Chapter 2
Inferences in Regression and Correlation Analysis

許湘伶

Applied Linear Regression Models
(Kutner, Nachtsheim, Neter, Li)
Inferences concerning:

- the regression parameters $\beta_0$ and $\beta_1$
- interval estimation of $\beta_0$ and $\beta_1$
- tests about them

✓ interval estimation of $E(Y)$ of the probability distribution of $Y$ for given $X$
✓ prediction intervals of a new observation $Y$
✓ confidence bands for the regression line
the analysis of variance approach to regression analysis
the general linear test approach
descriptive measure of association
the correlation coefficient
Assume that the normal error regression model is applicable:

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

- \( \beta_0 \) and \( \beta_1 \) are parameters
- \( X_i \) are known constants
- \( \varepsilon_i \sim N(0, \sigma^2) \): are independent
Inferences about the slope $\beta_1$ of the regression line

**illustration**

A market research analyst studying the relation between sales ($Y$) and advertising expenditures (廣告支出 $X$):

- to obtain an interval estimate of $\beta_1$
- provide information as to how many additional sales dollars

\[
H_0 : \beta_1 = 0; \\
H_a : \beta_1 \neq 0.
\]
Inferences about the slope $\beta_1$ of the regression line (cont.)

Figure 1: Regression Model when $\beta_1 = 0$.

- $\beta_1 = 0 \Rightarrow$ no linear association between $Y$ and $X$
- The regression line is horizontal.
- The means of $Y$: $E\{Y\} = \beta_0$.
- The probability distribution of $Y$ are identical at all levels of $X$. 
Point estimator $b_1$

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

The sample distribution of $b_1$ refers to the different values of $b_1$ that would be obtained with repeated sampling when the levels of the predictor variable $X$ are held constant from sample to sample.
Sampling distribution of $b_1$ (需$k_i$的特性)

For normal error regression model:

Mean: $E\{b_1\} = \beta_1$;

variance: $\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$

✓ $b_1$ is a linear combination of $Y_i$

$$b_1 = \sum k_i Y_i$$

$$k_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2} : \text{a function of } X_i$$
✓ Properties of $k_i$:

\[ \sum k_i = 0 \]
\[ \sum k_i X_i = 1 \]
\[ \sum k_i^2 = \frac{1}{\sum(X_i - \bar{X})^2} \]
Normality

- $Y_i$: independently, normally distributed
- A linear combination of independent normal random variables is normally distributed.
Sampling Distribution of $b_1$

Normality (cont.)

Normally properties

- **Mean:**
  \[ E\{ b_1 \} = \beta_1 \]

- **Variance:**
  \[ \sigma^2 \{ b_1 \} = \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2} \]
Normality

Normally properties (cont.)

- The unbiased estimator of $\sigma^2$:

\[
\text{MSE} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}
\]

Estimated Variance $\sigma^2 \{b_1\}$:

\[
s^2 \{b_1\} = \frac{\text{MSE}}{\sum (X_i - \bar{X})^2}
\]

The point estimator $s^2 \{b_1\}$ is an unbiased estimator of $\sigma^2 \{b_1\}$. 
Normality

Normally properties (cont.)

- The unbiased estimator of $\sigma^2$:

$$MSE = \frac{SSE}{n - 2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2}$$
Normality

Normally properties (cont.)

- The unbiased estimator of $\sigma^2$: 
  $$MSE = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$$

- Estimated Variance $\sigma^2\{b_1\}$:
  $$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2}$$

- The point estimator $s^2\{b_1\}$ is an unbiased estimator of $\sigma^2\{b_1\}$.
- The point estimator of $\sigma\{b_1\}$:
Sampling Distribution of $b_1$

**Normality**

**Normally properties (cont.)**

- The unbiased estimator of $\sigma^2$: 
  
  $$MSE = \frac{SSE}{n - 2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - 2}$$

- Estimated Variance $\sigma^2\{b_1\}$:
  
  $$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2}$$

- The point estimator $s^2\{b_1\}$ is an unbiased estimator of $\sigma^2\{b_1\}$.

