# Lecture 20
# Linear Regression and the Methods of Least Squares

**Chao Song**

College of Ecology
Lanzhou University

December 2, 2024

# Model and statistical inference

In all previous discussions on statistical inference, we assumed that the observable random variables $Y_1, Y_2, \ldots, Y_n$ were independent and identically distributed. One implication of this assumption is that the expected value of $E(Y_i)$ is constant.

In many inferential problems, such an assumption is unrealistic.

- Stopping distance of automobile will depend on the speed of the vehicle;
- Potency of an antibiotic depends on the time it has been stored;
- Elongation observed in a metal alloy depends on the force applied and the temperature.

In these cases, we are usually interested in undertaking a inferential procedure that can be used when a random variable $Y$, called the **dependent variable**, has a mean that is a function of one or more nonrandom variables $X_1, X_2, \ldots, X_k$, called the **independent variables**.

# Deterministic and probabilistic model

A deterministic model is when the dependent variable can be predicted from the independent variables without any uncertainty.

**Example**: Fick's law of diffusion postulates that the flux rate of a gas across concentration gradient is described as

$$J = D\phi$$

where $J$ is the flux rate, $D$ is diffusion coefficient, and $\phi$ is concentration gradient.

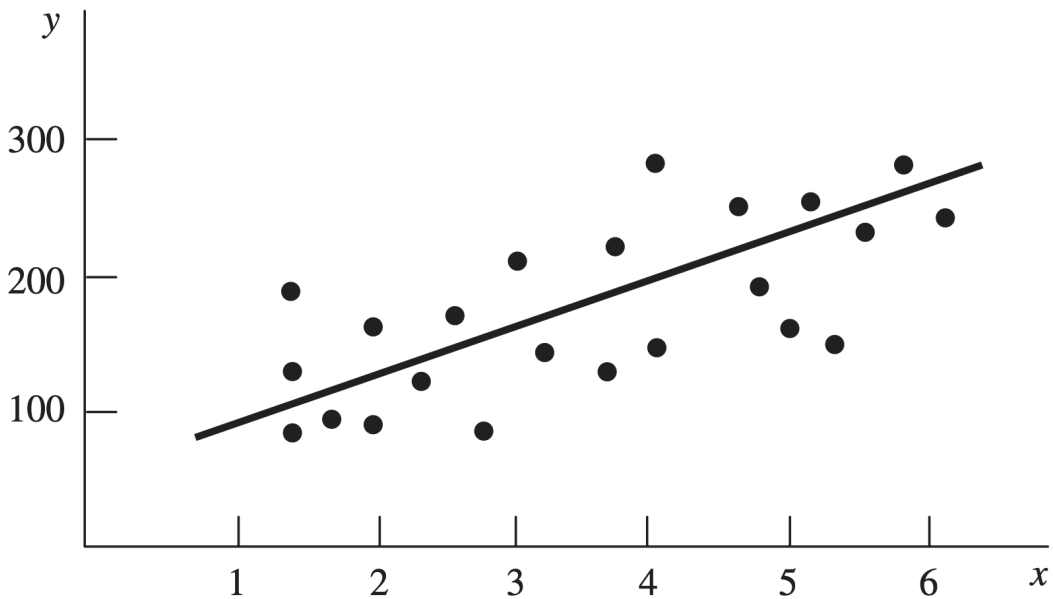
This is a deterministic model because it does not allow any uncertainty/error in predicting $Y$. This model implies that $Y$ always takes the same value $D\phi_0$ whenever $\phi = \phi_0$.

