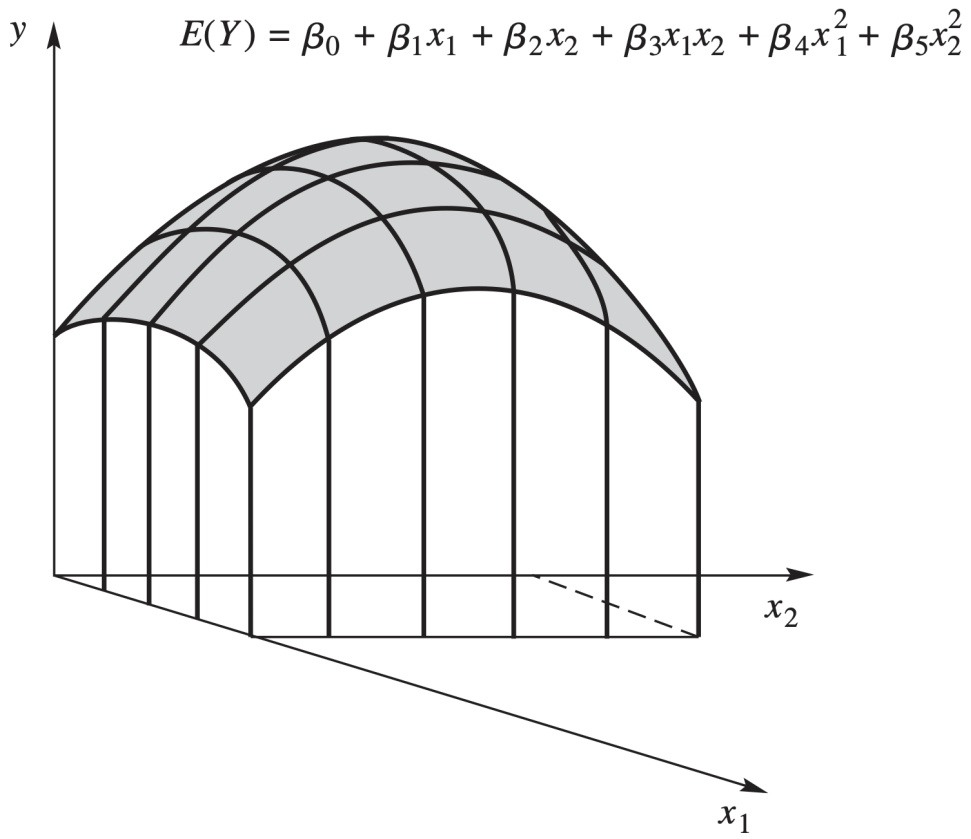
where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters, ε is a random variable and the variables $X_1, X_2, \ldots, X_k$ assume known value. We will assume $E(\varepsilon) = 0$ and hence that

$$E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

The term “linear” means that the mean of dependent variable $E(Y)$ is a linear function of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$. It is not necessarily a linear function of X. For example, $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ or $Y = \beta_0 + \beta_1 \ln(X)$ are also a linear model.

Linear model

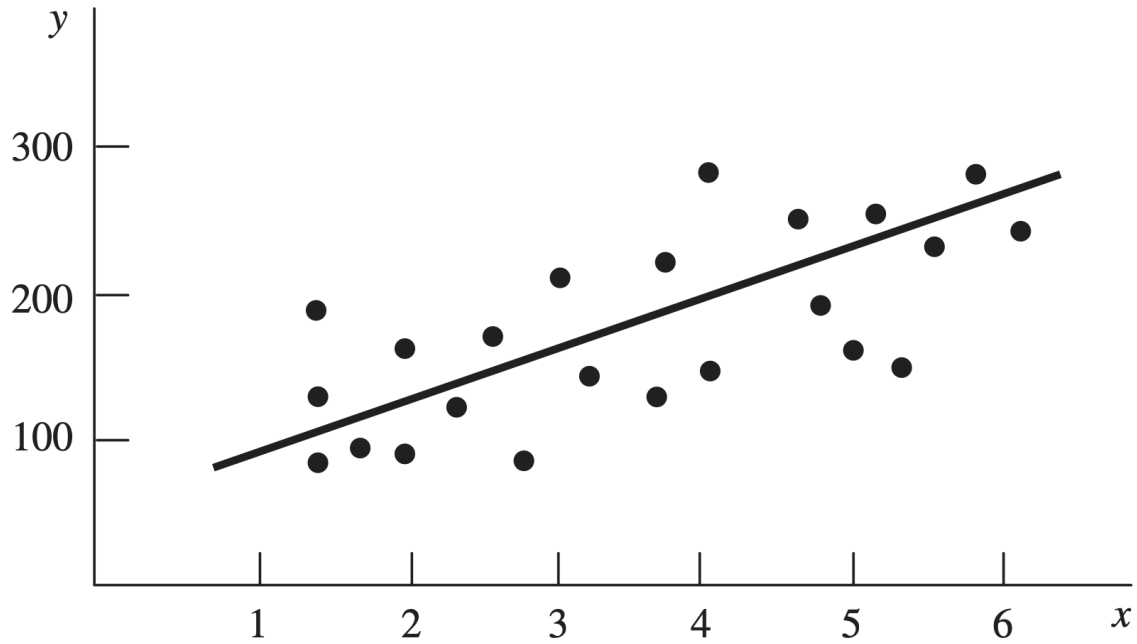
If the model is of the form $Y = \beta_0 + \beta_1 X$, where X is a continuous variable, the model is a simple linear regression. If the model contains multiple continuous independent variables, the model is called multiple linear regression. Below is an example of multiple linear regression:



