11 Linear regression Models

Complementary reading: Chapter 11 and Appendix A (WMS)

11.1 Introduction

IMPORTANCE: A problem that often arises in economics, engineering, medicine, and other areas is that of investigating the mathematical relationship between two (or more) variables. In such settings, the goal is often to model a continuous random variable Y as a function of one or more **independent variables**, say, $x_1, x_2, ..., x_k$. Mathematically, we can express this model as

$$Y = g(x_1, x_2, ..., x_k) + \epsilon,$$

where $g : \mathcal{R}^k \to \mathcal{R}$, and where the random variable ϵ satisfies certain conditions. This is called a **regression model**.

- The presence of the random error term ϵ conveys the fact that the relationship between the dependent variable Y and the independent variables through $g(x_1, x_2, ..., x_k)$ is not perfect.
- The independent variables $x_1, x_2, ..., x_k$ are assumed to be **fixed** (not random), and they are measured without error. If $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$, then

$$\begin{aligned} E(Y|x_1, x_2, ..., x_k) &= g(x_1, x_2, ..., x_k) \\ V(Y|x_1, x_2, ..., x_k) &= \sigma^2. \end{aligned}$$

LINEAR MODELS: In this course, we will consider models of the form

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}_{g(x_1, x_2, ..., x_k)} + \epsilon,$$

that is, g is a linear function of the **regression parameters** β_0, β_1, ..., β_k. We call this a **linear regression model**.

REMARK: In some problems, a **nonlinear regression model** may be appropriate. For example, suppose that Y measures plant growth (in cm, say) and x denotes time. We would expect the relationship to eventually "level off" as x gets large, as plants can not continue to grow forever. A popular model for this situation is the nonlinear model

$$Y = \underbrace{\frac{\beta_0}{1+\beta_1 e^{\beta_2 x}}}_{g(x)} + \epsilon.$$

Note that, if $\beta_2 < 0$, then

$$\lim_{x\to\infty} g(x) = \lim_{x\to\infty} \left(\frac{\beta_0}{1+\beta_1 e^{\beta_2 x}} \right) = \beta_0.$$

Therefore, if $\beta_2 < 0$, this g function has a horizontal asymptote at $y = \beta_0$, a characteristic that is consistent with the data we would likely observe.

DESCRIPTION: We call a regression model a **linear regression model** if the regression parameters enter the g function in a linear fashion. For example, each of the models is a linear regression model:

$$\begin{aligned} Y &= \underbrace{\beta_0 + \beta_1 x}_{g(x)} + \epsilon \\ Y &= \underbrace{\beta_0 + \beta_1 x + \beta_2 x^2}_{g(x)} + \epsilon \\ Y &= \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}_{g(x_1,x_2)} + \epsilon. \end{aligned}$$

These should be contrasted with the nonlinear model above, where the regression parameters β_0, β_1, and β_2 enter the g function nonlinearly. The term "linear" does not refer to the shape of the regression function g. It refers to the manner in which the regression parameters β_0, β_1, ..., β_k enter the g function.

GOALS: We will restrict attention to linear (regression) models. Our goals are to

- obtain estimates of the regression parameters and study the sampling distributions of these estimators
- perform statistical inference for the regression parameters and functions of them
- make predictions about future values of Y based on an estimated model.

11.2 Simple linear regression

TERMINOLOGY: A **simple linear regression model** includes only one independent variable x. The model is of the form

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

The regression function $g(x) = \beta_0 + \beta_1 x$ is a straight line with intercept β_0 and slope β_1. If $E(\epsilon) = 0$, then β_1 quantifies the change in $E(Y)$ brought about by a one-unit change in x.

TERMINOLOGY: When we say, "fit a regression model," we mean that we would like to estimate the regression parameters in the model with the observed data. Suppose that we collect (x_i, Y_i), $i = 1, 2, ..., n$, and postulate the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for each $i = 1, 2, ..., n$. Our first goal is to estimate β_0 and β_1. Formal assumptions for the error terms ϵ_i will be given later.

11.2.1 Least squares estimation

LEAST SQUARES: A widely-accepted method of estimating the model parameters β_0 and β_1 is that of least squares. The **method of least squares** says to choose the values of β_0 and β_1 that minimize

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Denote the least squares estimators by $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively. These are the values of β_0 and β_1 that minimize $Q(\beta_0, \beta_1)$. A two-variable minimization exercise can be used to find expressions for $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Taking partial derivatives of $Q(\beta_0, \beta_1)$, we obtain

$$\begin{aligned} \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) \stackrel{\text{set}}{=} 0 \\ \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i \stackrel{\text{set}}{=} 0. \end{aligned}$$

Solving for β_0 and β_1 gives the **least squares estimators**

$$\begin{aligned} \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{x} \\ \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}. \end{aligned}$$

11.2.2 Properties of the least squares estimators

INTEREST: We wish to investigate the properties of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as estimators of the true regression parameters β_0 and β_1 in the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$. To do this, we need to formally state our assumptions on the error terms ϵ_i. Specifically, we will assume that $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. This means that

- $E(\epsilon_i) = 0$, for $i = 1, 2, ..., n$
- $V(\epsilon_i) = \sigma^2$, for $i = 1, 2, ..., n$, that is, the variance is constant
- the random variables ϵ_i are independent
- the random variables ϵ_i are normally distributed.

OBSERVATION: Under the assumption that $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, it follows that

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

In addition, the random variables Y_i are **independent**. They are not identically distributed because the mean $\beta_0 + \beta_1 x_i$ is different for each x_i.

Fact 1. The least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are **unbiased estimators** of β_0 and β_1, respectively, that is,

$$\begin{aligned} E(\widehat{\beta}_0) &= \beta_0 \\ E(\widehat{\beta}_1) &= \beta_1. \end{aligned}$$

Proof. Algebraically,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

since

$$\begin{aligned}\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y}) &= \sum_{i=1}^n (x_i - \overline{x})Y_i - \sum_{i=1}^n (x_i - \overline{x})\overline{Y} \\ &= \sum_{i=1}^n (x_i - \overline{x})Y_i - \overline{Y}\sum_{i=1}^n (x_i - \overline{x})\end{aligned}$$

and $\sum_{i=1}^n (x_i - \overline{x}) = 0$. Therefore, if we let

$$c_i = \frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

for $i = 1, 2, ..., n$, we see that $\widehat{\beta}_1$ can be written as

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} = \sum_{i=1}^n c_i Y_i,$$

a **linear combination** of $Y_1, Y_2, ..., Y_n$. Taking expectations, we have

$$\begin{aligned}E(\widehat{\beta}_1) = E\left(\sum_{i=1}^n c_i Y_i\right) &= \sum_{i=1}^n c_i E(Y_i) \\ &= \sum_{i=1}^n c_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i.\end{aligned}$$

However, note that

$$\sum_{i=1}^n c_i = \sum_{i=1}^n \left[\frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] = \frac{\sum_{i=1}^n (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = 0$$

and

$$\sum_{i=1}^n c_i x_i = \sum_{i=1}^n \left[\frac{(x_i - \overline{x})x_i}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = 1.$$

Therefore, $E(\widehat{\beta}_1) = \beta_1$ as claimed. To show that $\widehat{\beta}_0$ is unbiased, we first note that

$$E(\widehat{\beta}_0) = E(\overline{Y} - \widehat{\beta}_1 \overline{x}) = E(\overline{Y}) - \overline{x}E(\widehat{\beta}_1).$$

However, $E(\widehat{\beta}_1) = \beta_1$ and

$$\begin{aligned} E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\beta_0 + \frac{1}{n}\sum_{i=1}^{n}\beta_1 x_i \\ &= \beta_0 + \beta_1 \overline{x}. \end{aligned}$$

Therefore,

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

as claimed. We have shown that the least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased.

NOTE: It is important to note that the only assumption we used in the preceding argument was that $E(\epsilon_i) = 0$. Therefore, a sufficient condition for the least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to be unbiased is that $E(\epsilon_i) = 0$. □

Fact 2. The least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have the following characteristics:

$$\begin{aligned} V(\widehat{\beta}_0) &= \sigma^2\left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \\ V(\widehat{\beta}_1) &= \sigma^2\left[\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \\ \text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) &= \sigma^2\left[\frac{-\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]. \end{aligned}$$

REMARK: For these formulae to hold, we need to use the assumptions that $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$, and ϵ_i independent (i.e., normality is not needed).

Proof. Recall that $\widehat{\beta}_1$ can be written as

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i,$$

where the constant

$$c_i = \frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

for $i = 1, 2, ..., n$. Therefore,

$$\begin{aligned} V(\widehat{\beta}_1) = V\left(\sum_{i=1}^{n} c_i Y_i\right) &= \sum_{i=1}^{n} c_i^2 V(Y_i) \\ &= \sigma^2 \sum_{i=1}^{n} \left[\frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]^2 \\ &= \frac{\sigma^2}{[\sum_{i=1}^{n}(x_i - \overline{x})^2]^2}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right] = \sigma^2\left[\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right], \end{aligned}$$

as claimed. The variance of $\widehat{\beta}_0$ is

$$\begin{aligned} V(\widehat{\beta}_0) &= V(\overline{Y} - \widehat{\beta}_1 \overline{x}) \\ &= V(\overline{Y}) + \overline{x}^2 V(\widehat{\beta}_1) - 2\overline{x}\mathrm{Cov}(\overline{Y}, \widehat{\beta}_1). \end{aligned}$$

Note that

$$\begin{aligned} V(\overline{Y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) &= \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

Also,

$$\begin{aligned} \mathrm{Cov}(\overline{Y}, \widehat{\beta}_1) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} c_i Y_i\right) \\ &= \frac{1}{n}\left[\sum_{i=1}^{n} \mathrm{Cov}(Y_i, c_i Y_i) + \sum_{i \neq j} \mathrm{Cov}(Y_i, c_j Y_j)\right] = \frac{1}{n}\sum_{i=1}^{n} c_i V(Y_i) = \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0. \end{aligned}$$

Therefore,

$$\begin{aligned} V(\widehat{\beta}_0) = \frac{\sigma^2}{n} + \sigma^2\left[\frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] &= \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \\ &= \sigma^2\left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 + n\overline{x}^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \\ &= \sigma^2\left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\right], \end{aligned}$$

as claimed. Finally, the covariance between $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is given by

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \mathrm{Cov}(\overline{Y} - \widehat{\beta}_1\overline{x}, \widehat{\beta}_1) = \mathrm{Cov}(\overline{Y}, \widehat{\beta}_1) - \overline{x}V(\widehat{\beta}_1).$$

We have already shown that $\text{Cov}(\overline{Y}, \widehat{\beta}_1) = 0$. Therefore,

$$\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\overline{x}V(\widehat{\beta}_1) = \sigma^2 \left[\frac{-\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right],$$

as claimed. □

Fact 3. The least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are normally distributed.