# 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

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

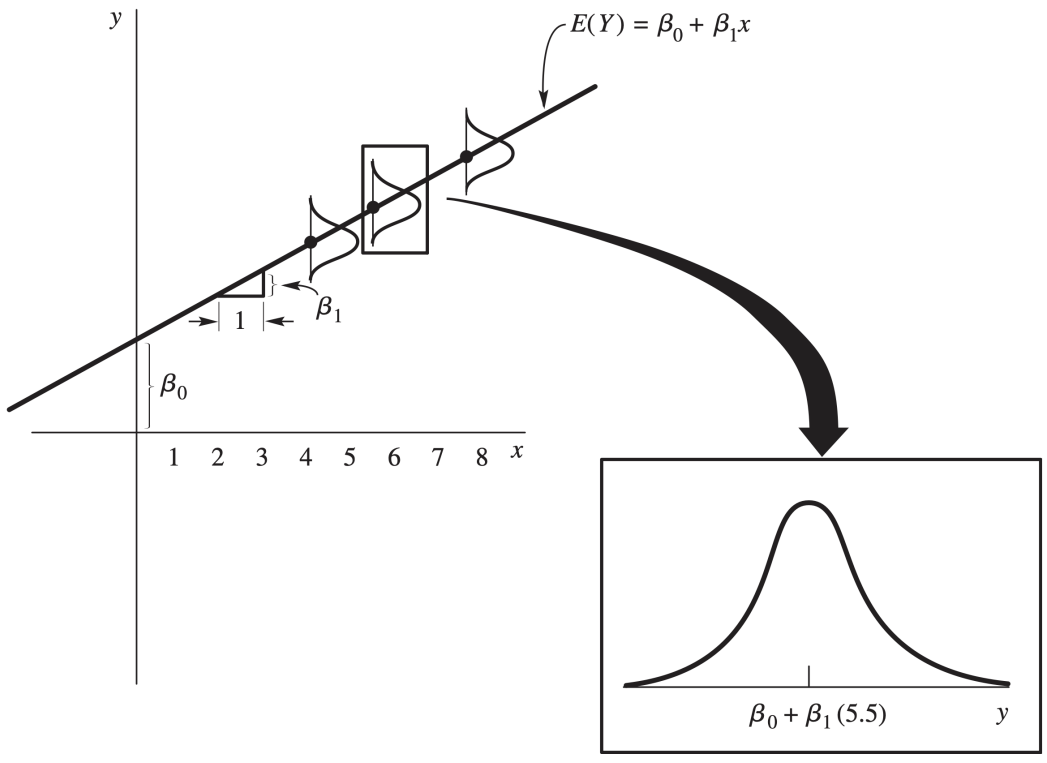
Each observation deviates from the mean by an unknown random error

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

where $\varepsilon$ 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 $\varepsilon$.



# 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

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

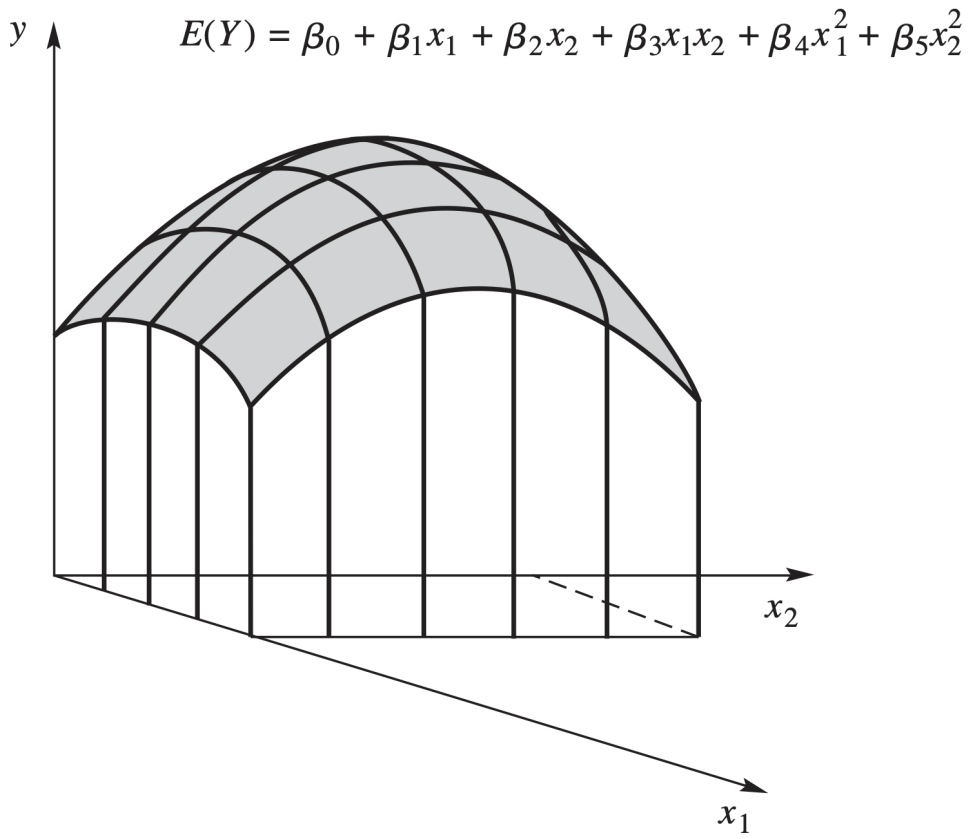
where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters, $\varepsilon$ is a random variable and the variables $X_1, X_2, \ldots, X_k$ assume known value. We will assume $E(\varepsilon) = 0$ and hence that