The method of least square

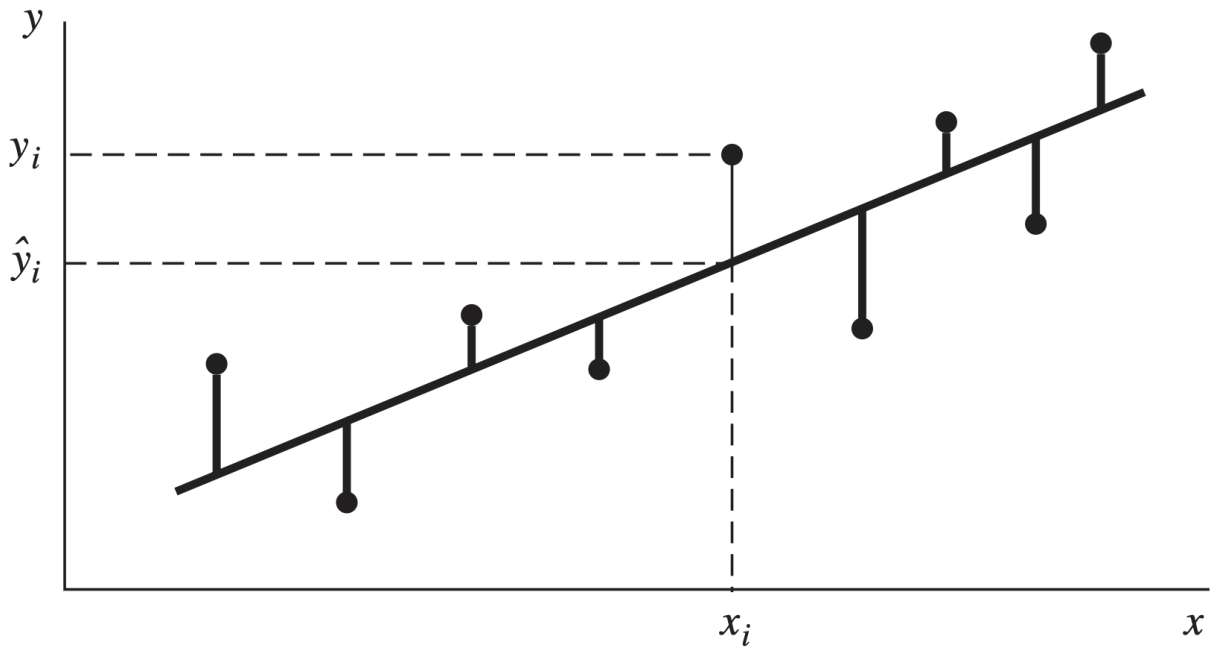
How do we estimate parameter in a linear model?

Intuitively, we want to fit a line through the data and we want the difference between the observed values and the corresponding points on the fitted line to be “small” in some overall sense.



The method of least square

A convenient way to accomplish this, and one that yields estimators with good properties, is to minimize the sum of squares of the vertical deviations from the fitted line. This method is called the method of **least squares**.



Graphic illustration of the method of least squares

Method of least squares

In a simple linear regression $Y = \beta_0 + \beta_1 X$, let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the estimates of model parameters, and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ denotes the predicted value of y_i based on the regression. The sum of squares of deviations to be minimized is

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}i)^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

The solution to the least square equations are

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

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

Properties of least square estimators

We have the following properties of the least square estimators:

- The least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased;
- $Var(\hat{\beta}_0) = c_{00}\sigma^2$, where $c_{00} = \sum x_i^2 / nS_{xx}$;
- $Var(\hat{\beta}_1) = c_{11}\sigma^2$, where $c_{11} = 1/S_{xx}$;
- $Cov(\hat{\beta}_0, \hat{\beta}_1) = c_{01}\sigma^2$, where $c_{01} - \overline{x}/S_{xx}$
- $s^2 = SSE/(n-2)$ is an unbiased estimator for σ^2.

All these properties are derived based on the assumption that we have a linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $E(\varepsilon_i) = 0$ and are independent. These properties do not require any distributional assumption about ε_i.

Properties of least square estimators

If we further assume that $\varepsilon_i \sim N(0, \sigma^2)$, then

- Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed;
- The random variable $(n-2)s^2/\sigma^2$ has $\chi^2(n-2)$;
- The statistic $s^2 = SSE/(n-2)$ is independent of both $\hat{\beta}_0$ and $\hat{\beta}_1$.

Inferences concerning the parameters β_i

In a linear regression, if the random error ε is normally distributed, we can show that $\hat{\beta}_i$ is an unbiased, normally distributed estimator of β_i with

$$Var(\hat{\beta}_0) = c_{00}\sigma^2, \quad c_{00} = \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}$$

$$Var(\hat{\beta}_1) = c_{11}\sigma^2, \quad c_{11} = \frac{1}{S_{xx}}$$

For each $\hat{\beta}_i$, we thus have

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{c_{ii}}\sigma} \sim N(0,1)$$

We have also shown $s^2 = SSE/(n-2)$ is independent of $\hat{\beta}_i$ and that

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$$

Inferences concerning the parameters β_i

This allows us to construct a test statistics for H_0: $\beta_i = \beta_{i0}$:

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{\sqrt{c_{ii}}\sigma} \Big/ \sqrt{\frac{(n-2)s^2}{\sigma^2} \Big/ (n-2)}$$

$$= \frac{\hat{\beta}_i - \beta_{i0}}{\sqrt{c_{ii}}s} \sim t(n-2)$$

This result also suggests that we can construct $100(1-\alpha)\%$ confidence interval for β_i as:

$$\hat{\beta}_i \pm t_{\alpha/2}(n-2)s\sqrt{c_{ii}}$$

Inferences concerning the parameters β_i

The t-test for each β_i can also be done based on F-distribution as

$$\frac{\frac{(\hat{\beta}_i-\beta_{i0})^2}{c_{ii}\sigma^2}/1}{\frac{(n-2)s^2}{\sigma^2}/(n-2)} = \frac{SSH/1}{SSE/(n-2)} \sim F_{1,n-2}$$

Here, we refer to $\frac{(\hat{\beta}_i-\beta_{i0})^2}{c_{ii}}$ as SSH and $(n-2)s^2$ as SSE. The F test statistic is constructed from the so called "sum of squares". This is an important concept in hypothesis testing in linear models.