Proof. Recall that $\widehat{\beta}_1$ can be written as

$$\widehat{\beta}_1 = \sum_{i=1}^n c_i Y_i,$$

where the constant

$$c_i = \frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

for $i = 1, 2, ..., n$. However, under our model assumptions,

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

Therefore, $\widehat{\beta}_1$ is normally distributed since it is a linear combination of $Y_1, Y_2, ..., Y_n$. That $\widehat{\beta}_0$ is also normally distributed follows because

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x},$$

a linear combination of $\overline{Y}$ and $\widehat{\beta}_1$, both of which are normally distributed. Therefore, $\widehat{\beta}_0$ is normally distributed as well. Note that we have used the normality assumption on the errors ϵ_i to prove this fact. □

SUMMARY: In the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, so far we have shown that

$$\widehat{\beta}_0 \sim \mathcal{N}(\beta_0, c_{00}\sigma^2) \quad \text{and} \quad \widehat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2),$$

where

$$c_{00} = \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2} \quad \text{and} \quad c_{11} = \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

11.2.3 Estimating the error variance

REVIEW: In the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, we have just derived the sampling distributions of the least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$. We now turn our attention to estimating σ^2, the **error variance**.

NOTE: In the simple linear regression model, we define the ith **fitted value** by

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the least squares estimators. We define the ith **residual** by

$$e_i = Y_i - \widehat{Y}_i.$$

We define the **error (residual) sum of squares** by

$$\text{SSE} \equiv \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

Fact 4. In the simple linear regression model,

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

is an unbiased estimator of σ^2, that is, $E(\widehat{\sigma}^2) = \sigma^2$.

Proof. See WMS, pp 580-581. We will prove this later under a more general setting. □

NOTATION: Your authors denote the unbiased estimator of σ^2 by S^2. I don't like this notation because we have always used S^2 to denote the sample variance of $Y_1, Y_2, ..., Y_n$.

Fact 5. If $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, then

$$\frac{\text{SSE}}{\sigma^2} = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

The proof of this fact is beyond the scope of this course.

Fact 6. If $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, then $\widehat{\sigma}^2$ is **independent** of both $\widehat{\beta}_0$ and $\widehat{\beta}_1$. The proof of this fact is also beyond the scope of this course.

11.2.4 Inference for β_0 and β_1

INTEREST: In the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, the regression parameters β_0 and β_1 are unknown. It is therefore of interest to (a) construct confidence intervals and (b) perform hypothesis tests for these parameters. In practice, inference for the slope parameter β_1 is of primary interest because of its connection to the independent variable x in the model. Inference for β_0 is usually less meaningful, unless one is explicitly interested in the mean of Y when $x = 0$.

INFERENCE FOR β_1: Under our model assumptions, recall that the least squares estimator

$$\widehat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2),$$

where $c_{11} = 1/\sum_{i=1}^n (x_i - \overline{x})^2$. Standardizing, we have

$$Z = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{c_{11}\sigma^2}} \sim \mathcal{N}(0, 1).$$

Recall also that

$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_1$, it follows that Z and W are also independent. Therefore,

$$t = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{c_{11}\widehat{\sigma}^2}} = \frac{(\widehat{\beta}_1 - \beta_1)/\sqrt{c_{11}\sigma^2}}{\sqrt{\frac{(n-2)\widehat{\sigma}^2}{\sigma^2}/(n-2)}} \sim t(n-2).$$

Because $t \sim t(n-2)$, t is a pivot and we can write

$$P\left(-t_{n-2,\alpha/2} < \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{c_{11}\widehat{\sigma}^2}} < t_{n-2,\alpha/2}\right) = 1 - \alpha,$$

where $t_{n-2,\alpha/2}$ denotes the upper $\alpha/2$ quantile of the $t(n-2)$ distribution. Rearranging the event inside the probability symbol, we have

$$P\left(\widehat{\beta}_1 - t_{n-2,\alpha/2}\sqrt{c_{11}\widehat{\sigma}^2} < \beta_1 < \widehat{\beta}_1 + t_{n-2,\alpha/2}\sqrt{c_{11}\widehat{\sigma}^2}\right) = 1 - \alpha,$$

which shows that

$$\widehat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{c_{11}\widehat{\sigma}^2}.$$

is a $100(1-\alpha)$ percent confidence interval for β_1. If our interest was to test

$$H_0 : \beta_1 = \beta_{1,0}$$

versus

$$H_a : \beta_1 \neq \beta_{1,0},$$

where $\beta_{1,0}$ is a fixed value (often, $\beta_{1,0} = 0$), we would use

$$t = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\sqrt{c_{11}\widehat{\sigma}^2}}$$

as a test statistic and

$$\text{RR} = \{t : |t| > t_{n-2,\alpha/2}\}$$

as a level α rejection region. One sided tests would use a suitably-adjusted rejection region. Probability values are computed as areas under the $t(n-2)$ distribution.

INFERENCE FOR β_0: A completely analogous argument shows that

$$t = \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{c_{00}\widehat{\sigma}^2}} \sim t(n-2),$$

where $c_{00} = \sum_{i=1}^{n} x_i^2/n \sum_{i=1}^{n}(x_i - \overline{x})^2$. Therefore, a $100(1-\alpha)$ percent confidence interval for β_0 is

$$\widehat{\beta}_0 \pm t_{n-2,\alpha/2}\sqrt{c_{00}\widehat{\sigma}^2}.$$

In addition, a level α test of

$$H_0 : \beta_0 = \beta_{0,0}$$

versus

$$H_a : \beta_0 \neq \beta_{0,0}$$

can be performed using

$$t = \frac{\widehat{\beta}_0 - \beta_{0,0}}{\sqrt{c_{00}\widehat{\sigma}^2}}$$

as a test statistic and

$$\text{RR} = \{t : |t| > t_{n-2,\alpha/2}\}$$

as a level α rejection region. One sided tests would use a suitably-adjusted rejection region. Probability values are computed as areas under the $t(n-2)$ distribution.

11.2.5 Confidence intervals for $E(Y \mid x^*)$

INTEREST: In the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, we first consider constructing confidence intervals for linear parametric functions of the form

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

where a_0 and a_1 are fixed constants.

ESTIMATION: Using the least squares estimators of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as point estimators for β_0 and β_1, respectively, a point estimator for θ is

$$\widehat{\theta} = a_0\widehat{\beta}_0 + a_1\widehat{\beta}_1.$$

It is easy to see that $\widehat{\theta}$ is an unbiased estimator for θ since

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

It is also possible to show that

$$V(\widehat{\theta}) \equiv \sigma^2_{\widehat{\theta}} = \sigma^2 \left[\frac{\frac{a_0^2}{n}\sum_{i=1}^n x_i^2 + a_1^2 - 2a_0a_1\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right].$$

Since $\widehat{\theta}$ is a linear combination of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, both of which are normally distributed, it follows that

$$\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2_{\widehat{\theta}}).$$

INFERENCE: The variance $\sigma^2_{\widehat{\theta}}$ depends on the unknown parameter σ^2. An estimate of $\sigma^2_{\widehat{\theta}}$ is given by

$$\widehat{\sigma}^2_{\widehat{\theta}} = \widehat{\sigma}^2 \left[\frac{\frac{a_0^2}{n}\sum_{i=1}^n x_i^2 + a_1^2 - 2a_0a_1\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right],$$

where

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2}.$$

Because $\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2_{\widehat{\theta}})$, we have by standardization,

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{N}(0, 1).$$

Recall also that

$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, it is independent of $\widehat{\theta}$ and hence Z and W are independent. Therefore,

$$t = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}_{\widehat{\theta}}} = \frac{(\widehat{\theta} - \theta)/\sigma_{\widehat{\theta}}}{\sqrt{\frac{(n-2)\widehat{\sigma}^2}{\sigma^2}/(n-2)}} \sim t(n-2).$$

Since t is a pivotal quantity, a $100(1-\alpha)$ percent confidence interval for θ is

$$\widehat{\theta} \pm t_{n-2,\alpha/2}\widehat{\sigma}_{\widehat{\theta}}.$$

In addition, tests of hypotheses concerning θ use the $t(n-2)$ distribution.

SPECIAL CASE: A special case of the preceding result is estimating the mean value of Y for a fixed value of x, say, x^*. In our simple linear regression model, we know that

$$E(Y|x^*) = \beta_0 + \beta_1 x^*,$$

which is just a linear combination of the form $\theta = a_0\beta_0 + a_1\beta_1$, where $a_0 = 1$ and $a_1 = x^*$. Therefore,

$$\widehat{\theta} \equiv \widehat{E(Y|x^*)} = \widehat{\beta}_0 + \widehat{\beta}_1 x^*$$

is an unbiased estimator of $\theta \equiv E(Y|x^*) = \beta_0 + \beta_1 x^*$ and its variance is

$$V(\widehat{\theta}) = \sigma^2_{\widehat{\theta}} = \sigma^2 \left[\frac{\frac{a_0^2}{n}\sum_{i=1}^n x_i^2 + a_1^2 - 2a_0 a_1 \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right] = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right].$$

Applying the preceding general results to this special case, a $100(1-\alpha)$ **percent confidence interval** for $E(Y|x^*) = \beta_0 + \beta_1 x^*$ is given by

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2} \sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right]}.$$

NOTE: The confidence interval for $E(Y|x^*) = \beta_0 + \beta_1 x^*$ will be different for different values of x^*; see pp 597 (WMS). It is easy to see that the width of the confidence interval will be smallest when $x^* = \overline{x}$ and will increase as the distance between x^* and $\overline{x}$ increases. That is, more precise inference for $\theta = E(Y|x^*) = \beta_0 + \beta_1 x^*$ will result when x^* is close to $\overline{x}$. When x^* is far away from $\overline{x}$, our precision may not be adequate. It is sometimes desired to estimate $E(Y|x^*) = \beta_0 + \beta_1 x^*$ for a value of x^* outside the range of x values in the observed data. This is called **extrapolation**. In order for these inferences to be valid, we must believe that the model is accurate even for x values outside the range where we have observed data. In some situations, this may be reasonable; in others, we may have no basis for making such a claim without data to support it.

11.2.6 Prediction intervals for Y^*

PREDICTION: For some research questions, we may not be interested in the mean $E(Y|x^*) = \beta_0 + \beta_1 x^*$, but rather in the actual value of Y we may observe when $x = x^*$. On the surface, this may sound like the same problem, but they are very different.

EXAMPLE: Suppose that we have adopted the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where Y = 1st year final course percentage in MATH 141 and x = SAT MATH score. Consider these (very different) questions:

- What is an estimate of the mean MATH 141 course percentage for those students who made a SAT math score of $x = 700$?
- What MATH 141 course percentage would you predict for your friend Joe, who made a SAT math score of $x = 700$?

The first question deals with **estimating** $E(Y|x^* = 700)$, a population mean. The second question deals with **predicting** the value of the random variable Y, say Y^*, that comes from a distribution with mean $E(Y|x^* = 700)$. Estimating $E(Y|x^* = 700)$ is much easier than predicting Y^*.

GOAL: Our goal is to construct a **prediction interval** for a new value of Y, which we denote by Y^*. Our point predictor for Y^*, when $x = x^*$, is

$$\widehat{Y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^*.$$

This point predictor is the same as the point estimator we used to estimate $E(Y|x^*) = \beta_0 + \beta_1 x^*$. However, we use a different symbol in this context to remind ourselves that we are predicting Y^*, not estimating $E(Y|x^*)$. We call $\widehat{Y}^*$ a **prediction** to make the distinction clear.