- The point estimator of $\sigma\{b_1\}$: $s\{b_1\}$
**Theorem 1**

The estimator $b_1$ has minimum variance among all unbiased linear estimators of:

$$\hat{\beta}_1 = \sum c_i Y_i$$

- $c_i$: arbitrary constants; $\sum c_i = 0$; $\sum c_i X_i = 1$

- **Unbiased:** $E\{\hat{\beta}_1\} = \beta_1$

- **Variance:** $\sigma^2\{\hat{\beta}_1\} = \sigma^2 \sum c_i^2$
  - $c_i = k_i + d_i$
  - $k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}$
  - $d_i$: arbitrary constants
  - $\sum k_i d_i = 0$
  - $\sigma^2 \sum k_i^2 = \sigma^2\{b_1\}$

- $\hat{\beta}_1$ is at a minimum when $\sum d_i^2 = 0 \iff$ all $d_i = 0 \iff c_i = k_i$
Theorem 2

\[
\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}
\]

If the observations \( Y_i \) come from the same normal population, \( (\bar{Y} - \mu)/s\{\bar{Y}\} \overset{d}{\sim} \)
Sampling distribution of \((b_1 - \beta_1)/s\{b_1\}\)

**Theorem 2**

\[
\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}
\]

- If the observations \(Y_i\) come from the same normal population, \((\bar{Y} - \mu)/s\{\bar{Y}\} \sim t_{n-1}\)
Sampling distribution of \((b_1 - \beta_1)/s\{b_1\}\)

**Theorem 2**

\[
\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}
\]

- If the observations \(Y_i\) come from the same normal population, \((\bar{Y} - \mu)/s\{\bar{Y}\} \overset{d}{\sim} t_{n-1}\)
- Two parameters: \(\beta_0, \beta_1\)
Sampling distribution of \( (b_1 - \beta_1)/s\{b_1\} \) (cont.)

Proof of \( \frac{b_1-\beta_1}{s\{b_1\}} \sim t_{n-2} \)

**Theorem 3**

For regression model (2.1),

\[ \frac{SSE}{\sigma^2} \sim \chi^2(n - 2), \]

**and is independent of** \( b_0 \) **and** \( b_1 \).
Sampling distribution of \((b_1 - \beta_1)/s\{b_1\}\) (cont.) (cont.)

**Theorem 4**

**t Distribution**

Let \(Z\) and \(\chi^2(\nu)\) be independent r.v. (standard normal, \(\chi^2\)). A \(t\) random variable as follows:

\[
t(\nu) = \frac{Z}{\left[\frac{\chi^2(\nu)}{\nu}\right]} \quad \text{where } Z \text{ and } \chi^2(\nu) \text{ are indep.}
\]
Sampling distribution of \((b_1 - \beta_1)/s\{b_1\}\) (cont.) (cont.)

1. \((b_1 - \beta_1)/\sigma\{b_1\} \sim Z\) (Standard Normal variable)

2. \(\frac{s^2\{b_1\}}{\sigma^2\{b_1\}} \sim \frac{\chi^2(n-2)}{n-2}\)

3. 

\[
\frac{b_1 - \beta_1}{s\{b_1\}} = \frac{b_1 - \beta_1}{\sigma\{b_1\}} \div \frac{s\{b_1\}}{\sigma\{b_1\}} \sim \frac{Z}{\sqrt{\frac{\chi^2(n-2)}{n-2}}}
\]

\(Z\) and \(\chi^2\) are independent;

4. \(Z\) is a function of \(b_1\)

5. \(b_1\) is independent of \(SSE/\sigma^2 \sim \chi^2\)

\[
\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}
\]
Confidence interval for $\beta_1$

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}$$

$$P\left\{ t(\alpha/2; n-2) \leq (b_1 - \beta_1)/s\{b_1\} \leq t(1 - \alpha/2; n-2) \right\} = 1 - \alpha$$

$$\Rightarrow P\left\{ b_1 - t(1 - \alpha/2; n-2)s\{b_1\} \leq \beta_1 \leq b_1 + t(1 - \alpha/2; n-2)s\{b_1\} \right\}$$

$$= 1 - \alpha$$

- $t(\alpha/2; n-2)$: $(\alpha/2)100$ percentile of the $t$ distribution with $n - 2$ d.f.
Confidence interval for $\beta_1$ (cont.)