While t-test can be used to test hypothesis concerning a single parameter, F-test constructed from various "sum of squares" provides a general way of hypothesis testing in linear models.

Inferences concerning linear functions of parameters

In addition to making inference about a single β_i, we frequently are interested in linear functions of model parameters. For example, we may wish to make inference about

$$\theta = a_0\beta_0 + a_1\beta_1$$

Because $\hat{\beta}_i$ are all normally distributed, $\hat{\theta}$ as a linear function of $\hat{\beta}_i$ also has a normal distribution. This allows us to construct a t statistic for testing H_0: $\theta = \theta_0$ as

$$T = \frac{\frac{\hat{\theta}-\theta_0}{\sqrt{c_\theta}\sigma}}{\sqrt{\frac{(n-2)s^2}{\sigma^2}/(n-2)}} = \frac{\hat{\theta}-\theta_0}{s\sqrt{c_\theta}} \sim t(n-2)$$

A $100(1-\alpha)\%$ confidence interval for θ is thus

$$\hat{\theta} \pm t_{\frac{\alpha}{2}}(n-2)s\sqrt{c_\theta}$$

Inference about predicted mean

An important application of making inference about linear functions of model parameters is to predict the mean of response variable at a new value of independent variable. Suppose we have already fitted a linear model. We want make inference about the mean of Y at $x = x^*$,

We estimate the mean of Y at x^* by $E(Y^*) = \hat{\beta}_0 + \hat{\beta}_1 x^*$. Note here $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is a linear function of model parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ where $a_0 = 1$ and $a_1 = x^*$. Thus, using results from previous slides:

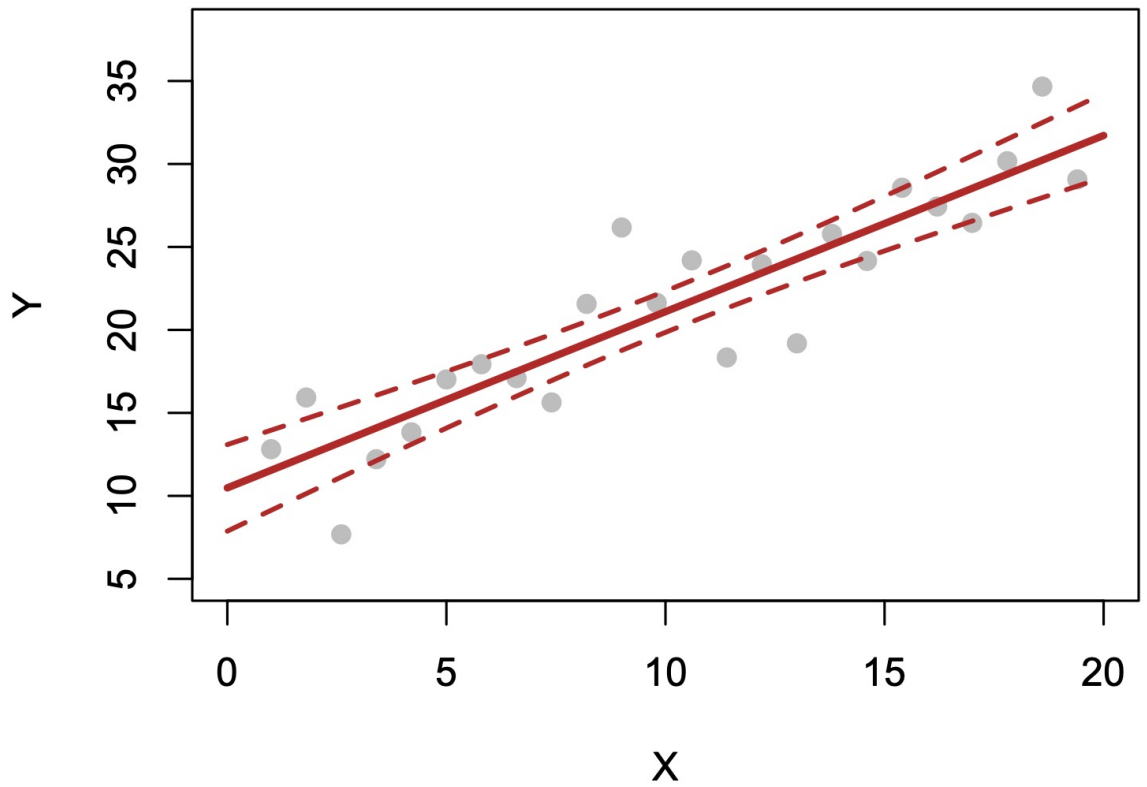
A $100(1-\alpha)$%Confidence interval for $E(Y^*) = \beta_0 + \beta_1 x^*$ is

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2}}(n-2)s\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$

Inference about predicted mean

On the width of the confidence interval for $E(Y)$:

- The width is the narrowest at $x = \overline{x}$;
- The width decreases with S_{xx}, suggesting that spreading x out helps improve the precision of predicting the mean.



Inference about predicted value of Y

In addition to making inferences about the mean of Y at x^*, can we make prediction about the value Y at x^*, namely Y^*?

Notice that Y^* **is a random variable, not a parameter**; predicting its value therefore represents a departure from previous objective of making inferences about model parameters.

In a linear model assuming normal error, Y^* is normally distributed with mean $\beta_0 + \beta_1 x^*$. It is thus reasonable to use $\hat{\beta}_0 + \hat{\beta}_1 x^*$ as a predictor of Y^*.