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

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

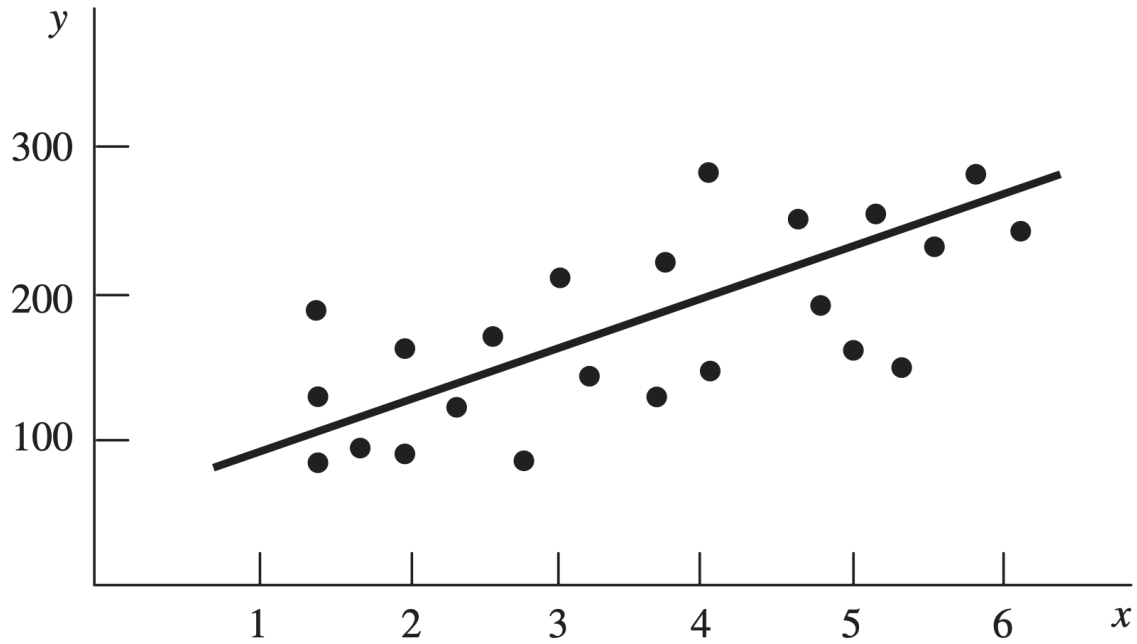
If the model is of the form $Y = \beta_0 + \beta_1 X$, where $X$ is a continuous variable, the model is a simple linear regression. If the model contains multiple continuous independent variables, the model is called multiple linear regression. Below is an example of multiple linear regression:



# The method of least square

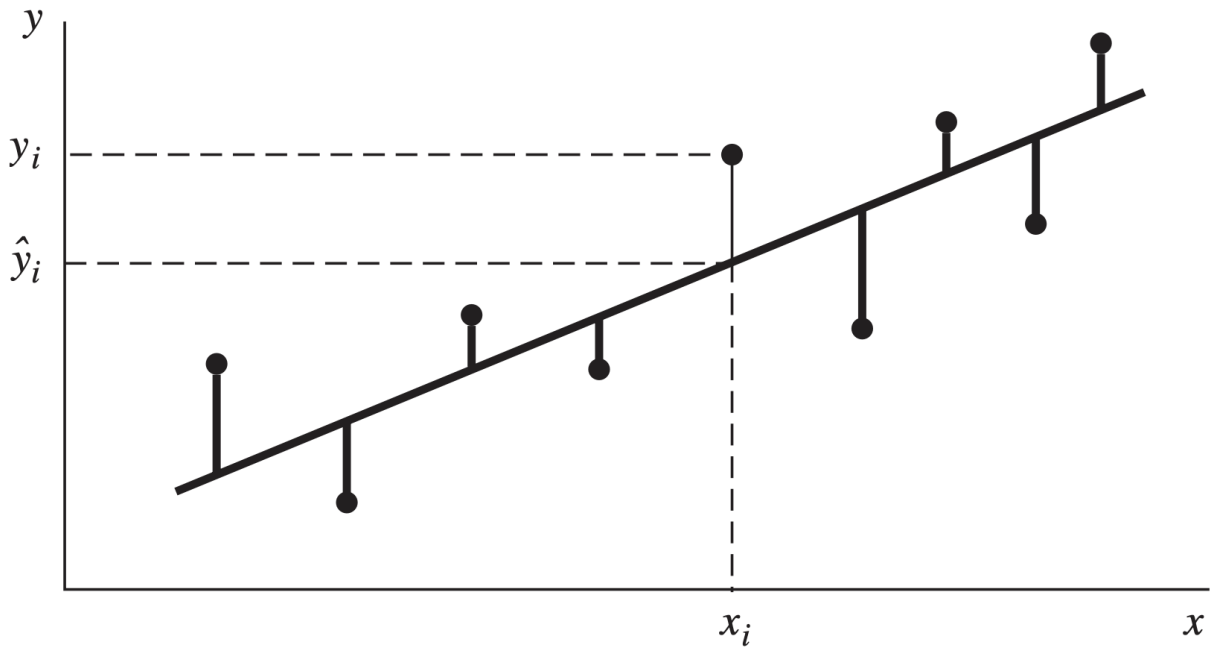
**How do we estimate parameter in a linear model?**