- Symmetric: $t(\alpha/2; n - 2) = -t(1 - \alpha/2; n - 2)$

The $1 - \alpha$ confidence limits for $\beta_1$ are:

$$b_1 \pm t(1 - \alpha/2; n - 2)s\{b_1\}$$
Confidence interval for $\beta_1$ (cont.)
Ex: Confidence interval for $\beta_1$

The Toluca Company: an estimate of $\beta_1$ with 95 percent confidence coefficient.

```r
### Example p46
toluca<-read.table("toluca.txt",header=T)
attach(toluca)

## method 1
n<-length(Size)
alpha<-0.05
tdf<-qt(1-alpha/2,n-2)
Sxx<-sum((Size-mean(Size))^2)
b1<-sum((Size-mean(Size))*(Hrs-mean(Hrs)))/Sxx
b0<-mean(Hrs)-b1*mean(Size)
c(b0,b1)
SSE<-sum((Hrs-(b0+b1*Size))^2)
MSE<-SSE/(n-2)
```

Ex: Confidence interval for $\beta_1$ (cont.)

```
sb1<-sqrt(MSE/sum((Size-mean(Size))^2)) #s{b1}
conf<-c(b1-tdf*sb1,b1+tdf*sb1)
conf
```

```r
## method 2
fitreg<-lm(Hrs~Size)
summary(fitreg)
confint(fitreg)
```

- $s^2 \{b_1\} = \frac{MSE}{\sum (X_i - \bar{X})^2} = 0.120404$
- $s \{b_1\} = 0.3470$
- $t(0.975, 23) = 2.069$
- The 95 percent confidence interval:

\[
2.85 \leq \beta_1 \leq 4.29
\]
Confidence interval for $\beta_1$

**Ex: Confidence interval for $\beta_1$ (cont.)**

TABLE 2.1

<table>
<thead>
<tr>
<th>Results for Toluca Company Example Obtained in Chapter 1.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 25$</td>
</tr>
<tr>
<td>$b_0 = 62.37$</td>
</tr>
<tr>
<td>$\hat{Y} = 62.37 + 3.5702X$</td>
</tr>
<tr>
<td>$\sum(X_i - \bar{X})^2 = 19,800$</td>
</tr>
<tr>
<td>$\sum(Y_i - \bar{Y})^2 = 307,203$</td>
</tr>
</tbody>
</table>

FIGURE 2.2

The regression equation is

$Y = 62.4 + 3.57 \times X$

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>62.37</td>
<td>26.18</td>
<td>2.38</td>
<td>0.026</td>
</tr>
<tr>
<td>X</td>
<td>3.5702</td>
<td>0.3470</td>
<td>10.29</td>
<td>0.000</td>
</tr>
</tbody>
</table>

$s = 48.82$  \hspace{1cm} R-sq = 82.2\%  \hspace{1cm} R-sq(adj) = 81.4\%$

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>252378</td>
<td>252378</td>
<td>105.88</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>23</td>
<td>54825</td>
<td>2384</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>307203</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tests concerning $\beta_1$

\[
\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}
\]

Two-Sided Test

\[H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0\]

The test statistic:

\[t^* = \frac{b_1}{s\{b_1\}}\]

The decision rule (the level of significance $\alpha$):

- If $|t^*| \leq t(1 - \alpha/2; n - 2)$, conclude $H_0$
- If $|t^*| > t(1 - \alpha/2; n - 2)$, conclude $H_a$
Tests concerning $\beta_1$

- $\alpha = 0.05$; $n = 25$
- $t(0.975; 23) = 2.069$
- $b_1 = 3.5702$
- $|t^*| = |3.5702/0.3470| = 10.29 > 2.069$

**conclude $H_a$:** $\beta_1 \neq 0$

- **p-value:** $P\{t(23) > t^* = 10.29\} < 0.0005$
Tests concerning $\beta_1$

- $\alpha = 0.05$; $n = 25$
- $t(0.975; 23) = 2.069$
- $b_1 = 3.5702$
- $|t^*| = \frac{|3.5702/0.3470|}{10.29 > 2.069}$

Conclude $H_a: \beta_1 \neq 0$

- p-value: $P\left\{ t(23) > t^* = 10.29 \right\} < 0.0005$
Confidence interval for $\beta_1$

**Tests concerning $\beta_1$**

**One-Sided Test**

$H_0 : \beta_1 \leq 0$ vs. $H_a : \beta_1 > 0$

The test statistic:

$$t^* = \frac{b_1}{s\{b_1\}}$$

The decision rule (the level of significance $\alpha$):

If $t^* \leq t(1 - \alpha; n - 2)$, conclude $H_0$

If $t^* > t(1 - \alpha; n - 2)$, conclude $H_a$
Tests concerning $\beta_1$

Two-Sided Test

$H_0 : \beta_1 = \beta_{10}$ vs. $H_a : \beta_1 \neq \beta_{10}$

The test statistic:

$$t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}}$$

The decision rule (the level of significance $\alpha$):

If $|t^*| \leq t(1 - \alpha/2; n - 2)$, conclude $H_0$

If $|t^*| > t(1 - \alpha/2; n - 2)$, conclude $H_a$
Inferences concerning $\beta_0$

**Sampling distribution of $b_0$**

- The point estimator $b_0$: $b_0 = \bar{Y} - b_1 \bar{X}$

**Theorem 5**

*For regression model (2.1), the sampling distribution of $b_0$ is normal, with mean and variance:*

\[
E\{b_0\} = \beta_0
\]

\[
\sigma^2\{b_0\} = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2} \right\}
\]
Sampling distribution of $b_0$ (cont.)