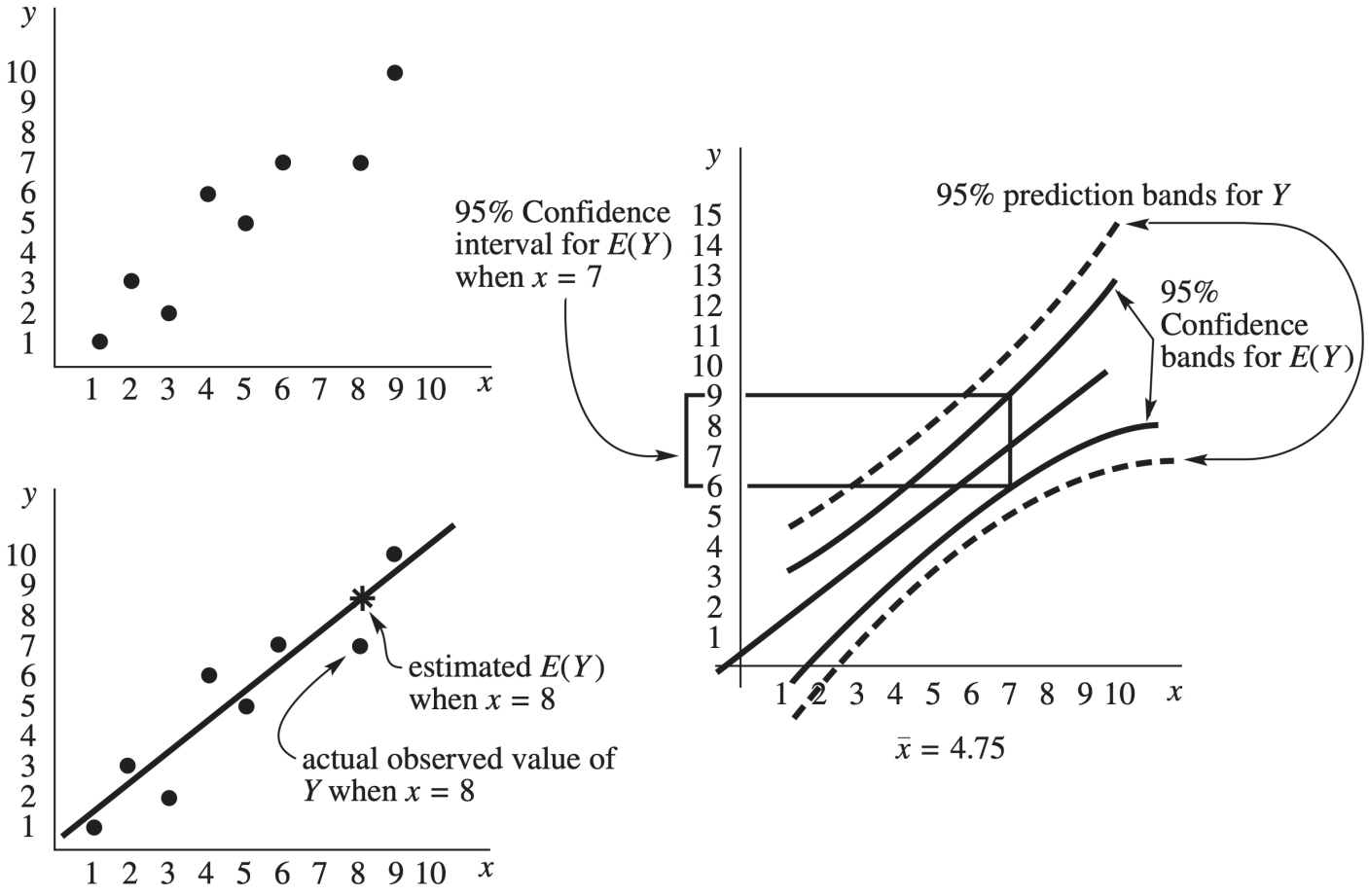
Inference about predicted value of Y

Using the same technique as deriving the t-test for a single parameter or linear functions of parameters, we can show that A $100(1-\alpha)\%$ prediction band for Y^* is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\frac{\alpha}{2}}(n-2)s\sqrt{1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}}$$

Inference about predicted value of Y

The length of the prediction interval for an actual value of Y is longer than the confidence interval for $E(Y)$ when both are determined at the same x^*.



Extending simple regression to multiple regression

To extend simple linear regression to multiple regression models, we need matrix representation of linear model.

A linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

can be written in the matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Fitting linear model by using matrices

Here, we briefly state how linear model is fit by using matrices. These results simply extends properties of simple linear regression to multiple regressions. Using matrix representation, the sum square of error (SSE) is

$$(\boldsymbol{Y} - \boldsymbol{X}\beta)^{\boldsymbol{T}}(\boldsymbol{Y} - \boldsymbol{X}\beta)$$

Taking derivative of SSE with respect to β and set it to zero, we obtain what is often referred to as the normal equation

$$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\beta = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}$$

Solving the normal equation, the solution is

$$\hat{\beta} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}$$

where $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-}$ is generalized inverse of $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$

Fitting linear model by using matrices

Example: We fit a simple linear regression $y = \beta_0 + \beta_1 x$. We observed a sequence of y as $0, 0, 1, 1, 3$ and x as $-2, -1, 0, 1, 2$.

Using algebraic results for simple linear regression, we estimate the regression parameters as

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{5}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{5}(x_i - \overline{x})^2} = 0.7$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 1$$

Next, we use matrix representation of the linear regression. We can see that the matrix representation of the regression yield the same estimates.

Fitting linear model by using matrices

In matrix representation, the data are

$$\boldsymbol{Y} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$$

It follows that

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}, \ \boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}, \ (\boldsymbol{X}^T\boldsymbol{X})^- = \begin{bmatrix} 1/5 & 0 \\ 0 & 1/10 \end{bmatrix}$$

Thus,

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} 1/5 & 0 \\ 0 & 1/10 \end{bmatrix} \begin{bmatrix} 5 \\ 7 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.7 \end{bmatrix}$$

Properties of least square estimators

In a multiple regression $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$, least square estimator has the same properties as the simple linear regression, just expressed in matrix form.