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

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

To obtain parameter estimates that minimize SSE, take partial derivatives and set them to zero.

$$\frac{\partial SSE}{\partial \beta_0} = \frac{\partial \sum_{i=1}^{n} \left[ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2}{\partial \beta_0} = -\sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

$$= -2\Big( \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \Big) = 0$$

# Method of least squares

$$\frac{\partial SSE}{\partial \beta_1} = \frac{\partial \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2}{\partial \beta_1} = -\sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]x_i$$

$$= -2\Big(\sum_{i=1}^{n} x_i y_i - n\hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2\Big) = 0$$

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

The least square estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased.

**Proof**: Recall the least square estimator for $\beta_1$ is

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i - \sum_{i=1}^{n}(x_i - \overline{x})\overline{y}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}
\end{aligned}$$

# Properties of least square estimator

$$
\begin{aligned}
E(\hat{\beta}_1) &= E\left[\frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right] \\
&= E\left[\frac{\sum_{i=1}^{n}(x_i-\overline{x})(\beta_0+\beta_1 x_i)}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right] \\
&= \beta_0\frac{\sum_{i=1}^{n}(x_i-\overline{x})}{S_{xx}} + \beta_1\frac{\sum_{i=1}^{n}(x_i-\overline{x})x_i}{S_{xx}} \\
&= \beta_1\frac{\sum_{i=1}^{n}(x_i-\overline{x})x_i - \sum_{i=1}^{n}(x_i-\overline{x})\overline{x}}{S_{xx}} \\
&= \beta_1\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{S_{xx}} \\
&= \beta_1,
\end{aligned}
$$

# Properties of least square estimator

We now find the expected value of $\hat{\beta}_0$. Recall the least square estimator $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, therefore

$$\begin{aligned} E(\hat{\beta}_0) &= E(\overline{y} - \hat{\beta}_1 \overline{x}) \\ &= E(\overline{y}) - E(\hat{\beta}_1)\overline{x} \\ &= \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} \\ &= \beta_0 \end{aligned}$$

We thus have proven that the least square estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ are **unbiased** estimator of $\beta_0$ and $\beta_1$ respectively.

# Properties of least square estimators

What are the variances of the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$?

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var\left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \\
&= \left(\frac{1}{S_{xx}}\right)^2 \sum_{i=1}^{n} Var\left[(x_i - \overline{x})y_i\right] \\
&= \left(\frac{1}{S_{xx}}\right)^2 \sum_{i=1}^{n} \left[(x_i - \overline{x})^2 Var(y_i)\right] \\
&= \left(\frac{1}{S_{xx}}\right)^2 S_{xx}\sigma^2 \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}
$$

# Properties of least square estimators

The variance of $\hat{\beta}_0$ is

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\overline{y} - \hat{\beta}_1 \overline{x}) \\ &= Var(\overline{y}) + \overline{x}^2 Var(\hat{\beta}_1) - 2Cov(\overline{y}, \hat{\beta}_1 \overline{x}) \\ &= Var(\overline{y}) + \overline{x}^2 Var(\hat{\beta}_1) - 2\overline{x} Cov(\overline{y}, \hat{\beta}_1) \end{aligned}$$