- An estimator of $\sigma^2\{b_0\}$:
  \[
s^2\{b_0\} = MSE \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2} \right]
  \]

- An estimator of $\sigma\{b_0\}$: $s\{b_0\}$

Theorem 6

\[
\frac{b_0 - \beta_0}{s\{b_0\}} \sim t_{n-2}
\]
Confidence interval for $\beta_0$

The $1 - \alpha$ confidence limits for $\beta_0$ are:

$$b_0 \pm t(1 - \alpha/2; n - 2)s\{b_0\}$$

## method 1

```r
sb0<-sqrt(MSE*(1/n+mean(Size)^2/sum((Size-mean(Size))^2)))
b0<-mean(Hrs)-b1*mean(Size)
tdf0<-qt(1-0.1/2,n-2)
conf0<-c(b0-tdf0*sb0,b0+tdf0*sb0)
conf0
```

## method 2

```r
fitreg<-lm(Hrs~Size)
confint(fitreg,level = 0.9)
```
Example C.I. for $\beta_0$

Toluca Company

- Range of $X$: $[20, 120]$
- $t(0.95; 23) = 1.714$
- $s^2\{b_0\} = MSE \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2} \right] = 685.34$
- $s\{b_0\} = 26.18$
- The 90% C.I. for $\beta_0$:

$$17.5 \leq \beta_0 \leq 107.2$$
Example C.I. for $\beta_0$ (cont.)

- It does not necessarily provide information about the "setup" cost since we are not certain whether a linear regression model is appropriate when the scope of the model is extended to $X = 0$. 
Departures from Normality

- If the probability distribution of $Y$ are not exactly normal but do not depart seriously, the sampling distributions of $b_0$ and $b_1$ will approximately normal, and the use of the $t$ distribution will provide approximately the specified confidence coefficient or level of significance.
Departures from Normality (cont.)

- Even if the distribution of $Y$ are far from normal, the estimators $b_0$ and $b_1$ generally have the property of *asymptotic normality*—their distributions approach normality under very general conditions as the sample size increases.

**Skewness**

The coefficient of skewness is a measure for the degree of symmetry in the variable distribution.

- Negatively skewed distribution
  - Skewed to the left
  - Skewness $< 0$

- Normal distribution
  - Symmetrical
  - Skewness $= 0$

- Positively skewed distribution
  - Skewed to the right
  - Skewness $> 0$

**Kurtosis**

The coefficient of kurtosis is a measure for the degree of peakedness/flatness in the variable distribution.

- Platykurtic distribution
  - Low degree of peakedness
  - Kurtosis $< 0$

- Normal distribution
  - Mesokurtic distribution
  - Kurtosis $= 0$

- Leptokurtic distribution
  - High degree of peakedness
  - Kurtosis $> 0$
Power of Tests

\[ H_0 : \beta_1 = \beta_{10} \text{ vs. } H_a : \beta_1 \neq \beta_{10} \]

Test statistic: \[ t^* = \frac{\hat{\beta}_1 - \beta_{10}}{s\{\hat{\beta}_1\}} \]

The power of this test for \( \alpha \) level: the decision rule will lead to conclusion \( H_a \) when \( H_a \) holds

\[ \text{Power} = P\left\{|t^*| > t(1 - \alpha/2; n - 2)|\delta\right\} \]
Power of Tests (cont.)

- the noncentrality measure i.e., a measure of how far the true value of $\beta_1$ is from $\beta_{10}$

$$\delta = \frac{|\beta_1 - \beta_{10}|}{\sigma\{b_1\}}$$

(Appendix Table B.5)
Some considerations on making inferences concerning $\beta_0$ and $\beta_1$

Common objective

To estimate the mean for one or more probability distributions of $Y$.

Illustration

A study of the relation between level of piecework (按件計酬的工作) pay ($X$) and worker productivity (生産力 $Y$).