- Parameter estimates are unbiased $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$;
- $Var(\beta_i) = c_{ii}\sigma^2$, where c_{ii} is the element in row i and column i of the matrix $(\boldsymbol{X}^T\boldsymbol{X})^-$;
- $Cov(\beta_i, \beta_j) = c_{ij}\sigma^2$ where c_{ij} is the element in row i and column j of the matrix $\boldsymbol{X}^T\boldsymbol{X}^-$;
- An unbiased estimator of σ^2 is $s^2 = SSE/(n-k-1)$, where $SSE = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^T)^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y}$;
- Each β_i is normally distributed;
- $(n-k-1)s^2/\sigma^2$ has a $\chi^2(n-k-1)$ distribution;
- All β_i and s^2 are independent.

Inferences in multiple linear regression

In simple linear regression, we use t-distribution to construct confidence interval for parameters or linear functions of parameters. For example, $100(1-\alpha)\%$ confidence interval for $\theta = a_0\beta_0 + a_1\beta_1$

$$\hat{\theta} \pm t_{\frac{\alpha}{2}}(n-2)s\sqrt{\frac{a_0^2\frac{\sum x_i^2}{n} + a_1^2 - 2a_0a_1\overline{x}}{S_{xx}}}$$

In multiple regression, we derive the confidence intervals in the same way, just in matrix representation:

$$\boldsymbol{a}^T\boldsymbol{\beta} \pm t_{\frac{\alpha}{2}}(n-k-1)s\sqrt{\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{a}}$$

Inferences in multiple linear regression

In a simple linear regression, a $100(1-\alpha)\%$ prediction interval is constructed from the t-distribution. At x^*, the prediction interval is

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2}}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$

In multiple regression, prediction interval is derived the same way. The prediction interval expressed in matrix form is

$$\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{\beta} \pm t_{\frac{\alpha}{2}}(n-k-1)s\sqrt{1 + \boldsymbol{a}^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-}\boldsymbol{a}}$$

where $\boldsymbol{a}^{\boldsymbol{T}} = \left[1, x_1^*, x_2^*, \ldots, x_k^*\right]$

Inferences in multiple linear regression

Hypotheses about the value of a parameter or a linear function of parameters can be written generally as

$$\boldsymbol{\Lambda\beta} = \boldsymbol{d}$$

Example: In a linear regression $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, the hypothesis $H_0 : \beta_1 = 0$ can be written in matrix form where

$$\boldsymbol{\Lambda} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \ d = 0$$

$H_0 : \beta_0 = 0$ and $\beta_1 = \beta_2$ can be written in matrix form where

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \ \boldsymbol{d} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Inferences in multiple linear regression

In a simple linear regression, hypotheses concerning a single parameter or linear function of parameters is tested using t-statistic, or equivalently, F-statistics constructed from sum of squares.

$$\frac{\frac{(\hat{\beta}_i-\beta_{i0})^2}{c_{ii}\sigma^2}/1}{\frac{(n-2)s^2}{\sigma^2}/(n-2)} = \frac{\frac{1}{\sigma^2}SSH/1}{\frac{1}{\sigma^2}SSE/(n-2)} \sim F_{1,n-2}$$

The same procedure can be extended to multiple linear regression:

$$SSH = (\Lambda\hat{\beta} - \boldsymbol{d})^T(\Lambda(\boldsymbol{X}^T\boldsymbol{X})^-\Lambda^T)^-(\Lambda\hat{\beta} - \boldsymbol{d})$$

$$SSE = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}$$

$$\frac{\frac{1}{\sigma^2}SSH/m}{\frac{1}{\sigma^2}SSE/(n-k-1)} \sim F_{m,n-k-1}$$

where m is the number of independent hypotheses.

Inferences in multiple linear regression

In general, hypotheses in linear regression models are tested using F statistic. Since the F statistic is constructed from various sum of square, the results of hypotheses testings in linear regressions are usually presented in a so-called ANOVA table.

- SSH or SSE are usually labelled sum of squares;
- Sum of squares divided by the corresponding degrees of freedom is mean squares;
- F-statistics is typically constructed from ratios of mean squares.

The analysis of variance

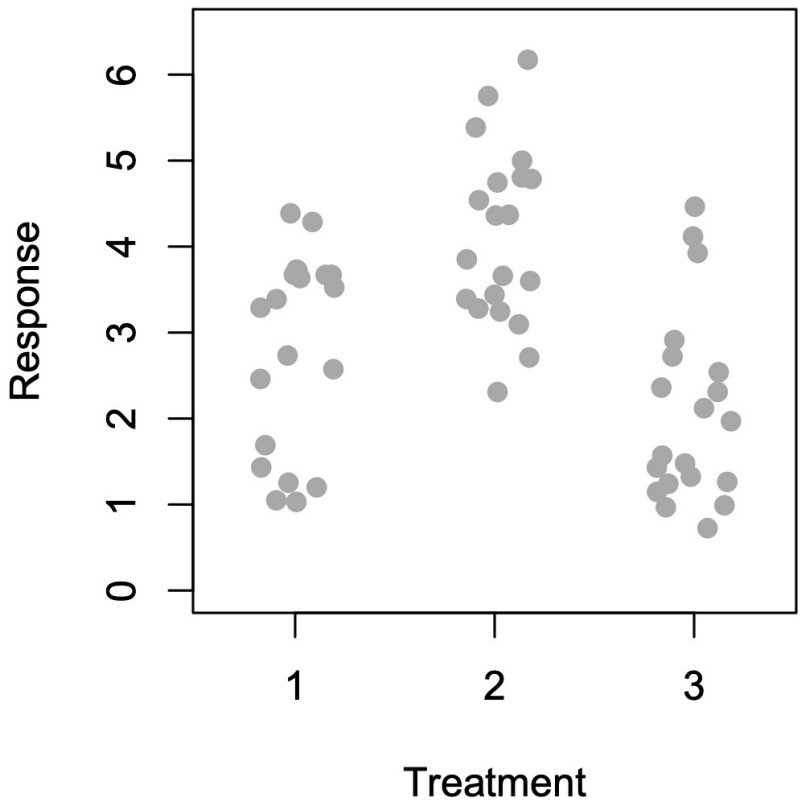
A common scenario we encounter in data analysis is comparing a response variable under several treatments