We thus must find $Var(\overline{y})$ and $Cov(\overline{y}, \hat{\beta}_1)$. Here,

$$Var(\overline{y}) = Var(\frac{1}{n}\sum_{i=1}^{n} y_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(y_i) = \frac{\sigma^2}{n}$$

$$Cov(\overline{y}, \hat{\beta}_1) = Cov\left[\sum_{i=1}^{n}\frac{1}{n}y_i, \sum_{i=1}^{n}\frac{x_i - \overline{x}}{S_{xx}}y_i\right] = \sum_{i=1}^{n}\frac{x_i - \overline{x}}{nS_{xx}} Var(y_i) = 0.$$

Therefore, we have

$$Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \overline{x}^2\frac{\sigma^2}{S_{xx}} = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}$$

# Properties of least square estimators

Note that the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_0$ are not independent.

$$
\begin{aligned}
Cov(\hat{\beta}_1, \hat{\beta}_0) &= Cov(\hat{\beta}_1, \overline{y} - \hat{\beta}_1 \overline{x}) \\
&= Cov(\hat{\beta}_1, \overline{y}) - \overline{x} Var(\hat{\beta}_1) \\
&= Cov(\sum_{i=1}^{n} \frac{x_i - \overline{x}}{S_{xx}} y_i, \sum_{i=1}^{n} \frac{1}{n} y_i) - \overline{x} Var(\hat{\beta}_1) \\
&= \sum_{i=1}^{n} \frac{x_i - \overline{x}}{n S_{xx}} Var(y_i) - \overline{x} Var(\hat{\beta}_1) \\
&= 0 - \overline{x} \frac{\sigma^2}{S_{xx}} \\
&= -\frac{\overline{x} \sigma^2}{S_{xx}}
\end{aligned}
$$

Thus, $\hat{\beta}_0$ and $\hat{\beta}_0$ are negatively correlated unless $\overline{x} = 0$.

# Method of least squares

There is a remaining parameter $\sigma^2$, we typically estimate it by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{SSE}{n-2}$$

This is an unbiased estimator because

$$\begin{aligned} E(SSE) &= E\Big[\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\Big] \\ &= E\Big[\sum_{i=1}^{n}(y_i - \overline{y} + \hat{\beta}_1\overline{x} - \hat{\beta}_1 x_i)^2\Big] \\ &= E\Big[\sum(y_i - \overline{y})^2 + \hat{\beta}_1^2\sum(x_i - \overline{x})^2 - 2\hat{\beta}_1\sum(x_i - \overline{x})(y_i - \overline{y})\Big] \end{aligned}$$

# Method of least squares

Note here that

$$\sum(x_i - \overline{x})(y_i - \overline{y}) = \sum(x_i - \overline{x})^2 \hat{\beta}_1$$
$$\sum(y_i - \overline{y})^2 = \sum y_i^2 - n\overline{y}^2$$

Plug these two equations in $E(SSE)$, we have

$$E(SSE) = E\left[\sum y_i^2 - n\overline{y}^2 + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1^2 S_{xx}\right]$$
$$= \sum E(y_i^2) - nE(\overline{y}^2) - S_{xx}E(\hat{\beta}_1^2)$$

Here, each component can be calculated as

$$E(y_i^2) = Var(y_i) + [E(y_i)]^2 = \sigma^2 + (\beta_0 + \beta_1 x_i)^2$$
$$E(\overline{y}^2) = Var(\overline{y}) + [E(\overline{y})]^2 = \frac{\sigma^2}{n} + (\beta_0 + \beta_1 \overline{x})^2$$
$$E(\hat{\beta}_1^2) = Var(\hat{\beta}_1) + [E(\hat{\beta}_1)]^2 = \frac{\sigma^2}{S_{xx}} + \beta_1^2$$

# Method of least squares

We therefore have

$$\begin{aligned}
E(SSE) &= n\sigma^2 + \sum(\beta_0 + \beta_1 x_i)^2 - \sigma^2 - n(\beta_0 + \beta_1 \overline{x})^2 - \sigma^2 - S_{xx}\beta_1^2 \\
&= (n-2)\sigma^2 + \sum y_i^2 - n\overline{y}^2 - \sum(x_i - \overline{x})^2\beta_1^2 \\
&= (n-2)\sigma^2 + \sum(y_i - \overline{y})^2 - \sum(x_i - \overline{x})^2\beta_1^2 \\
&= (n-2)\sigma^2 + \sum(\beta_1 x_i - \beta_1 \overline{x})^2 - \sum(x_i - \overline{x})^2\beta_1^2 \\
&= (n-2)\sigma^2
\end{aligned}$$

Therefore, $SSE/(n-2)$ provides a unbiased estimator for $\sigma^2$.