- The mean productivity at high and medium levels of piecework pay may be of particular interest for purposes of analyzing the benefits obtained from an increase in the pay
Some considerations on making inferences concerning $\beta_0$ and $\beta_1$

- $X_h$: the level of $X$ for which we wish to estimate the mean response
  - may be a value which occurred in the sample
  - other value of the predictor variable within the scope (範圍) of the model
- $E\{Y_h\}$: the mean response when $X = X_h$
- $\hat{Y}_h$: the point estimator of $E\{Y_h\}$:

$$\hat{Y}_h = b_0 + b_1 X_h$$

What is the sampling distribution of $\hat{Y}_h$?
Sampling distribution of $\hat{Y}_h$

For normal error regression model (2,1), the sampling distribution of $\hat{Y}_h$ is normal:

- **Mean:**
  \[
  E\{\hat{Y}_h\} = E\{Y_h\}
  \]

- **Variance:**
  \[
  \sigma^2\{\hat{Y}_h\} = \sigma^2 \left[ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]
  \]

- $\hat{Y}_h$ is a linear combination of the observations $Y_i$.
- $\hat{Y}_h$ is an unbiased estimator of $E\{Y_h\}$.
Some considerations on making inferences concerning $\beta_0$ and $\beta_1$

Sampling distribution of $\hat{Y}_h$ (cont.)

Figure 2: Effect on $\hat{Y}_h$ of Variation in $b_1$ from Sample to Sample in Two Samples with Same Means $\bar{Y}$ and $\bar{X}$
Some considerations on making inferences concerning $\beta_0$ and $\beta_1$

Properties of $\hat{Y}_h$

- $X_h = 0$:
  - $\text{Var}(\hat{Y}_h) = \text{Var}(b_0)$
  - $s^2\{\hat{Y}_h\} = s^2\{b_0\}$

- $b_1$ and $\bar{Y}$ are uncorrelated $\iff \sigma\{\bar{Y}, b_1\} = 0$
Some considerations on making inferences concerning $\beta_0$ and $\beta_1$

Properties of $\hat{Y}_h$ (cont.)

- When $MSE$ is substituted for $\sigma^2$, the estimated variance of $\hat{Y}_h$ ($s^2(\hat{Y}_h)$)

\[
s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]
\]

- The estimated standard deviation of $\hat{Y}_h$: $s\{\hat{Y}_h\}$
Sampling Distribution of \( (\hat{Y}_h - E\{ Y_h \}) / s\{ \hat{Y}_h \} \)

**Theorem 7**

\[
\frac{\hat{Y}_h - E\{ Y_h \}}{s\{ \hat{Y}_h \}} \sim t_{n-2}
\]

The 1 \(-\alpha\) confidence limits for \( E\{ Y_h \} \) are:

\[
\hat{Y}_h \pm t(1 - \alpha/2; n - 2)s\{ \hat{Y}_h \}
\]
Example 1

- Toluca Company example
- find a 90 percent confidence interval for $E\{Y_h\}$ when the lot size is $X_h = 65$ units.
  - $\hat{Y}_h = 62.37 + 3.5702(65) = 294.4$
  - $s^2\{\hat{Y}_h\} = 2,384 \left[ \frac{1}{25} + \frac{(65-70.00)^2}{19,800} \right] = 98.37 \Rightarrow s\{\hat{Y}_h\} = 9.918$
  - $t(.95; 23) = 1.7141$

$$277.4 \leq E\{Y_h\} \leq 311.4$$

- Conclude: the mean number of work hours required when lots of 65 units are produced is somewhere between 277.4 and 311.4 hours.
- the estimate of the mean number of work hours is moderately precise.
Interval Estimation of $E\{Y_h\}$
Code for Example 1

```r
### Example p54
toluca<-read.table("toluca.txt",header=T)
attract(toluca)

## method 1
n<-length(Size)
alpha<-0.1
tdf90<-qt(1-alpha/2,n-2)
Sxx<-sum((Size-mean(Size))^2)
b1<-sum((Size-mean(Size))*(Hrs-mean(Hrs)))/Sxx
b0<-mean(Hrs)-b1*mean(Size)
c(b0,b1)
SSE<-sum((Hrs-(b0+b1*Size))^2)
MSE<-SSE/(n-2)
sb1<-sqrt(MSE/sum((Size-mean(Size))^2)) #s{b1}
Xh<-65
hYh<-b0+b1*Xh
sYh<-sqrt(MSE*(1/n+(Xh-mean(Size))^2)/sum((Size-mean(Size))^2)))
```
Code for Example 1 (cont.)

```
EYhCI<-c(hYh-tdf90*sYh,hYh+tdf90*sYh)

## method 2
fitreg<-lm(Hrs~Size,data=toluca)
summary(fitreg)
predXCI<-predict(fitreg,data.frame(Size = c(Xh)),
    interval = "confidence", se.fit = F,level = 0.9)
plot(Size, Hrs)
abline(fitreg,col="blue")

# now the confidence interval for $X_h=specific level$
points(Xh, predXCI[, "fit"],col="red",pch=15)
points(Xh, predXCI[, "lwr"], lty = "dotted",col="red")
points(Xh, predXCI[, "upr"], lty = "dotted",col="red")
```
Example 2