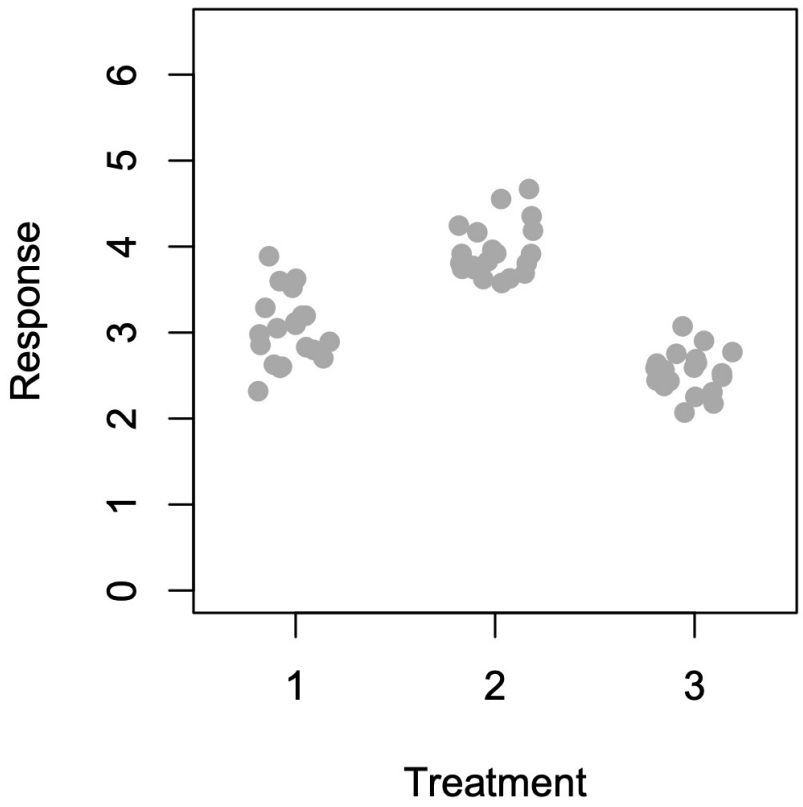
- Comparing plant growth rate under three levels of fertilizations;
- Comparing soil respiration under ambient temperature and warming;
- Evaluating student performance under different teaching methods.

In these examples, we refer to the independent variable that defines the groups as **factors**, and different values of the factor is called its **levels**. To analyze the difference among group means, we typically use a method called **analysis of variance**, or simply **ANOVA**.

The analysis of variance

To examine whether the means in each treatment are all the same or not, an intuitive approach is to compare the variation among treatment means and the variation within a treatment group.





One-way ANOVA

Suppose we have k groups of observations, each sampled from a normal population with means $\mu_1, \mu_2, \ldots, \mu_k$ and a common variance σ^2. Each group has n_k observations. To test H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$, we compare variation among group means and variation within a group.

Variation among group means, often referred to as SST or sum square of treatments, is calculated as

$$SST = \sum_{i=1}^{k} n_i(\overline{Y_{i.}} - \overline{Y})^2$$

Variation within a group, which is referred to as sum square of errors or SSE, is calculated as

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{i.}})^2$$

One-way ANOVA

The hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ is tested by a F-statistic

$$F = \frac{SST/(k-1)}{SSE/(n-k)} \sim F_{k-1,n-k}$$

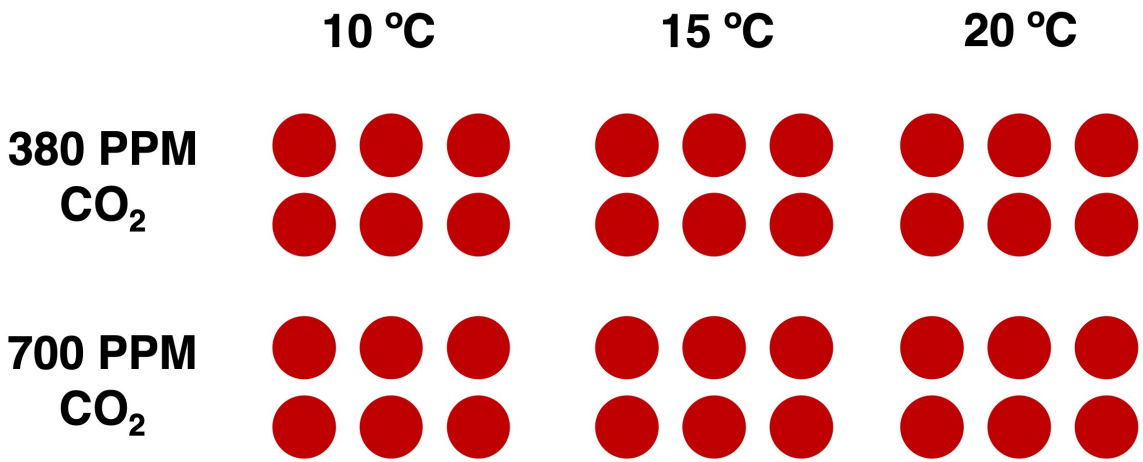
Conventionally, the results of ANOVA are presented in an ANOVA table

Source	df	SS	MS	F
Treatments	$k-1$	SST	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	TSS		

Two-way ANOVA

We can consider the effects of more than one factor. Suppose we have two factors A and B. They have a and b levels respectively. We assume that each observation is $N(\mu_{ij}, \sigma^2)$. This type of data typically allows us to examine