TERMINOLOGY: Define the random variable

$$U = Y^* - \widehat{Y}^*.$$

We call U the **prediction error**. Note that

$$\begin{aligned} E(U) = E(Y^* - \widehat{Y}^*) &= E(Y^*) - E(\widehat{Y}^*) \\ &= (\beta_0 + \beta_1 x^*) - E(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \\ &= (\beta_0 + \beta_1 x^*) - (\beta_0 + \beta_1 x^*) = 0. \end{aligned}$$

That is, the prediction error U is an unbiased estimator of 0. The variance of U is

$$V(U) = V(Y^* - \widehat{Y}^*) = V(Y^*) + V(\widehat{Y}^*) - 2\text{Cov}(Y^*, \widehat{Y}^*).$$

Under our model assumptions, we know that $V(Y^*) = \sigma^2$. In addition,

$$V(\widehat{Y}^*) = V(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right],$$

which is the same as the variance of $\widehat{E(Y|x^*)}$. Finally,

$$\text{Cov}(Y^*, \widehat{Y}^*) = 0,$$

because of the independence assumption. More specifically, $\widehat{Y}^*$ is a function of $Y_1, Y_2, ..., Y_n$, the observed data. The value Y^* is a new value of Y, and, hence, is independent of $Y_1, Y_2, ..., Y_n$. Therefore,

$$\begin{aligned} V(U) &= V(Y^* - \widehat{Y}^*) = V(Y^*) + V(\widehat{Y}^*) - 2\text{Cov}(Y^*, \widehat{Y}^*) \\ &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right]. \end{aligned}$$

We finally note that the prediction error $U = Y^* - \widehat{Y}^*$ is normally distributed because it is a linear combination of Y^* and $\widehat{Y}^*$, both of which are normally distributed. We have shown that

$$U = Y^* - \widehat{Y}^* \sim \mathcal{N}\left\{0,\ \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]\right\}.$$

Standardizing, we have

$$Z = \frac{Y^* - \widehat{Y}^*}{\sqrt{\sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} \sim \mathcal{N}(0, 1).$$

Recall also that

$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, it is independent of $\widehat{Y}^*$ and hence Z and W are independent. Therefore,

$$t = \frac{Z}{\sqrt{W/(n-2)}} = \frac{Y^* - \widehat{Y}^*}{\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} \sim t(n-2).$$

Using t as a pivot, we can write

$$P\left(-t_{n-2,\alpha/2} < \frac{Y^* - \widehat{Y}^*}{\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} < t_{n-2,\alpha/2}\right) = 1 - \alpha,$$

where $t_{n-2,\alpha/2}$ denotes the upper $\alpha/2$ quantile of the $t(n-2)$ distribution. Rearranging the event inside the probability symbol, we have

$$P\Bigg(\widehat{Y}^* - t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]} < Y^* < \widehat{Y}^* + t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}\Bigg) = 1 - \alpha.$$

We call

$$\widehat{Y}^* \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}$$

is a $100(1-\alpha)$ **percent prediction interval** for Y^*.

NOTE: It is of interest to compare the confidence interval for $E(Y|x^*)$, given by

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]},$$

to the prediction interval for Y^*, given by

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}.$$

As we can see, the prediction interval when $x = x^*$ will always be wider than the corresponding confidence interval for $E(Y|x^*)$. This is a result of the additional uncertainty which arises from having to predict the value of a new random variable.

11.2.7 Example

Example 11.1. A botanist is studying the absorption of salts by living plant cells. She prepares $n = 30$ dishes containing potato slices and adds a bromide solution to each dish. She waits a duration of time x (measured in hours) and then analyzes the potato slices for absorption of bromide ions (y, measured in mg/1000g). Here are the data.

Dish	x	y	Dish	x	y	Dish	x	y
1	16.4	5.2	11	65.5	15.3	21	121.6	23.0
2	18.2	1.0	12	68.6	11.2	22	121.8	22.3
3	21.6	4.8	13	75.4	16.9	23	122.4	24.6
4	22.3	2.7	14	76.3	12.3	24	124.4	22.4
5	24.1	1.1	15	88.0	15.3	25	128.0	28.1
6	29.7	3.5	16	92.0	19.9	26	128.0	20.5
7	34.6	8.7	17	96.6	21.1	27	131.2	26.5
8	35.2	10.1	18	98.1	19.5	28	140.7	31.3
9	56.5	11.4	19	103.9	20.7	29	145.8	29.1
10	58.7	10.8	20	115.9	22.4	30	149.5	32.6

Table 11.1: Botany data. Absorption of bromide ions (y, measured in mg/1000g) and time (x, measured in hours).

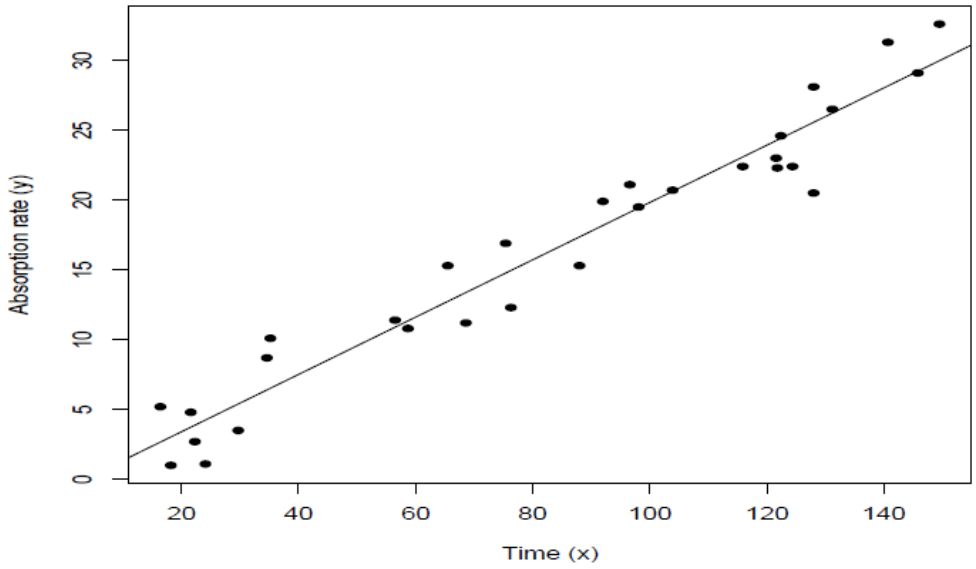

Figure 11.1: Botany data. Absorption of bromide ions (y, measured in mg/1000g) versus time (x, measured in hours). The least squares regression line has been superimposed.

REGRESSION MODEL: From the scatterplot in Figure 11.1, the linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., 30$, appears to be appropriate. Fitting this model in R, we get the output:

```
> summary(fit)
Call: lm(formula = absorp ~ time)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.700374   0.894462  -0.783     0.44
time         0.205222   0.009509  21.582   <2e-16 ***

Residual standard error: 2.236 on 28 degrees of freedom
Multiple R-squared: 0.9433,  Adjusted R-squared: 0.9413
F-statistic: 465.8 on 1 and 28 DF,  p-value: < 2.2e-16
```

OUTPUT: The `Estimate` output gives the least squares estimates $\widehat{\beta}_0 \approx -0.700$ and $\widehat{\beta}_1 \approx 0.205$. The equation of the least squares regression line is therefore

$$\widehat{Y} = -0.700 + 0.205x,$$

or, in other words,

$$\widehat{\texttt{ABSORPTION}} = -0.700 + 0.205\texttt{TIME}.$$

The `Std.Error` output gives

$$\begin{aligned} 0.894462 &= \widehat{\text{se}}(\widehat{\beta}_0) = \sqrt{c_{00}\widehat{\sigma}^2} \\ 0.009509 &= \widehat{\text{se}}(\widehat{\beta}_1) = \sqrt{c_{11}\widehat{\sigma}^2}, \end{aligned}$$

which are the estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively, where

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{30-2} = (2.236)^2 \approx 5.00$$

is the square of the `Residual standard error`. The `t value` output gives the t statistics

$$\begin{aligned} t = -0.783 &= \frac{\widehat{\beta}_0 - 0}{\sqrt{c_{00}\widehat{\sigma}^2}} \\ t = 21.582 &= \frac{\widehat{\beta}_1 - 0}{\sqrt{c_{11}\widehat{\sigma}^2}}, \end{aligned}$$

which test $H_0 : \beta_0 = 0$ versus $H_a : \beta_0 \neq 0$ and $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, respectively. Two-sided probability values are in `Pr(>|t|)`. We see that

- there is insufficient evidence against $H_0 : \beta_0 = 0$ (p-value $= 0.44$).
- there is strong evidence against $H_0 : \beta_1 = 0$ (p-value < 0.0001). This means that the absorption rate Y is (significantly) linearly related to duration time x.

CONFIDENCE INTERVALS: Ninety-five percent confidence intervals for β_0 and β_1 are

$$\begin{aligned} \widehat{\beta}_0 \pm t_{28,0.025}\widehat{\text{se}}(\widehat{\beta}_0) &\Longrightarrow -0.700 \pm 2.048(0.894) \Longrightarrow (-2.53, 1.13) \\ \widehat{\beta}_1 \pm t_{28,0.025}\widehat{\text{se}}(\widehat{\beta}_1) &\Longrightarrow 0.205 \pm 2.048(0.010) \Longrightarrow (0.18, 0.23). \end{aligned}$$

We are 95 percent confident that β_0 is between -2.53 and 1.13. We are 95 percent confident that β_1 is between 0.18 and 0.23.

PREDICTION: Suppose that we are interested estimating $E(Y|x)$ and predicting a new Y when $x^* = 80$ hours. We use R to compute the following:

```
> predict(fit,data.frame(time=80),level=0.95,interval="confidence")
     fit      lwr      upr
15.71735 14.87807 16.55663
> predict(fit,data.frame(time=80),level=0.95,interval="prediction")
     fit      lwr      upr
15.71735 11.06114 20.37355
```

- Note that

$$\widehat{E(Y|x^*)} = \widehat{Y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^* = -0.700 + 0.205(80) \approx 15.71735.$$

- A 95 percent **confidence interval** for $E(Y|x^* = 80)$ is $(14.88, 16.56)$. When the duration time is $x = 80$ hours, we are 95 percent confident that the mean absorption is between 14.88 and 16.56 mg/1000g.

- A 95 percent **prediction interval** for Y^*, when $x = 80$, is $(11.06, 20.37)$. When the duration time is $x = 80$ hours, we are 95 percent confident that the absorption for a new dish will be between 11.06 and 20.37 mg/1000g. □

11.3 Correlation

RECALL: In the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, it is assumed that the independent variable x is fixed. This assumption is plausible in designed experiments, say, where the investigator has control over which values of x will be included in the experiment. For example,

- x = dose of a drug, Y = change in blood pressure for a human subject

- x = concentration of toxic substance, Y = number of mutant offspring observed for a pregnant rat
- x = time, Y = absorption of bromide ions.