- Toluca Company example
- Estimate $E\{ Y_h \}$ for lots with $X_h = 100$ units with a 90 percent confidence interval
  - $\hat{Y}_h = 62.37 + 3.5702(100) = 419.4$
  - $s^2\{ \hat{Y}_h \} = 2,384 \left[ \frac{1}{25} + \frac{(100-70.00)^2}{19,800} \right] = 203.72$
  - $s\{ \hat{Y}_h \} = 14.27$
  - $t(.95; 23) = 1.7141$

$$394.9 \leq E\{ Y_h \} \leq 443.9$$

- Conclude: the confidence interval is somewhat wider than that for Example 1, since the $X_h$ level here is substantially farther from the mean $\bar{X} = 70$ than the $X_h = 65$. 
Sampling distribution of $\hat{Y}_h$

- The variance of $\hat{Y}_h$ is smallest when $X_h = \bar{X}$.
- In an experiment to estimate the mean response at a particular level $X_h$ of the predictor variable, the precision of the estimate will be greatest if the observations on $X$ are spaced so that $\bar{X} = X_h$.
- The usual relationship between C.I. and tests applies in inferences concerning the mean response.
- The two-sided confidence limits can be utilized for two-sided tests concerning the mean response at $X_h$. Alternatively, a regular decision rule can be set up.
- The confidence limits for a mean response $E\{ Y_h \}$ are not sensitive to moderate departures from the assumption that the error terms are normally distributed.
Toluca Company: the next lot to be produced consists of 100 units; wishes to predict the number of work hours for this particular lot.

Estimated the regression relation between company sales and number of persons 16 or more years old from data for the past 10 years; wishes to predict next year’s company sales.
College admissions example

- the relevant parameters of the regression model are known:

\[ \beta_0 = 0.10 \quad \beta_1 = 0.95 \]

\[ E\{Y\} = 0.10 + 0.95X \]

\[ \sigma = 0.12 \]

- An applicant whose high school GPA is \( X_h = 3.5 \):

\[ E\{Y_h\} = 0.10 + 0.95(3.5) = 3.425 \]

- \( E\{Y_h\} \pm 3\sigma: \)

\[ 3.425 \pm 3(0.12) \Rightarrow 3.065 \leq Y_{h(new)} \leq 3.785 \]
The basic idea of a prediction interval is to choose a range in the distribution of $Y$ wherein most of the observations will fall, and then to declare that the next observation will fall in this range.

When the regression parameters of normal error regression model (2.1) are known, the $1 - \alpha$ prediction limits for $Y_{h(new)}$ are:

$$E\{Y_h\} \pm z(1 - \alpha/2)\sigma$$
Prediction Interval for $Y_{h(new)}$ when Parameters Unknown

Figure 3: Figure 2.5: Prediction of $Y_{h(new)}$ when Parameters Unknown
Prediction Interval for $Y_{h(new)}$ when Parameters Unknown (cont.)

- Since we cannot be certain of the location of the distribution of $Y$, prediction limits for $Y_{h(new)}$ clearly must take account of two elements (Figure 2.5)
  - Variation in possible location of the distribution of $Y$
  - Variation within the probability distribution of $Y$
- Prediction limits for a new observation $Y_{h(new)}$ at $X_h$(given) are obtained:

**Theorem 8**

$$
\frac{Y_{h(new)} - \hat{Y}_h}{s\{pred\}} \sim t_{n-2} \quad \text{for normal error regression model (2.1)}
$$
Prediction Interval for $Y_{h(new)}$ when Parameters Unknown (cont.)