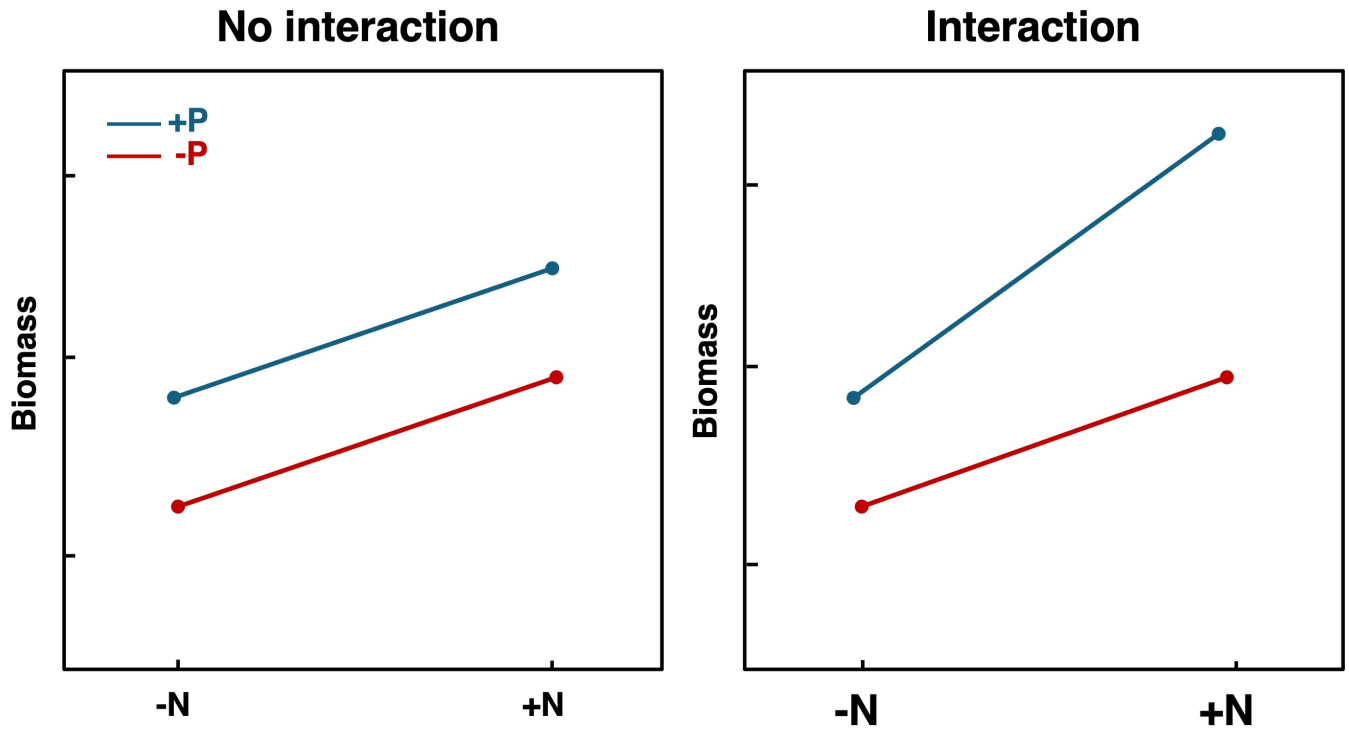
- The main effects of A and B;
- The interaction between A and B.



(A typical experiment set up for two-way ANOVA)

Two-way ANOVA

Two-way ANOVA allows us to identify interactions, which means that the effect of one factor depends on the level of another factor.



(An illustration of interaction in two-way ANOVA.)

Two-way ANOVA

The main effect and interaction can be tested by comparing sum squares of each factor and their interaction with the sum squares of error.

Let Y_{ijk} be the kth replicates at level i of factor A and level j of factor j. Suppose there are a levels of factor A, b levels of factor B, and n_{ij} replicates within each A and B treatment combination.

$$SSA = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\overline{Y_{i\cdot\cdot}} - \overline{Y})^2$$

$$SSB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\overline{Y_{\cdot j\cdot}} - \overline{Y})^2$$

$$SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\overline{Y_{ij\cdot}} - \overline{Y_{i\cdot\cdot}} - \overline{Y_{\cdot j\cdot}} + \overline{Y})^2$$

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \overline{Y_{ij\cdot}})^2$$

Two-way ANOVA

The results of two-way ANOVA can be presented in an ANOVA table.

Source	df	SS	MS	F
A	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$F = \frac{MSA}{MSE}$
B	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$F = \frac{MSB}{MSE}$
AB	$(a-1)(b-1)$	SSAB	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F = \frac{MSAB}{MSE}$
Error	$n-ab$	SSE	$MSE = \frac{SSE}{n-ab}$	
Total	$n-1$	TSS		