In other settings, it is unreasonable to think that the researcher can "decide" beforehand which values of x will be observed. Consider the following examples:

- X = weight, Y = height of a human subject
- X = average heights of plants in a plot, Y = yield
- X = STAT 513 HW score, Y = STAT 513 final exam score.

In each of these instances, the independent variable X is best regarded as **random**. It is unlikely that the researcher can control (fix) its value.

IMPORTANT: When both X and Y are best regarded as random, it is conventional to model the observed data as realizations of (X, Y), a bivariate random vector. A popular model for (X, Y) is the bivariate normal distribution.
RECALL: The random vector (X, Y) is said to have a **bivariate normal distribution** if its (joint) pdf is given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-Q/2}$$

for all $(x, y)' \in \mathcal{R}^2$, where

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right].$$

Under the bivariate normal model, recall from STAT 511 that

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

where

$$\begin{aligned} \beta_0 &= \mu_Y - \beta_1\mu_X \\ \beta_1 &= \rho\left(\frac{\sigma_Y}{\sigma_X}\right). \end{aligned}$$

IMPORTANT: Note that because

$$\beta_1 = \rho\left(\frac{\sigma_Y}{\sigma_X}\right),$$

the correlation ρ and the (population) slope parameter β_1 have the same sign.

ESTIMATION: Suppose that $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is an iid sample of size n from a bivariate normal distribution with marginal means μ_X and μ_Y, marginal variances σ_X^2 and σ_Y^2, and correlation ρ. The likelihood function is given by

$$\begin{aligned} L(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) &= \prod_{i=1}^{n} f_{X,Y}(x_i, y_i) \\ &= \left(\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\right)^n e^{-\sum_{i=1}^n Q_i/2}, \end{aligned}$$

where

$$Q_i = \frac{1}{1-\rho^2}\left[\left(\frac{x_i-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x_i-\mu_X}{\sigma_X}\right)\left(\frac{y_i-\mu_Y}{\sigma_Y}\right) + \left(\frac{y_i-\mu_Y}{\sigma_Y}\right)^2\right].$$

The maximum likelihood estimators are

$$\widehat{\mu}_X = \overline{X}, \qquad \widehat{\mu}_Y = \overline{Y}, \qquad \widehat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2, \qquad \widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2,$$

and

$$\widehat{\rho} = r = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

HYPOTHESIS TEST: In the bivariate normal model, suppose that it is desired to test

$$H_0 : \rho = 0$$

versus

$$H_a : \rho \neq 0.$$

Since ρ and β_1 always have the same sign, mathematically, this is equivalent to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0.$$

That is, we can use the statistic

$$t = \frac{\widehat{\beta}_1}{\sqrt{c_{11}\widehat{\sigma}^2}}$$

to test $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$. A level α rejection region is

$$\text{RR} = \{t : |t| > t_{\alpha/2,n-2}\}.$$

One sided tests can be performed similarly.

RESULT: Simple calculations show that

$$t = \frac{\widehat{\beta}_1}{\sqrt{c_{11}\widehat{\sigma}^2}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Therefore, the test of $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$ (or any other suitable H_a) can be performed using only the calculated value of r.

REMARK: Even though the tests

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0$$

and

$$H_0 : \rho = 0$$

versus

$$H_a : \rho \neq 0$$

are carried out in the exact same manner, it is important to remember that the interpretation of the results is very different, depending on which test we are performing.

- In the first test, we are determining whether or not there is a **linear relationship** between Y and x. The independent variable x is best regarded as fixed.

- In the second test, we are actually determining whether or not the random variables X and Y are **independent**. Recall that in the bivariate normal model,

$$X \text{ and } Y \text{ independent} \Longleftrightarrow \rho = 0.$$

REMARK: In some problems, it may be of interest to test

$$H_0 : \rho = \rho_0$$

versus

$$H_a : \rho \neq \rho_0$$

(or any other suitable H_a), where $\rho_0 \neq 0$. In this case, there is no equivalence between the two tests (as when $\rho_0 = 0$) that we saw before. We are forced to use a different test (i.e., one that is based on large sample theory).

ASYMPTOTIC RESULT: Suppose that $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is an iid sample of size n from a bivariate normal distribution with marginal means μ_X and μ_Y, marginal variances σ_X^2 and σ_Y^2, and correlation ρ. Let

$$r = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}$$

denote the maximum likelihood estimator of ρ. For large n, the statistic

$$W = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \sim \mathcal{AN} \left[\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right), \frac{1}{n-3} \right].$$

IMPLEMENTATION: This asymptotic result above can be used to test

$$H_0 : \rho = \rho_0$$

versus

$$H_a : \rho \neq \rho_0$$

(or any other suitable H_a), where $\rho_0 \neq 0$. The test statistic is the standardized value of W, computed under H_0, that is,

$$Z = \frac{\frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) - \frac{1}{2} \ln \left(\frac{1+\rho_0}{1-\rho_0} \right)}{1/\sqrt{n-3}}.$$

An approximate level α rejection region is

$$\text{RR} = \{z : |z| > z_{\alpha/2}\},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. One sided tests can be performed similarly.

11.4 Multiple linear regression models

11.4.1 Introduction

PREVIEW: We have already considered the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$. Our interest now is to extend this basic model to include multiple independent variables $x_1, x_2, ..., x_k$. Specifically, we consider models of the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$. We call this a **multiple linear regression model**.

- There are now $p = k + 1$ regression parameters β_0, β_1, ..., β_k. These are unknown and are to be estimated with the observed data.

- Schematically, we can envision the observed data as follows:

Individual	Y	x_1	x_2	$\cdots$	x_k
1	Y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	Y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	Y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

  That is, each of the n individuals contributes a response Y and a value of each of the independent variables $x_1, x_2, ..., x_k$.

- We continue to assume that $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$.

- We also assume that the independent variables $x_1, x_2, ..., x_k$ are fixed and measured without error. Therefore, Y is normally distributed with

$$\begin{aligned} E(Y|x_1, x_2, ..., x_k) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \\ V(Y|x_1, x_2, ..., x_k) &= \sigma^2. \end{aligned}$$

PREVIEW: To fit the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

we will still use the **method of least squares**. However, simple computing formulae for the least squares estimators of β_0, β_1, ..., β_k are no longer available (as they were in the simple linear regression model). It is advantageous to express multiple linear regression models in terms of matrices and vectors. This greatly streamlines notation and makes calculations tractable.

11.4.2 Matrix representation

MATRIX REPRESENTATION: Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$. Define

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

With these definitions, the model above can be expressed equivalently as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

In this equivalent representation,

- $\mathbf{Y}$ is an $n \times 1$ (random) vector of responses
- $\mathbf{X}$ is an $n \times p$ (fixed) matrix of independent variable measurements ($p = k + 1$)
- $\boldsymbol{\beta}$ is a $p \times 1$ (fixed) vector of unknown population regression parameters
- $\boldsymbol{\epsilon}$ is an $n \times 1$ (random) vector of unobserved errors.

LEAST SQUARES: The notion of least squares is the same as it was in the simple linear regression model. To fit a multiple linear regression model, we want to find the values of $\beta_0, \beta_1, ..., \beta_k$ that minimize

$$Q(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})]^2,$$

or, in matrix notation, the value of $\boldsymbol{\beta}$ that minimizes

$$Q = Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Because $Q(\boldsymbol{\beta})$ is a scalar function of the $p = k + 1$ elements of $\boldsymbol{\beta}$, it is possible to use calculus to determine the values of the p elements that minimize it. Formally, we can take the p partial derivatives with respect to each of $\beta_0, \beta_1, ..., \beta_k$ and set these equal to zero; i.e.,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial Q}{\partial \beta_0} \\ \frac{\partial Q}{\partial \beta_1} \\ \vdots \\ \frac{\partial Q}{\partial \beta_k} \end{pmatrix} \stackrel{\text{set}}{=} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

These are called the **normal equations**. Solving the normal equations for $\beta_0, \beta_1, ..., \beta_k$ gives the least squares estimators, which we denote by $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$.

NORMAL EQUATIONS: Using the calculus of matrices makes this much easier; in particular, the normal equations above can be expressed as

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}.$$

Provided that $\mathbf{X}'\mathbf{X}$ is full rank, the (unique) solution is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

This is the **least squares estimator** of $\boldsymbol{\beta}$. The fitted regression model is

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}},$$

or, equivalently,

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}.$$

NOTE: For the least squares estimator $\widehat{\boldsymbol{\beta}}$ to be unique, we need $\mathbf{X}$ to be of **full column rank**; i.e., $r(\mathbf{X}) = p = k + 1$. That is, there must be no linear dependencies among the columns of $\mathbf{X}$. If $r(\mathbf{X}) < p$, then $\mathbf{X}'\mathbf{X}$ does not have a unique inverse. In this case, the normal equations can not be solved uniquely. We will henceforth assume that $\mathbf{X}$ is of full column rank.

11.4.3 Random vectors: Important results

IMPORTANCE: Because multiple linear regression models are best presented in terms of (random) vectors and matrices, it is important to extend the notions of mean, variance, and covariance to random vectors. Doing so allows us to examine sampling distributions and the resulting inference that arises in multiple linear regression models.

TERMINOLOGY: Suppose that $Z_1, Z_2, ..., Z_n$ are random variables. We call

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}$$

a **random vector**. The multivariate probability density function (pdf) of $\mathbf{Z}$ is denoted by $f_{\mathbf{Z}}(\mathbf{z})$. The function $f_{\mathbf{Z}}(\mathbf{z})$ describes probabilistically how the random variables $Z_1, Z_2, ..., Z_n$ are jointly distributed.

- If $Z_1, Z_2, ..., Z_n$ are independent variables, then

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} f_{Z_i}(z_i),$$

  where $f_{Z_i}(z_i)$ is the marginal pdf of Z_i.

- If $Z_1, Z_2, ..., Z_n$ are iid from a common marginal pdf, say, $f_Z(z)$, then

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} f_Z(z_i).$$

TERMINOLOGY: Suppose $Z_1, Z_2, ..., Z_n$ are random variables with means $E(Z_i) = \mu_i$ and variances $V(Z_i) = \sigma_i^2$, for $i = 1, 2, ..., n$, and covariances $\text{Cov}(Z_i, Z_j) = \sigma_{ij}$ for $i \neq j$. The **mean** of a random vector $\mathbf{Z}$ is given by

$$E(\mathbf{Z}) = E\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} E(Z_1) \\ E(Z_2) \\ \vdots \\ E(Z_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}.$$

The **variance** of $\mathbf{Z}$ is

$$V(\mathbf{Z}) = V\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix} = \mathbf{V}.$$

- $\mathbf{V}$ is an $n \times n$ matrix. It is also called the **variance-covariance matrix** of $\mathbf{Z}$.
- $\mathbf{V}$ consists of the n variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ on the diagonal and the $2\binom{n}{2}$ covariance terms $\text{Cov}(Z_i, Z_j)$, for $i \neq j$, on the off-diagonal.
- Since $\text{Cov}(Z_i, Z_j) = \text{Cov}(Z_j, Z_i)$, $\mathbf{V}$ is **symmetric**; i.e., $\mathbf{V}' = \mathbf{V}$.