- The $1 - \alpha$ prediction limits for $Y_{h(new)}$:
  
  $$ \hat{Y}_h \pm t(1 - \alpha/2; n - 2) s\{\text{pred}\} $$

- The difference may be viewed as the prediction error, with $\hat{Y}_{h(new)}$ serving as the best point estimate of the value of the new observation $Y_{h(new)}$

- $\sigma^2\{\text{pred}\}$: the variance of the prediction error

  $$ \sigma^2\{\text{pred}\} = \sigma^2\{Y_{h(new)} - \hat{Y}_h\} = \sigma^2 + \sigma^2\{\hat{Y}_h\} $$
Prediction Interval for $Y_{h(new)}$ when Parameters Unknown (cont.)

- $\sigma^2\{\text{pred}\}$ has two components:
  - The variance of the distribution of $Y$ at $X = X_h$; $\sigma^2$
  - The variance of the sampling distribution of $\hat{Y}_h$; $\sigma^2\{\hat{Y}_h\}$

- An unbiased estimator of $\sigma^2\{\text{pred}\}$:

\[
s^2\{\text{pred}\} = MSE + s^2\{\hat{Y}_h\} = MSE \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]
\]
Example

- Toluca Company: $X_h = 100$
- 90 percent prediction interval: $t(0.95; 23) = 1.714$

$$\hat{Y}_h = 419.4 \quad s^2\{\hat{Y}_h\} = 203.72 \quad MSE = 2,384$$

$$\Rightarrow s^2\{\text{pred}\} = 2,384 + 203.72 = 2,587.72$$

$$s\{\text{pred}\} = 50.87$$

- The 90 percent prediction interval for $Y_{h(new)}$:

$$332.2 \leq Y_{h(new)} \leq 506.6$$
## Example (p59)

```r
defitreg<-lm(Hrs~Size, data=toluca)
Xh<-100
fitsnew<-predict.lm(fitreg, data.frame(Size = c(Xh)),
                    se.fit = T, level = 0.9)
s2pred<-fitsnew$se.fit^2+fitsnew$residual.scale^2
nc(c(s2pred, sqrt(s2pred)))
```
This prediction interval is rather wide and may not be useful for planning worker requirements for the next lot.

The interval can still be useful for control purposes.

If the actual work hours fall outside the prediction limits, some alerts may have occurred a change in the production process.
Comparing $Y_{h(new)}$ and $E\{Y_h\}$

- Toluca Company:
  The C.I. for $Y_{h(new)}$ is wider than the C.I. for $E\{Y_h\}$:
  $\because$ predicting the work hours required for a new lot, $\Rightarrow$
  encounter the variability in $\hat{Y}_h$ from sample to sample as well
  as the lot-to-lot variation within the probability distribution
  of $Y$

- (2.38a):
  The prediction interval is wider the further $X_h$ is from $\bar{X}$
Prediction of Mean of $m$ New Observations for Given $X_h$

- Predict the mean of $m$ new observations on $Y$ for a given $X_h$
- $Y$: the mean of the new observations to be predicted as $\bar{Y}_{h(new)}$
- the new $Y$ observations are independent
- The appropriate $1 - \alpha$ prediction limits:

$$\hat{Y}_h \pm t(1 - \alpha/2; n - 2) s\{\text{predmean}\}$$

$$s^2\{\text{predmean}\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\}$$

$$\Leftrightarrow s^2\{\text{predmean}\} = MSE\left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

- Two components for $s^2\{\text{predmean}\}$
Prediction of Mean of \( m \) New Observations for Given \( X_h \) (cont.)

- the variance of the mean of \( m \) observations from the probability distribution of \( Y \) at \( X = X_h \)
- The variance of the sampling distribution of \( \hat{Y}_h \).
Prediction of Mean of \( m \) New Observations for Given \( X_h \) (cont.)

Example

- Toluca Company: \( X_h = 100 \)
- 90 percent prediction interval for the mean number of work hours \( \bar{Y}_{h(new)} \) in three new production runs
- Previous results:
  \( \hat{Y}_h = 419.4; \quad s^2\{Y_h\} = 203.72 \)
  \( MSE = 2,384; \quad t(0.95; 23) = 1.714; \)

\[
\Rightarrow s^2\{\text{predmean}\} = \frac{2,384}{3} + 203.72 = 998.4
\]

\[s\{\text{predmean}\} = 31.60\]
Prediction of New Observation

Prediction of Mean of $m$ New Observations for Given $X_h$
(cont.)