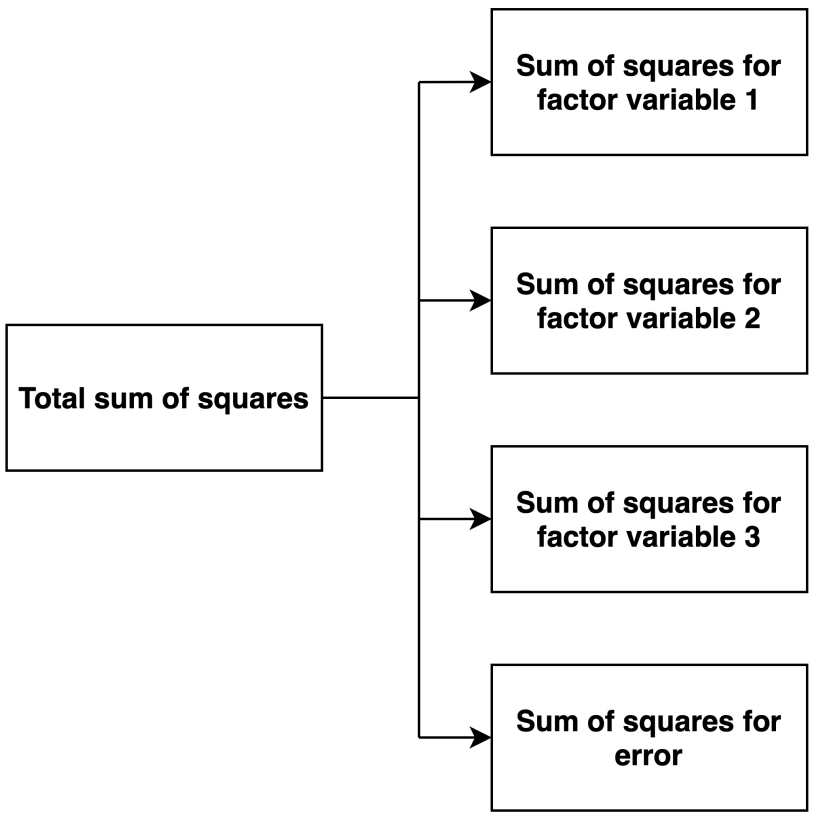
Two-way ANOVA

Example: An ecologists grew three varieties of oats under four nitrogen fertilization levels. She measured the crop yield and used a two-way ANOVA to analyze the effects of oats variety and nitrogen fertilization level. What conclusions would you draw based on the results below?

Source	Df	SS	MS	F	P-value
Variety	2	1786.4	893.2	1.7949	0.175
Nitrogen	3	20020.5	6673.5	13.41	8.367×10^{-7}
Variety:Nitrogen	6	321.7	53.6	0.1078	0.9952
Error	60	29857.3	497.6		

The analysis of variance

The intuitive approach we have introduced so far can be extended to cases with more than two factors. In essence, ANOVA decomposes the total sum of squares into sum of squares for each factor.



ANOVA as a linear model

ANOVA uses decomposition of sum of squares to compare means in groups defined by factors. Similarly, we have seen that we use the F-statistics constructed from sum of squares for hypothesis testing in linear models. This similarity is not coincidence. In fact, ANOVA is a type of linear model.

Recall a linear model is defined as a model where the response variable is a linear function of parameters, i.e.,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

ANOVA can be written as a linear model by defining x as a dummy or indicator variable.

ANOVA as a linear model

A one-way ANOVA with k groups can be written as

$$y_{ij} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Here, x_i is an indicator variable where $x_i = 1$ if the observation is in group i and $x_i = 0$ if the observation is not in group i.

A few comments on the model parameters:

- The model can be written more concisely as $y_{ij} = \beta_i, \ i = 1, 2, \ldots, k$;
- There are $k + 1$ parameters in the model;
- The model can also be written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$.

ANOVA as a linear model

Example: We compare the biomass of plants grown under two nitrogen levels and observed the following data:

N level	Biomass
Low	12, 11, 10
High	20, 22, 21

Here, we can write the model as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\boldsymbol{y} = \begin{bmatrix} 12 \\ 11 \\ 10 \\ 20 \\ 22 \\ 21 \end{bmatrix}; \ \boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}; \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}; \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix}.$$

ANOVA as a linear model

There are multiple ways to write the same linear model. The model in the above example can also be written as as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\boldsymbol{y} = \begin{bmatrix} 12 \\ 11 \\ 10 \\ 20 \\ 22 \\ 21 \end{bmatrix} ; \ \boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} ; \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} ; \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix}.$$

Here, β_1 is the mean of group 1 and β_2 is the difference in group mean between group 2 and group 1. This type of coding is referred to as the reference level coding and is used in most statistical software.

ANOVA as a linear model

A two-way ANOVA model with interaction can be written as

$$y_{ijk} = \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$
$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

where α_i is the effect of level i of factor A, β_j is the effect of level j of factor B, and γ_{ij} is the interaction effect.

Comments on the two-way ANOVA model

- A two-way ANOVA do not necessarily has all terms as in the model above. For example, if there are reasons to suggest that there are no interactions, you do not need the interaction term.
- Hierarchical principle: if you include a interaction term, you should usually include the main effects involved in that interaction.

ANOVA as a linear model

Since ANOVA is a linear model, the procedures for parameter estimation and hypotheses testing we derived for linear models can be readily applied here.

- Parameters can be estimated as $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Y}$;
- Hypothesis can be tested by F-tests. The results are the same as the ANOVA table we presented in previous slides.

$$SSH = (\Lambda\hat{\beta} - \boldsymbol{d})^T(\Lambda(\boldsymbol{X}^T\boldsymbol{X})^-\Lambda^T)^-(\Lambda\hat{\beta} - \boldsymbol{d})$$

$$SSE = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}$$

$$\frac{\frac{1}{\sigma^2}SSH/m}{\frac{1}{\sigma^2}SSE/(n-k-1)} \sim F_{m,n-k-1}$$

Hypothesis testing in ANOVA

In a one-way ANOVA, the factor has 3 levels and 3 replicates in each group.

Using reference level coding, how do you write the model in matrix form?

$$\boldsymbol{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{bmatrix}; \; \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}; \; \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

Hypothesis testing in ANOVA

How do you test the hypothesis that the factor has no effects?

If the factor has no effects, the group means are equal. Therefore,

$$H_0 : \beta_2 = 0 \textit{ and } \beta_3 = 0$$

$$H_A : \beta_2 \neq 0 \textit{ or } \beta_3 \neq 0.$$

In matrix form, this hypothesis is written as $\boldsymbol{\Lambda\beta} = \mathbf{0}$, where

$$\Lambda = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Summary

Although comprehending the theory of linear model is not feasible in a single lecture, understanding the following principles helps us use it correctly in practice:

- Least squares works the same way for linear regression and ANOVA, providing a unifying framework to understand these methods;
- Matrix representation of linear model helps understand the meaning of parameters, which is critical for subsequent inference;
- Least squares coupled with normally distributed iid error is the theoretical foundation for statistical inference;
- In linear models, statistical inference are typically done with F-test, which can be presented in the form a an ANOVA table.