TERMINOLOGY: Suppose that

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}$$

are random vectors. The **covariance** between $\mathbf{Y}$ and $\mathbf{Z}$ is

$$\text{Cov}(\mathbf{Y}, \mathbf{Z}) = \begin{pmatrix} \text{Cov}(Y_1, Z_1) & \text{Cov}(Y_1, Z_2) & \cdots & \text{Cov}(Y_1, Z_m) \\ \text{Cov}(Y_2, Z_1) & \text{Cov}(Y_2, Z_2) & \cdots & \text{Cov}(Y_2, Z_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_n, Z_1) & \text{Cov}(Y_n, Z_2) & \cdots & \text{Cov}(Y_n, Z_m) \end{pmatrix}_{n \times m}.$$

RESULTS: Suppose $\mathbf{Z}$ is a random vector with mean $E(\mathbf{Z}) = \boldsymbol{\mu}$ and variance-covariance matrix $V(\mathbf{Z}) = \mathbf{V}$. Suppose $\mathbf{a}$ is a nonrandom (constant) vector and that $\mathbf{A}$ and $\mathbf{B}$ are nonrandom (constant) matrices.

1. $E(\mathbf{a} + \mathbf{BZ}) = \mathbf{a} + \mathbf{B}E(\mathbf{Z}) = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}$
2. $V(\mathbf{a} + \mathbf{BZ}) = \mathbf{B}V(\mathbf{Z})\mathbf{B}' = \mathbf{BVB}'$
3. $\text{Cov}(\mathbf{AY}, \mathbf{BZ}) = \mathbf{A}\text{Cov}(\mathbf{Y}, \mathbf{Z})\mathbf{B}'$.

TERMINOLOGY: Let $\mathbf{Y}$ be an $n \times 1$ random vector with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{V}$. Let $\mathbf{A}$ be an $n \times n$ nonrandom matrix. We call $\mathbf{Y}'\mathbf{AY}$ a **quadratic form**. The mean of a quadratic form is

$$E(\mathbf{Y}'\mathbf{AY}) = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{AV}),$$

where $\text{tr}(\cdot)$ means "trace," that is, $\text{tr}(\mathbf{AV})$ is the sum of the diagonal elements of $\mathbf{AV}$.

REMARK: It is important to see that a quadratic form $\mathbf{Y}'\mathbf{AY}$ is a scalar random variable. Therefore, its mean $E(\mathbf{Y}'\mathbf{AY})$ is a scalar constant. Quadratic forms are important in the theory of linear (regression) models. It turns out that **sums of squares** (which appear in analysis of variance tables) can always be written as quadratic forms.

11.4.4 Multivariate normal distribution

TERMINOLOGY: Suppose that $Z_1, Z_2, ..., Z_n$ are iid $\mathcal{N}(0,1)$ random variables. The joint pdf of $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)'$, for all $\mathbf{z} \in \mathcal{R}^n$, is given by

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} f_Z(z_i) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n z_i^2/2} = (2\pi)^{-n/2} \exp(-\mathbf{z}'\mathbf{z}/2). \end{aligned}$$

If $\mathbf{Z}$ has a pdf given by $f_{\mathbf{Z}}(\mathbf{z})$, we say that $\mathbf{Z}$ has a **standard multivariate normal distribution**; i.e., a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{I}$. Here,

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

That is, $\mathbf{0}$ is an $n \times 1$ zero vector and $\mathbf{I}$ is the $n \times n$ identity matrix. We write $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$. Note that

$$Z_1, Z_2, ..., Z_n \sim \text{iid } \mathcal{N}(0,1) \Longleftrightarrow \mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}).$$

TERMINOLOGY: The random vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)'$ is said to have a **multivariate normal distribution** with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{V}$ if its joint pdf is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right\},$$

for all $\mathbf{y} \in \mathcal{R}^n$. We write $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$.

FACTS:

- If $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)' \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$, then $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, for each $i = 1, 2, ..., n$.
- If $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{a}_{m \times 1}$ and $\mathbf{B}_{m \times n}$ are nonrandom, then

$$\mathbf{U} = \mathbf{a} + \mathbf{B}\mathbf{Y} \sim \mathcal{N}_m(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{V}\mathbf{B}').$$

APPLICATION: Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. Equivalently, we can write this model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Note that

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{0} = \mathbf{X}\boldsymbol{\beta}$$

and

$$V(\mathbf{Y}) = V(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = V(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}.$$

Because $\mathbf{Y}$ is a linear combination of $\boldsymbol{\epsilon}$, which is normally distributed by assumption, it follows that

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}). \; \square$$

11.4.5 Estimating the error variance

REVIEW: Consider the multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Recall that the least squares estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Our next task is to estimate the error variance σ^2.
TERMINOLOGY: We define the **error (residual) sum of squares** as

$$\begin{aligned} \text{SSE} &= (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= (\mathbf{Y} - \widehat{\mathbf{Y}})'(\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{e}'\mathbf{e}. \end{aligned}$$

- The $n \times 1$ vector $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ contains the least squares **fitted values**.
- The $n \times 1$ vector $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ contains the least squares **residuals**.

TERMINOLOGY: Consider the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and define

$$\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

$\mathbf{M}$ is called the **hat matrix**. Many important quantities in linear regression can be written as functions of the hat matrix. For example, the vector of fitted values can be written as

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{M}\mathbf{Y}.$$

The vector of residuals can be written as

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{M}\mathbf{Y} = (\mathbf{I} - \mathbf{M})\mathbf{Y}.$$

The error (residual) sum of squares can be written as

$$\text{SSE} = (\mathbf{Y} - \widehat{\mathbf{Y}})'(\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}.$$

Note that $\text{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}$ is a quadratic form.

FACTS: The matrix $\mathbf{M}$ possesses the following properties:

- $\mathbf{M}$ is symmetric, i.e., $\mathbf{M}' = \mathbf{M}$.
- $\mathbf{M}$ is idempotent, i.e., $\mathbf{M}^2 = \mathbf{M}$.
- $\mathbf{MX} = \mathbf{X}$, i.e., $\mathbf{M}$ projects each column of $\mathbf{X}$ onto itself.

RESULT: Consider the multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Let $p = k + 1$ denote the number of regression parameters in the model. The quantity

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-p}$$

is an **unbiased estimator** of σ^2, that is, $E(\widehat{\sigma}^2) = \sigma^2$.

Proof. Recall that $\text{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}$. Because $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $V(\mathbf{Y}) = \sigma^2\mathbf{I}$, we have

$$\begin{aligned} E(\text{SSE}) &= E[\mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}] \\ &= (\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{M})\mathbf{X}\boldsymbol{\beta} + \text{tr}[(\mathbf{I} - \mathbf{M})\sigma^2\mathbf{I}]. \end{aligned}$$

The first term $(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{M})\mathbf{X}\boldsymbol{\beta} = 0$ because

$$(\mathbf{I} - \mathbf{M})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{M}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}.$$

Because the $\text{tr}(\cdot)$ function is linear,

$$\begin{aligned} \text{tr}[(\mathbf{I} - \mathbf{M})\sigma^2\mathbf{I}] &= \sigma^2[\text{tr}(\mathbf{I}) - \text{tr}(\mathbf{M})] \\ &= \sigma^2\{n - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\}. \end{aligned}$$

Since $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ for any matrices $\mathbf{A}$ and $\mathbf{B}$, taking $\mathbf{A} = \mathbf{X}$ and $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, we can write the last expression as

$$\begin{aligned}\sigma^2\{n - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\} &= \sigma^2\{n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}]\} \\ &= \sigma^2[n - \text{tr}(\mathbf{I}_p)] = \sigma^2(n-p),\end{aligned}$$

since $\mathbf{I}_p = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$ is $p \times p$. We have shown that $E(\text{SSE}) = \sigma^2(n-p)$. Thus,

$$E(\widehat{\sigma}^2) = E\left(\frac{\text{SSE}}{n-p}\right) = \frac{\sigma^2(n-p)}{n-p} = \sigma^2,$$

showing that $\widehat{\sigma}^2$ is an unbiased estimator of σ^2. $\square$

RESULT: Consider the multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Let $p = k + 1$ denote the number of regression parameters in the model. Under these model assumptions,

$$\frac{\text{SSE}}{\sigma^2} = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

The proof of this result is beyond the scope of this course.

11.4.6 Sampling distribution of $\hat{\boldsymbol{\beta}}$

GOAL: Consider the multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. We now investigate the **sampling distribution** of the least squares estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

MEAN AND VARIANCE: The mean of $\widehat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned}E(\widehat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.\end{aligned}$$

This shows that $\widehat{\boldsymbol{\beta}}$ is an **unbiased estimator** of $\boldsymbol{\beta}$. The variance of $\widehat{\boldsymbol{\beta}}$ is

$$\begin{aligned} V(\widehat{\boldsymbol{\beta}}) &= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{Y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

NORMALITY: Since $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is a linear combination of $\mathbf{Y}$, which is (multivariate) normal under our model assumptions, it follows that $\widehat{\boldsymbol{\beta}}$ is normally distributed as well. Therefore, we have shown that

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]. \ \square$$

IMPLICATIONS: The following results are direct consequences of our recent discussion:

1. $E(\widehat{\beta}_j) = \beta_j$, for $j = 0, 1, ..., k$; that is, the least squares estimators are unbiased.

2. $V(\widehat{\beta}_j) = c_{jj}\sigma^2$, for $j = 0, 1, ..., k$, where

   $$c_{jj} = (\mathbf{X}'\mathbf{X})^{-1}_{jj}$$

   is the corresponding jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. An estimate of $V(\widehat{\beta}_j)$ is

   $$\widehat{V}(\widehat{\beta}_j) = c_{jj}\widehat{\sigma}^2 = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{jj},$$

   where

   $$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-p}.$$

3. $\text{Cov}(\widehat{\beta}_i, \widehat{\beta}_j) = c_{ij}\sigma^2$, where

   $$c_{ij} = (\mathbf{X}'\mathbf{X})^{-1}_{ij}$$

   is the corresponding ith row, jth column entry of $(\mathbf{X}'\mathbf{X})^{-1}$, for $i, j = 0, 1, ..., k$. An estimate of $\text{Cov}(\widehat{\beta}_i, \widehat{\beta}_j)$ is

   $$\widehat{\text{Cov}}(\widehat{\beta}_i, \widehat{\beta}_j) = c_{ij}\widehat{\sigma}^2 = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{ij}.$$

4. Marginally, $\widehat{\beta}_j \sim \mathcal{N}(\beta_j, c_{jj}\sigma^2)$, for $j = 0, 1, ..., k$.

11.4.7 Inference for regression parameters

IMPORTANCE: Consider our multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. Confidence intervals and hypothesis tests for β_j can help us assess the importance of using the independent variable x_j in a model with the other independent variables. That is, inference regarding β_j is always **conditional** on the other variables being included in the model.