# Properties of least square estimators

We have derived 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 $\sigma^2$.

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 $\varepsilon_i$.

# Properties of least square estimators

If we further assume that $\varepsilon_i \sim N(0, \sigma^2)$, then $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Note that both $\hat{\beta}_0$. and $\hat{\beta}_1$ are linear combinations of $y_i$:

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

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

Recall that any linear combination of normal distributed random variables still have a normal distribution. Therefore, if $\varepsilon$ is normal, the least square estimators are both normally distributed.

# Properties of least estimators

Without explicitly proving them, we state the following properties of least square estimators in a simple linear regression.

If the error $\varepsilon$ has a normal distribution, we have

- 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 $\beta_i$

In a linear regression, if the random error $\varepsilon$ is normally distributed, we have established that $\hat{\beta}_i$ is an unbiased, normally distributed estimator of $\beta_i$ 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 $\beta_i$

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 $\beta_i$ as:

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

# Inferences concerning the parameters $\beta_i$

Recall the definition of t-distribution:

$$T = \frac{Z}{\sqrt{V/k}} \sim t(k)$$

where $Z \sim N(0,1)$ and $V \sim \chi^2(k)$ and are independent. F-distribution is defined as

$$F = \frac{V_1/d_1}{V_2/d_2} \sim F_{d_1,d_2}$$

where $V_1$ and $V_2$ are independent $\chi^2$ variables with degrees of freedom $d_1$ and $d_2$, respectively.

We thus can see that $T^2$ has a F-distribution with df 1 and $k$. A hypothesis test based on t-distribution can be equivalently done with a corresponding F-distribution.

# Inferences concerning the parameters $\beta_i$

The t-test for each $\beta_i$ 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 $\beta_i$, we frequently are interested in linear functions of model parameters. For example, we may wish to make inference about

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

We use $\hat{\theta} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1$ as an unbiased estimator of $\theta$ because

$$E(\hat{\theta}) = E(a_0\hat{\beta}_0 + a_1\hat{\beta}_1) = a_0E(\hat{\beta}_0) + a_1E(\hat{\beta}_1) = a_0\beta_0 + a_1\beta_1 = \theta$$

Because $\hat{\beta}_i$ are all normally distributed, $\hat{\theta}$ as a linear function of $\hat{\beta}_i$ also has a normal distribution. Its variance is

$$\begin{aligned} Var(\hat{\theta}) &= Var(a_0\hat{\beta}_0 + a_1\hat{\beta}_1) \\ &= a_0^2 Var(\hat{\beta}_0) + a_1^2 Var(\hat{\beta}_1) + 2a_0a_1 Cov(\hat{\beta}_0, \hat{\beta}_1) \\ &= \frac{a_0^2\frac{\sum x_i^2}{n} + a_1^2 - 2a_0a_1\overline{x}}{S_{xx}}\sigma^2 = c_\theta\sigma^2 \end{aligned}$$

# Inferences concerning linear functions of parameters

We have shown that $\hat{\theta} \sim N(\theta, c_\theta \sigma^2)$. Thus

$$\frac{\hat{\theta} - \theta_0}{\sqrt{c_\theta}\sigma} \sim N(0, 1)$$

In linear regression, we also have

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

Using the results above, we derive 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 $\theta$ 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:

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x^*) - E(Y^*)}{\sqrt{c_{E(Y)}}\sigma} \sim t_{\alpha/2}(n-2)$$

where

$$c_{E(Y)} = \frac{\frac{\sum x_i^2}{n} + x^{*2} - 2x^*\overline{x}}{S_{xx}}$$

# Inference about predicted mean

Note that $c_{E(Y)}$ can be further simplified as

$$
\begin{aligned}
c_{E(Y)} &= \frac{\frac{\sum x_i^2}{n} + x^{*2} - 2x^*\overline{x}}{S_{xx}} = \frac{\frac{\sum x_i^2}{n} + x^{*2} - 2x^*\overline{x} + \overline{x}^2 - \overline{x}^2}{S_{xx}} \\
&= \frac{\frac{\sum x_i^2 - n\overline{x}^2}{n} + x^{*2} - 2x^*\overline{x} + \overline{x}^2}{S_{xx}} = \frac{\frac{S_{xx}}{n} + (x^* - \overline{x})^2}{S_{xx}} \\
&= \frac{1}{n} + \frac{(x^* - \overline{x})^2}{X_{xx}}
\end{aligned}
$$

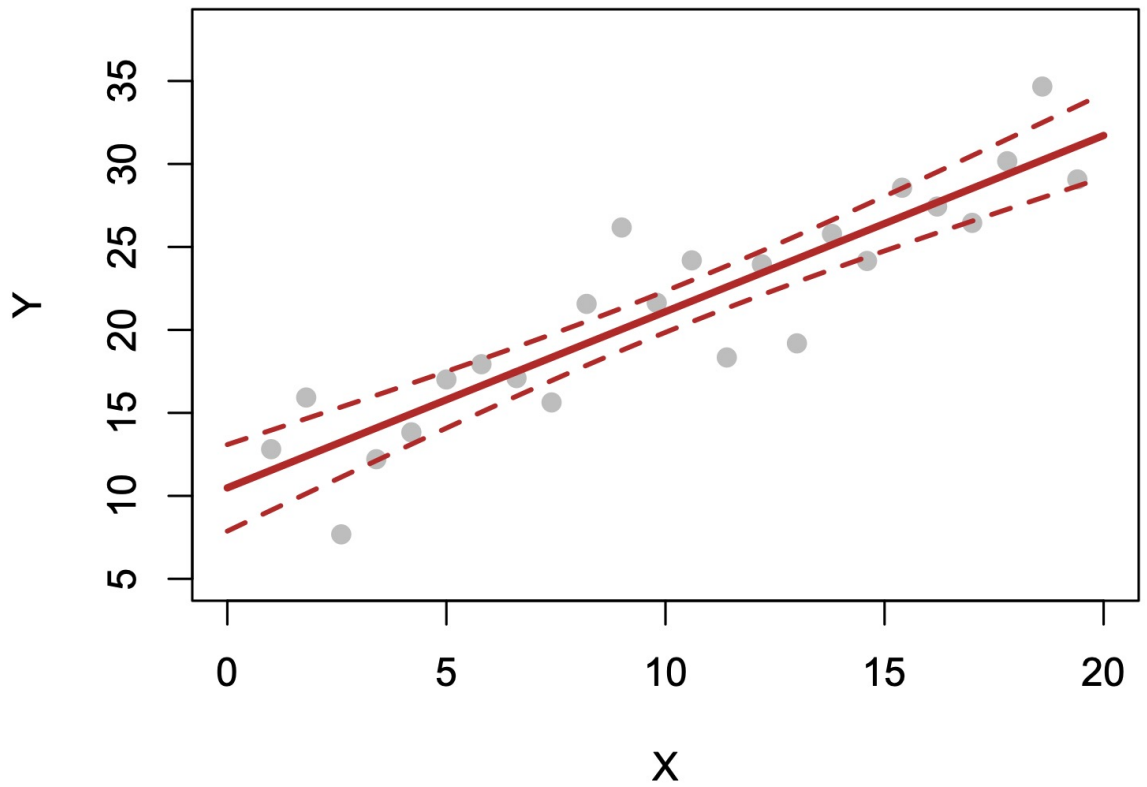
A $100(1-\alpha)\%$Confidence interval for $E(Y^*) = \beta_0 + \beta_1 x^*$ is

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

# 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 $Y$

Let $\varepsilon^*$ be the error of the prediction, i.e., $\varepsilon^* = Y^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$. Because $Y^*$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are all normally distributed, $\varepsilon^*$ is also normally distributed. Here,

$$\begin{aligned}
E(\varepsilon^*) &= E\left[Y^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*)\right] \\
&= E(Y^*) - E(\hat{\beta}_0) - E(\hat{\beta}_1)x^* \\
&= \beta_0 + \beta_2 x^* - \beta_0 - \beta_1 x^* = 0 \\
Var(\varepsilon^*) &= Var\left[Y^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*)\right] \\
&= Var(Y^*) + Var(\hat{\beta}_0 + \hat{\beta}_1 x^*) \\
&= \sigma^2 + \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)\sigma^2 \\
&= \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)\sigma^2
\end{aligned}$$

Note that $Y^*$ is a future prediction that is not employed in estimating $\hat{\beta}_0$ and $\hat{\beta}_1$. Thus, $Y^*$ is independent of $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

# Inference about predicted value of $Y$

We thus has shown that

$$Y^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*) \sim N\left[0, \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)\sigma^2\right]$$