CONFIDENCE INTERVALS: Since $\widehat{\beta}_j \sim \mathcal{N}(\beta_j, c_{jj}\sigma^2)$, for $j = 0, 1, 2, ..., k$, it follows, from standardization, that

$$Z_j = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}\sigma^2}} \sim \mathcal{N}(0, 1).$$

Recall also that

$$W = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_j$, it follows that Z and W are also independent. Therefore,

$$t = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}\widehat{\sigma}^2}} = \frac{(\widehat{\beta}_j - \beta_j)/\sqrt{c_{jj}\sigma^2}}{\sqrt{\frac{(n-p)\widehat{\sigma}^2}{\sigma^2}/(n-p)}} \sim t(n-p).$$

Because $t \sim t(n-p)$, t is a pivot and we can write

$$P\left(-t_{n-p,\alpha/2} < \frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}\widehat{\sigma}^2}} < t_{n-p,\alpha/2}\right) = 1 - \alpha,$$

where $t_{n-p,\alpha/2}$ denotes the upper $\alpha/2$ quantile of the $t(n-p)$ distribution. Rearranging the event inside the probability symbol, we have

$$P\left(\widehat{\beta}_j - t_{n-p,\alpha/2}\sqrt{c_{jj}\widehat{\sigma}^2} < \beta_j < \widehat{\beta}_j + t_{n-p,\alpha/2}\sqrt{c_{jj}\widehat{\sigma}^2}\right) = 1 - \alpha.$$

This shows that

$$\widehat{\beta}_j \pm t_{n-p,\alpha/2}\sqrt{c_{jj}\widehat{\sigma}^2}.$$

is a $100(1-\alpha)$ **percent confidence interval** for β_j.

HYPOTHESIS TESTS: Suppose that we want to test

$$H_0 : \beta_j = \beta_{j,0}$$

versus

$$H_a : \beta_j \neq \beta_{j,0},$$

where $\beta_{j,0}$ is a fixed value (often, $\beta_{j,0} = 0$). We use

$$t = \frac{\widehat{\beta}_j - \beta_{j,0}}{\sqrt{c_{jj}\widehat{\sigma}^2}}$$

as a test statistic and

$$\text{RR} = \{t : |t| > t_{n-p,\alpha/2}\}$$

as a level α rejection region. One sided tests would use a suitably-adjusted rejection region. Probability values are computed as areas under the $t(n-p)$ distribution.

11.4.8 Confidence intervals for $E(Y|x^*)$

RECALL: In the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, we learned how to obtain confidence intervals for the mean response $E(Y|x^*) = \beta_0 + \beta_1 x^*$. Extending this to multiple linear regression models is straightforward.

GOAL: Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, or, equivalently,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Our goal is to construct confidence intervals for linear parametric functions of the form

$$\theta = a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k = \mathbf{a}'\boldsymbol{\beta},$$

where

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

INFERENCE: A point estimator for $\theta = \mathbf{a}'\boldsymbol{\beta}$ is

$$\widehat{\theta} = \mathbf{a}'\widehat{\boldsymbol{\beta}},$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. It is easy to see that $\widehat{\theta}$ is an **unbiased estimator** for θ since

$$E(\widehat{\theta}) = E(\mathbf{a}'\widehat{\boldsymbol{\beta}}) = \mathbf{a}'E(\widehat{\boldsymbol{\beta}}) = \mathbf{a}'\boldsymbol{\beta} = \theta.$$

The variance of $\widehat{\theta}$ is given by

$$V(\widehat{\theta}) = V(\mathbf{a}'\widehat{\boldsymbol{\beta}}) = \mathbf{a}'V(\widehat{\boldsymbol{\beta}})\mathbf{a} = \mathbf{a}'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a} = \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}.$$

Since $\widehat{\theta} = \mathbf{a}'\widehat{\boldsymbol{\beta}}$ is a linear combination of $\widehat{\boldsymbol{\beta}}$, which is normally distributed, $\widehat{\theta}$ is also normally distributed. Therefore, we have shown that

$$\widehat{\theta} \sim \mathcal{N}[\theta, \sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}].$$

Standardizing, we have

$$Z = \frac{\widehat{\theta} - \theta}{\sqrt{\sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}} \sim \mathcal{N}(0, 1).$$

It also follows that

$$t = \frac{\widehat{\theta} - \theta}{\sqrt{\widehat{\sigma}^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}} \sim t(n - p),$$

where $p = k + 1$ and

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n - p}.$$

Since t is a pivotal quantity, a $100(1 - \alpha)$ percent confidence interval for $\theta = \mathbf{a}'\boldsymbol{\beta}$ is

$$\widehat{\theta} \pm t_{n-p,\alpha/2} \sqrt{\widehat{\sigma}^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}.$$

In addition, tests of hypotheses concerning θ use the $t(n - p)$ distribution.

SPECIAL CASE: A special case of the preceding result is estimating the mean value of Y for a fixed value of $\mathbf{x} = (x_1, x_2, ..., x_k)'$, say,

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_k^* \end{pmatrix}.$$

In our multiple linear regression model, we know that

$$E(Y|\mathbf{x}^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_k x_k^*,$$

which is just a linear combination of the form $\theta = a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k = \mathbf{a}'\boldsymbol{\beta}$, where

$$\mathbf{a} = \begin{pmatrix} 1 \\ x_1^* \\ \vdots \\ x_k^* \end{pmatrix}.$$

Therefore,

$$\widehat{\theta} \equiv \widehat{E(Y|\mathbf{x}^*)} = \widehat{\beta}_0 + \widehat{\beta}_1 x^* + \widehat{\beta}_2 x_2^* + \cdots + \widehat{\beta}_k x_k^* = \mathbf{a}'\widehat{\boldsymbol{\beta}}$$

is an unbiased estimator of $\theta = E(Y|\mathbf{x}^*)$, and its variance is

$$V(\widehat{\theta}) = \sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a},$$

where $\mathbf{a}$ is as given above. Applying the preceding general results to this special case, a $100(1-\alpha)$ **percent confidence interval** for $E(Y|\mathbf{x}^*)$, the mean of Y when $\mathbf{x} = \mathbf{x}^*$, is

$$\widehat{\theta} \pm t_{n-p,\alpha/2}\sqrt{\widehat{\sigma}^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}.$$

11.4.9 Prediction intervals for Y^*

RECALL: In the simple linear regression model, we learned how to obtain prediction intervals for a new response Y^*. Extending this to multiple linear regression models is straightforward.

GOAL: Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, or, equivalently,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Suppose that we would like to predict the value of a new response Y^*, for a fixed value of $\mathbf{x} = (x_1, x_2, ..., x_k)'$, say,

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_k^* \end{pmatrix}.$$

Our point predictor for Y^*, based on the least squares fit, is

$$\widehat{Y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \widehat{\beta}_2 x_2^* + \cdots + \widehat{\beta}_k x_k^* = \mathbf{a}'\widehat{\boldsymbol{\beta}},$$

where $\mathbf{a} = (1, x_1^*, x_2^*, ..., x_k^*)'$ and $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Define the error in prediction by $U = Y^* - \widehat{Y}^*$. Analogously to the simple linear regression case,

$$U = Y^* - \widehat{Y}^* \sim \mathcal{N}\{0, \sigma^2[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}]\}.$$

Using the fact that $(n-p)\widehat{\sigma}^2/\sigma^2 \sim \chi^2(n-p)$, it follows that

$$t = \frac{Y^* - \widehat{Y}^*}{\sqrt{\widehat{\sigma}^2\,[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}]}} \sim t(n-p).$$

Therefore,

$$\widehat{Y}^* \pm t_{n-p,\alpha/2}\sqrt{\widehat{\sigma}^2\,[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}]},$$

is a $100(1-\alpha)$ **percent prediction interval** for Y^*.

REMARK: Comparing the prediction interval for Y^* to the analogous $100(1-\alpha)$ percent confidence interval for $E(Y|\mathbf{x}^*)$, we see that the intervals are again identical except the prediction interval has an extra "1" in the estimated standard error. This results from the extra variability that arises when predicting Y^* as opposed to estimating $E(Y|\mathbf{x}^*)$.

11.4.10 Example

Example 11.2. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study from the LaTrobe Valley of Victoria, Australia, samples of cheddar cheese were analyzed for their chemical composition and were subjected to taste tests. For each specimen, the taste Y was obtained by combining the scores from several tasters. Data were collected on the following variables:

$$\begin{aligned} Y &= \text{taste score (TASTE)} \\ x_1 &= \text{concentration of acetic acid (ACETIC)} \\ x_2 &= \text{concentration of hydrogen sulfide (H2S)} \\ x_3 &= \text{concentration of lactic acid (LACTIC).} \end{aligned}$$

Variables ACETIC and H2S were both measured on the log scale. The variable LACTIC has not been transformed. Table 11.2 contains concentrations of the various chemicals in $n = 30$ specimens of cheddar cheese and the observed taste score.

Specimen	TASTE	ACETIC	H2S	LACTIC	Specimen	TASTE	ACETIC	H2S	LACTIC
1	12.3	4.543	3.135	0.86	16	40.9	6.365	9.588	1.74
2	20.9	5.159	5.043	1.53	17	15.9	4.787	3.912	1.16
3	39.0	5.366	5.438	1.57	18	6.4	5.412	4.700	1.49
4	47.9	5.759	7.496	1.81	19	18.0	5.247	6.174	1.63
5	5.6	4.663	3.807	0.99	20	38.9	5.438	9.064	1.99
6	25.9	5.697	7.601	1.09	21	14.0	4.564	4.949	1.15
7	37.3	5.892	8.726	1.29	22	15.2	5.298	5.220	1.33
8	21.9	6.078	7.966	1.78	23	32.0	5.455	9.242	1.44
9	18.1	4.898	3.850	1.29	24	56.7	5.855	10.20	2.01
10	21.0	5.242	4.174	1.58	25	16.8	5.366	3.664	1.31
11	34.9	5.740	6.142	1.68	26	11.6	6.043	3.219	1.46
12	57.2	6.446	7.908	1.90	27	26.5	6.458	6.962	1.72
13	0.7	4.477	2.996	1.06	28	0.7	5.328	3.912	1.25
14	25.9	5.236	4.942	1.30	29	13.4	5.802	6.685	1.08
15	54.9	6.151	6.752	1.52	30	5.5	6.176	4.787	1.25

Table 11.2: Cheese data. ACETIC, H2S, and LACTIC are independent variables. The response variable is TASTE.

REGRESSION MODEL: Suppose the researchers postulate that each of the three chemical composition variables x_1, x_2, and x_3 is important in describing the taste. In this case, they might initially consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. We now use R to fit this model using the method of least squares. Here is the output:

```
> summary(fit)
Call: lm(formula = taste ~ acetic + h2s + lactic)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -28.877     19.735  -1.463  0.15540
acetic         0.328      4.460   0.074  0.94193
h2s            3.912      1.248   3.133  0.00425 **
lactic        19.670      8.629   2.279  0.03109 *

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared: 0.6518,     Adjusted R-squared: 0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.810e-06
```

OUTPUT: The `Estimate` output gives the values of the least squares estimates:

$$\widehat{\beta}_0 \approx -28.877 \qquad \widehat{\beta}_1 \approx 0.328 \qquad \widehat{\beta}_2 \approx 3.912 \qquad \widehat{\beta}_3 \approx 19.670.$$

Therefore, the fitted least squares regression model is

$$\widehat{Y} = -28.877 + 0.328x_1 + 3.912x_2 + 19.670x_3,$$

or, in other words,

$$\widehat{\texttt{TASTE}} = -28.877 + 0.328\texttt{ACETIC} + 3.912\texttt{H2S} + 19.670\texttt{LACTIC}.$$

The `Std.Error` output gives

$$\begin{aligned}
19.735 &= \widehat{\text{se}}(\widehat{\beta}_0) = \sqrt{c_{00}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{00}} \\
4.460 &= \widehat{\text{se}}(\widehat{\beta}_1) = \sqrt{c_{11}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{11}} \\
1.248 &= \widehat{\text{se}}(\widehat{\beta}_2) = \sqrt{c_{22}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{22}} \\
8.629 &= \widehat{\text{se}}(\widehat{\beta}_3) = \sqrt{c_{33}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{33}},
\end{aligned}$$

where

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{30-4} = (10.13)^2 \approx 102.63$$

is the square of the `Residual standard error`. The `t value` output gives the t statistics

$$\begin{aligned}
t = -1.463 &= \frac{\widehat{\beta}_0 - 0}{\sqrt{c_{00}\widehat{\sigma}^2}} \\
t = 0.074 &= \frac{\widehat{\beta}_1 - 0}{\sqrt{c_{11}\widehat{\sigma}^2}} \\
t = 3.133 &= \frac{\widehat{\beta}_2 - 0}{\sqrt{c_{22}\widehat{\sigma}^2}} \\
t = 2.279 &= \frac{\widehat{\beta}_3 - 0}{\sqrt{c_{33}\widehat{\sigma}^2}}.
\end{aligned}$$

These t statistics can be used to test $H_0 : \beta_i = 0$ versus $H_0 : \beta_i \neq 0$, for $i = 0, 1, 2, 3$. Two-sided probability values are in `Pr(>|t|)`. At the $\alpha = 0.05$ level,

- we do not reject $H_0 : \beta_0 = 0$ (p-value $= 0.155$). **Interpretation:** In the model which includes all three independent variables, the intercept term β_0 is not statistically different from zero.

- we do not reject $H_0 : \beta_1 = 0$ (p-value $= 0.942$). **Interpretation:** `ACETIC` does not significantly add to a model that includes `H2S` and `LACTIC`.

- we reject $H_0 : \beta_2 = 0$ (p-value $= 0.004$). **Interpretation:** `H2S` does significantly add to a model that includes `ACETIC` and `LACTIC`.

- we reject $H_0 : \beta_3 = 0$ (p-value $= 0.031$). **Interpretation:** `LACTIC` does significantly add to a model that includes `ACETIC` and `H2S`.

CONFIDENCE INTERVALS: Ninety-five percent confidence intervals for the regression parameters β_0, β_1, β_2, and β_3, respectively, are

$$\begin{aligned}
\widehat{\beta}_0 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_0) &\Longrightarrow -28.877 \pm 2.056(19.735) \Longrightarrow (-69.45, 11.70) \\
\widehat{\beta}_1 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_1) &\Longrightarrow 0.328 \pm 2.056(4.460) \Longrightarrow (-8.84, 9.50) \\
\widehat{\beta}_2 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_2) &\Longrightarrow 3.912 \pm 2.056(1.248) \Longrightarrow (1.35, 6.48) \\
\widehat{\beta}_3 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_3) &\Longrightarrow 19.670 \pm 2.056(8.629) \Longrightarrow (1.93, 37.41).
\end{aligned}$$

PREDICTION: Suppose that we are interested estimating $E(Y|\mathbf{x}^*)$ and predicting a new Y when `ACETIC` $= 5.5$, `H2S` $= 6.0$, and `LACTIC` $= 1.4$, so that

$$\mathbf{x}^* = \begin{pmatrix} 5.5 \\ 6.0 \\ 1.4 \end{pmatrix}.$$

We use R to compute the following:

```
> predict(fit,data.frame(acetic=5.5,h2s=6.0,lactic=1.4),level=0.95,interval="confidence")
     fit      lwr      upr
23.93552 20.04506 27.82597
> predict(fit,data.frame(acetic=5.5,h2s=6.0,lactic=1.4),level=0.95,interval="prediction")
     fit      lwr      upr
23.93552 2.751379 45.11966
```

- Note that

$$\begin{aligned}
\widehat{E(Y|\mathbf{x}^*)} = \widehat{Y}^* &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \widehat{\beta}_2 x_2^* + \widehat{\beta}_3 x_3^* \\
&= -28.877 + 0.328(5.5) + 3.912(6.0) + 19.670(1.4) \approx 23.936.
\end{aligned}$$

- A 95 percent **confidence interval** for $E(Y|\mathbf{x}^*)$ is $(20.05, 27.83)$. When `ACETIC` $= 5.5$, `H2S` $= 6.0$, and `LACTIC` $= 1.4$, we are 95 percent confident that the mean taste rating is between 20.05 and 27.83.

- A 95 percent **prediction interval** for Y^*, when $\mathbf{x} = \mathbf{x}^*$, is $(2.75, 45.12)$. When `ACETIC` $= 5.5$, `H2S` $= 6.0$, and `LACTIC` $= 1.4$, we are 95 percent confident that the taste rating for a new cheese specimen will be between 2.75 and 45.12.

11.5 The analysis of variance for linear regression

IMPORTANCE: The fit of a linear regression model (simple or linear) can be summarized in an **analysis of variance (ANOVA)** table. An ANOVA table provides a partition of the variability in the observed data. This partition, in turn, allows us to assess the overall fit of the model.

MODEL: Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, and let $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ denote the hat matrix. Recall that $\widehat{\mathbf{Y}} = \mathbf{M}\mathbf{Y}$ and $\mathbf{e} = (\mathbf{I} - \mathbf{M})\mathbf{Y}$ denote the vectors of least squares fitted values and residuals, respectively.

SUMS OF SQUARES: Start with the simple quadratic form $\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{I}\mathbf{Y}$. Note that

$$\begin{aligned}
\mathbf{Y}'\mathbf{Y} &= \mathbf{Y}'(\mathbf{M} + \mathbf{I} - \mathbf{M})\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{M}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{M}\mathbf{M}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{M})(\mathbf{I} - \mathbf{M})\mathbf{Y} \\
&= \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} + \mathbf{e}'\mathbf{e}.
\end{aligned}$$

This equation can be expressed equivalently as

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \widehat{Y}_i^2 + \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

TERMINOLOGY: We call

- $\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{n} Y_i^2$ the **uncorrected total** sum of squares
- $\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} = \sum_{i=1}^{n} \widehat{Y}_i^2$ the **uncorrected regression (model)** sum of squares
- $\mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$ the **error (residual)** sum of squares.

CORRECTED VERSIONS: When we fit a linear regression model, we are often interested in the regression coefficients that are attached to independent variables; i.e.,

$\beta_1, \beta_2, ..., \beta_k$. We generally are not interested in the intercept term β_0, the overall mean of Y (ignoring the independent variables). Therefore, it is common to "remove" the effects of fitting the intercept term β_0. This removal is accomplished by subtracting $n\overline{Y}^2$ from both sides of the last equation. This gives

$$\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = \sum_{i=1}^{n} \widehat{Y}_i^2 - n\overline{Y}^2 + \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2,$$

or, equivalently,

$$\underbrace{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}_{\text{SSE}}.$$

We call

- SST the **corrected total** sum of squares
- SSR the **corrected regression (model)** sum of squares
- SSE the **error (residual)** sum of squares.

QUADRATIC FORMS: To enhance our understanding of the partitioning of sums of squares, we express the SST = SSR + SSE partition in terms of quadratic forms. The basic **uncorrected** partition is given by

$$\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{M}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}.$$

To write the corrected partition, we subtract $n\overline{Y}^2 = \mathbf{Y}'n^{-1}\mathbf{J}\mathbf{Y}$ from both sides of the last equation, where

$$\mathbf{J} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}_{n \times n}$$

is the $n \times n$ matrix of ones. This gives

$$\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'n^{-1}\mathbf{J}\mathbf{Y} = \mathbf{Y}'\mathbf{M}\mathbf{Y} - \mathbf{Y}'n^{-1}\mathbf{J}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}$$

or, equivalently,

$$\underbrace{\mathbf{Y}'(\mathbf{I} - n^{-1}\mathbf{J})\mathbf{Y}}_{\text{SST}} = \underbrace{\mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y}}_{\text{SSR}} + \underbrace{\mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}}_{\text{SSE}}.$$

ANOVA TABLE: The general form of an ANOVA table for linear regression (simple or multiple) is given below:

Source	df	SS	MS	F
Regression	$p-1$	SSR	$\text{MSR} = \frac{\text{SSR}}{p-1}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Error	$n-p$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-p}$	
Total	$n-1$	SST		

NOTES:

- The corrected partition SSR + SSE = SST appears in the column labeled "SS" (**sum of squares**).
- The column labeled "df" gives the **degrees of freedom** for each quadratic form. Mathematically,

$$\begin{aligned} p-1 &= r(\mathbf{M}-n^{-1}\mathbf{J}) \\ n-p &= r(\mathbf{I}-\mathbf{M}) \\ n-1 &= r(\mathbf{I}-n^{-1}\mathbf{J}). \end{aligned}$$

  That is, the degrees of freedom are the ranks of the quadratic form matrices in

$$\mathbf{Y}'(\mathbf{I}-n^{-1}\mathbf{J})\mathbf{Y} = \mathbf{Y}'(\mathbf{M}-n^{-1}\mathbf{J})\mathbf{Y} + \mathbf{Y}'(\mathbf{I}-\mathbf{M})\mathbf{Y}.$$

  Note also that the degrees of freedom add down (as the SS do).
- The column labeled "MS" contains the **mean squares**

$$\begin{aligned} \text{MSR} &= \frac{\text{SSR}}{p-1} \\ \text{MSE} &= \frac{\text{SSE}}{n-p}. \end{aligned}$$

  That is, the mean squares are the SS divided by the corresponding degrees of freedom. Note that

$$\widehat{\sigma}^2 = \text{MSE} = \frac{\text{SSE}}{n-p}$$

  is our unbiased estimator of the error variance σ^2 in the underlying model.
- The ANOVA table F statistic will be discussed next.

F STATISTIC: The F statistic in the ANOVA table is used to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

versus

$$H_a : \text{at least one of the } \beta_j \text{ is nonzero.}$$

In other words, F tests whether or not at least one of the independent variables $x_1, x_2, ..., x_k$ is important in describing the response Y. If H_0 is rejected, we do not know which one or how many of the β_j's are nonzero; only that at least one is. In this light, one could argue that this test is not all that meaningful.