Using the same technique as deriving the t-test for a single parameter or linear functions of parameters, we have

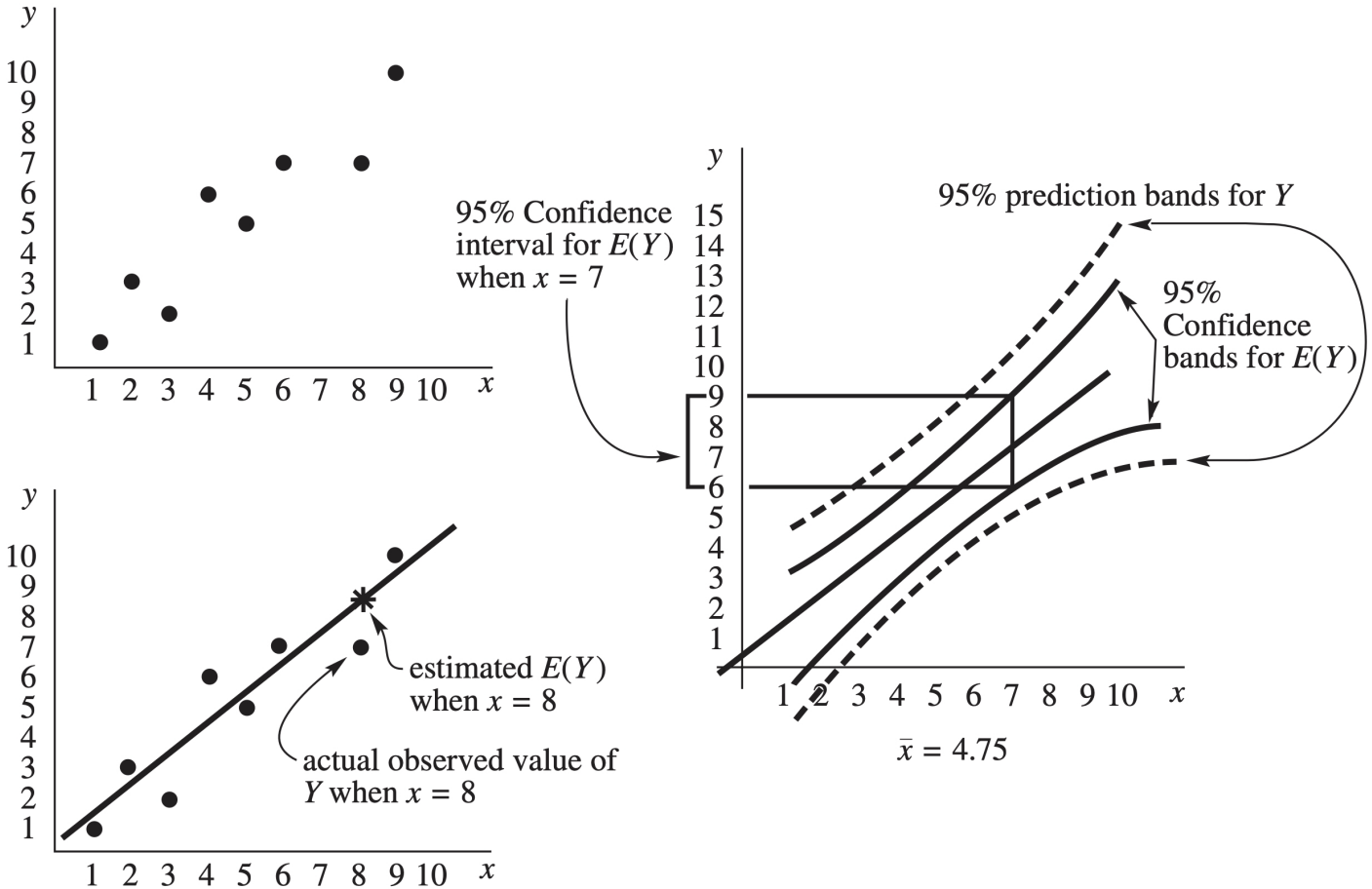
$$\frac{Y^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}} \sim t(n-2)$$

A $100(1-\alpha)\%$ prediction band for $Y^*$ is thus

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

# 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 $\beta$ 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 $\sigma^2$ 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 $\beta_i$ is normally distributed;
- $(n-k-1)s^2/\sigma^2$ has a $\chi^2(n-k-1)$ distribution;
- All $\beta_i$ 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

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

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

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

# 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.