JUSTIFICATION: When H_0 is true,

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2(p-1), \quad \frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-p),$$

and SSR and SSE are independent. These facts would be proven in a more advanced course. Therefore, when H_0 is true,

$$F = \frac{\frac{\text{SSR}/\sigma^2}{p-1}}{\frac{\text{SSE}/\sigma^2}{n-p}} = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} = \frac{\text{MSR}}{\text{MSE}} \sim F(p-1, n-p).$$

The test above uses a one-sided, upper tail rejection region. Specifically, a level α rejection region is

$$\text{RR} = \{F : F > F_{p-1,n-p,\alpha}\},$$

where $F_{p-1,n-p,\alpha}$ denotes the upper α quantile of the F distribution with $p-1$ (numerator) and $n-p$ (denominator) degrees of freedom. Probability values are computed as areas to the right of F on the $F(p-1, n-p)$ distribution.

TERMINOLOGY: Since

$$\text{SST} = \text{SSR} + \text{SSE},$$

the proportion of the total variation in the data explained by the model is

$$R^2 = \frac{\text{SSR}}{\text{SST}}.$$

The statistic R^2 is called the **coefficient of determination**. The larger the R^2, the more variation that is being explained by the regression model.

Example 11.2 (continued). In Example 11.2, we fit the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. The ANOVA table, obtained using SAS, is shown below.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4994.50861	1664.83620	16.22	<.0001
Error	26	2668.37806	102.62993		
Corrected Total	29	7662.88667			

The F statistic is used to test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

versus

$$H_a : \text{at least one of the } \beta_j \text{ is nonzero.}$$

ANALYSIS: Based on the F statistic ($F = 16.22$), and the corresponding probability value (p-value < 0.0001), we conclude that at least one of `ACETIC`, `H2S`, and `LACTIC` is important in describing taste (that is, we reject H_0). The coefficient of determination is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{4994.51}{7662.89} \approx 0.652.$$

That is, about 65.2 percent of the variability in the taste data is explained by the independent variables. If we analyze these data using R, we get the following:

```
anova.fit<-anova(lm(taste~acetic+h2s+lactic))
anova.fit
Response: taste
           Df   Sum Sq  Mean Sq  F value     Pr(>F)
acetic      1  2314.14  2314.14  22.5484  6.528e-05 ***
h2s         1  2147.11  2147.11  20.9209  0.0001035 ***
lactic      1   533.26   533.26   5.1959  0.0310870 *
Residuals  26  2668.38   102.63
```

NOTE: The convention used by R is to "split up" the (corrected) regression sum of squares

$$\text{SSR} = 4994.50861$$

into sums of squares for each of the three independent variables `ACETIC`, `H2S`, and `LACTIC`, as they are added sequentially to the model (these are called **sequential sums of squares**). The sequential sums of squares for the independent variables add to the SSR (up to rounding error) for the model, that is,

$$\begin{aligned} \text{SSR} = 4994.51 &= 2314.14 + 2147.11 + 533.26 \\ &= \text{SS}(\texttt{ACETIC}) + \text{SS}(\texttt{H2S}) + \text{SS}(\texttt{LACTIC}). \end{aligned}$$

In words,

- SS(`ACETIC`) is the sum of squares added when compared to a model that includes only an intercept term.
- SS(`H2S`) is the sum of squares added when compared to a model that includes an intercept term and `ACETIC`.
- SS(`LACTIC`) is the sum of squares added when compared to a model that includes an intercept term, `ACETIC`, and `H2S`.

11.6 Reduced versus full model testing

SETTING: Consider the (full) multiple regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, or, equivalently,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. We now consider the question of whether or not a smaller model is adequate for the data. That is, can we remove some of the independent variables and write a smaller model that does just as well at describing the data as the full model?

REMARK: Besides their ease of interpretation, smaller models confer statistical benefits. Remember that for each additional independent variable we add to the model, there is an associated regression parameter that has to be estimated. For each additional regression parameter that we have to estimate, we lose a degree of freedom for error. Remember that MSE, our estimator for the error variance σ^2 uses the degrees of freedom for error in its computation. Thus, the fewer error degrees of freedom we have, the less precise estimate we have of σ^2. With an imprecise estimate of σ^2, hypothesis tests, confidence intervals, and prediction intervals are less informative.

TERMINOLOGY: We call

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_g x_{ig} + \beta_{g+1} x_{i(g+1)} + \cdots + \beta_k x_{ik} + \epsilon_i$$

the **full model** because it includes all of the independent variables $x_1, x_2, ..., x_k$. We call

$$Y_i = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_g x_{ig} + \epsilon_i$$

a **reduced model** because it includes only the independent variables $x_1, x_2, ..., x_g$, where $g < k$, that is, independent variables $x_{g+1}, x_{g+2}, ..., x_k$ are not included in the reduced model.

MATRIX NOTATION: In matrix notation, the full model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1g} & x_{1(g+1)} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2g} & x_{2(g+1)} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{ng} & x_{n(g+1)} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_g \\ \beta_{g+1} \\ \vdots \\ \beta_k \end{pmatrix}.$$

In matrix notation, the reduced model is

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where

$$\mathbf{X}_0 = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1g} \\ 1 & x_{21} & x_{22} & \cdots & x_{2g} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{ng} \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_g \end{pmatrix}.$$

That is, the matrix $\mathbf{X}_0$ is simply $\mathbf{X}$ with the last $(k-g)$ columns removed.

TESTING PROBLEM: In order to determine whether or not the extra independent variables $x_{g+1}, x_{g+2}, ..., x_k$ should be included in the regression, we are interested in testing the reduced model versus the full model, that is,

$$H_0 : \mathbf{Y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

versus

$$H_a : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

In terms of the regression parameters in the full model, we are essentially testing

$$H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$$

versus

$$H_a : \text{not } H_0.$$

INTUITION: Define the hat matrices for the reduced and full models by $\mathbf{M}_0 = \mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'$ and $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, respectively. We know that

$$\begin{aligned} \text{SSR}_F &= \mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y} \\ \text{SSR}_R &= \mathbf{Y}'(\mathbf{M}_0 - n^{-1}\mathbf{J})\mathbf{Y} \end{aligned}$$

are the (corrected) regression sum of squares for the full and reduced models, respectively. Since the regression sum of squares SSR can never decrease by adding independent variables, it follows that

$$\text{SSR}_F = \mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y} \geq \mathbf{Y}'(\mathbf{M}_0 - n^{-1}\mathbf{J})\mathbf{Y} = \text{SSR}_R.$$

In the light of this, our intuition should suggest the following:

- If $\text{SSR}_F = \mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y}$ and $\text{SSR}_R = \mathbf{Y}'(\mathbf{M}_0 - n^{-1}\mathbf{J})\mathbf{Y}$ are "close," then the additional independent variables $x_{g+1}, x_{g+2}, ..., x_k$ do not add too much to the regression, and the reduced model is adequate at describing the data.

- if $\text{SSR}_F = \mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y}$ and $\text{SSR}_R = \mathbf{Y}'(\mathbf{M}_0 - n^{-1}\mathbf{J})\mathbf{Y}$ are not "close," then the additional independent variables $x_{g+1}, x_{g+2}, ..., x_k$ add a significant amount to the regression. This suggests that the reduced model does an insufficient job of describing the data when compared to the full model.

- We therefore make our decision by examining the size of

$$\text{SSR}_F - \text{SSR}_R = \mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y} - \mathbf{Y}'(\mathbf{M}_0 - n^{-1}\mathbf{J})\mathbf{Y} = \mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}.$$

  If this difference is "large," then the reduced model does not do a good job of describing the data (when compared to the full model).

- We are assuming that the full model already does a good job of describing the data; we are trying to find a smaller model that does just as well.

TEST STATISTIC: Theoretical arguments in linear models show that when the reduced model is correct,

$$F = \frac{\mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}/(k-g)}{\text{MSE}_F} \sim F(k-g, n-p),$$

where $p = k + 1$ and MSE_F is the mean squared error computed from the full model. Therefore, a level α rejection region for testing

$$H_0 : \mathbf{Y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

versus

$$H_a : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

is given by

$$\text{RR} = \{F : F > F_{k-g,n-p,\alpha}\},$$

where $F_{k-g,n-p,\alpha}$ is the upper α quantile of the $F(k-g, n-p)$ distribution.

Example 11.2 (continued). In Example 11.2, consider the full model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

Suppose we believe that a simple linear regression model with `ACETIC` (x_1) only does just as well as the full model at describing `TASTE`. In this case, the reduced model is

$$Y_i = \gamma_0 + \gamma_1 x_{i1} + \epsilon_i.$$

IMPLEMENTATION: To test the reduced model versus the full model, we first compute the ANOVA tables from both model fits. The ANOVA table from the full model fit (using SAS) is

Analysis of Variance: Full Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4994.50861	1664.83620	16.22	<.0001
Error	26	2668.37806	102.62993		
Corrected Total	29	7662.88667			

The ANOVA table from the reduced model fit (using SAS) is

Analysis of Variance: Reduced Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2314.14151	2314.14151	12.11	0.0017
Error	28	5348.74515	191.02661		
Corrected Total	29	7662.88667			

Therefore, the difference in the (corrected) regression sum of squares is

$$\begin{aligned} \mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y} &= \text{SSR}_F - \text{SSR}_R \\ &= 4994.50861 - 2314.14151 = 2680.367 \end{aligned}$$

and the test statistic is

$$F = \frac{\mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}/(k-g)}{\text{MSE}_F} = \frac{2680.367/(3-1)}{102.62993} \approx 13.058.$$

A level $\alpha = 0.05$ rejection region is

$$\text{RR} = \{F : F > F_{2,26,0.05} = 3.369\}.$$

I used the R command `qf(0.95,2,26)` to compute $F_{2,26,0.05}$. Because the test statistic F falls in the rejection region, we reject H_0 at the $\alpha = 0.05$ level. We conclude that the reduced model does not do as well as the full model in describing `TASTE`. The probability value for the test is

$$\text{p-value} = P(F_{2,26} > 13.058) \approx 0.0001,$$

computed using the `1-pf(13.058,2,26)` in R.
IMPORTANT: It is interesting to note that the sum of squares

$$\begin{aligned} 2680.367 &= \mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y} \\ &= \text{SS}(\texttt{H2S}) + \text{SS}(\texttt{LACTIC}) = 2147.11 + 533.26. \end{aligned}$$

That is, we can obtain $\mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}$ by adding the sequential sum of squares corresponding to the independent variables not in the reduced model.
REMARK: It is possible to implement this test completely in R. Here is the output:

```
> fit.full<-lm(taste~acetic+h2s+lactic)
> fit.reduced<-lm(taste~acetic)
> anova(fit.reduced,fit.full,test="F")
Model 1: taste ~ acetic
Model 2: taste ~ acetic + h2s + lactic
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     28 5348.7
2     26 2668.4  2    2680.4 13.058 0.0001186 ***
```

ANALYSIS: R's convention is to produce the F statistic

$$F = \frac{\mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}/(k-g)}{\text{MSE}_F} = \frac{2680.367/(3-1)}{102.62993} \approx 13.058$$

automatically with the corresponding p-value in `Pr(>F)`.