- The prediction interval for $\bar{Y}_{h(new)}$:
  \[ 365.2 \leq \bar{Y}_{h(new)} \leq 473.6 \]

- The total number of work hours:
  \[ 1,095.6 \leq \text{Total work hours} \leq 1,420.8 \]
Partition of Total Sum of Squares

- The analysis of variance approach is based on the partitioning of sums of squares and degrees of freedom associated with $Y$.

- The variation is measured: the deviations of the $Y_i$ around their mean $\bar{Y}$:

$$Y_i - \bar{Y}$$
Partition of Total Sum of Squares (cont.)
Partition of Total Sum of Squares (cont.)

- Total variation: \( \text{SSTO: total sum of squares} \)
  \[
  \text{SSTO} = \sum (Y_i - \bar{Y})^2
  \]
  - \( Y_i \) are the same \( \Rightarrow \) \( \text{SSTO} = 0 \)
  - The greater the variation among the \( Y_i \), the larger is \( \text{SSTO} \).

- \( \text{SSE: error sum of squares} \)
  \[
  \text{SSE} = \sum (Y_i - \hat{Y}_i)^2
  \]
  - \( Y_i \) fall on the fitted regression line \( \Rightarrow \) \( \text{SSE} = 0 \)
  - The greater the variation of the \( Y_i \) around the fitted regression line, the larger is \( \text{SSE} \).
Partition of Total Sum of Squares (cont.)

- **SSR**: regression sum of squares

\[ SSR = \sum (\hat{Y}_i - \bar{Y}_i)^2 \]

- The regression line is horizontal \( \Rightarrow \) \( SSR = 0 \), otherwise \( SSR > 0 \)
- A measure associated with the regression line
- The larger \( SSR \) is in relation to \( SSTO \), the greater is the effect of the regression relation in accounting for the total variation in the \( Y_i \) observations.
Formal Development of Partitioning

- The total deviation:

\[ Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i \]

Two components:

✓ The deviation of the fitted value $\hat{Y}_i$ around the mean $\bar{Y}$.
✓ The deviation of the observation $Y_i$ around the fitted regression line.
Formal Development of Partitioning (cont.)

\[
\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y})^2
\]

- \[2 \sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 0\]
- \[SSR = b_1^2 \sum(X_i - \bar{X})^2\]
- Degrees of freedom: (df)

<table>
<thead>
<tr>
<th>SS</th>
<th>df</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSTO</td>
<td>(n - 1)</td>
<td>(\because \sum(Y_i - \bar{Y}) = 0)</td>
</tr>
<tr>
<td>SSR</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>SSE</td>
<td>(n - 2)</td>
<td>(\because \beta_0, \beta_1 \text{ in } \hat{Y}_i)</td>
</tr>
</tbody>
</table>
### Mean Squares

- **Mean square (MS):**
  A sum of squares divided by its associated df

<table>
<thead>
<tr>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR</td>
<td>1</td>
<td>MSR = ( \frac{SSR}{1} ) (regression mean square)</td>
</tr>
<tr>
<td>SSE</td>
<td>( n - 2 )</td>
<td>MSE = ( \frac{SSE}{n - 2} ) (error mean square)</td>
</tr>
</tbody>
</table>

- Mean squares are **not additive.**

\[
\frac{SSTO}{n - 1} \neq MSR + MSE
\]
**ANOVA table**

**Table 1 : ANOVA Table for Simple Linear Regression**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>$SS$</th>
<th>df</th>
<th>$MS$</th>
<th>$E{MS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$SSR = \sum(\hat{Y}_i - \bar{Y})^2$</td>
<td>1</td>
<td>MSR</td>
<td>$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$</td>
</tr>
<tr>
<td></td>
<td>$SSE = \sum(Y_i - \hat{Y}_i)^2$</td>
<td>$n-2$</td>
<td>MSE</td>
<td>$\frac{SSE}{n-2}$</td>
</tr>
<tr>
<td></td>
<td>$SSTO = \sum(Y_i - \bar{Y})^2$</td>
<td>$n-1$</td>
<td></td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Modified ANOVA table:

\[ SSTO = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n \bar{Y}^2 \]

- **SSTOU**: total uncorrected sum of squares:

\[ SSTOU = \sum Y_i^2 \]

- correction for the mean sum of squares:

\[ SS(\text{correction for mean}) = n \bar{Y}^2 \]
### Table 2: Modified ANOVA Table for Simple Linear Regression

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>$SS$</th>
<th>$df$</th>
<th>$MS$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$SSR = \sum (\hat{Y}_i - \bar{Y})^2$</td>
<td>1</td>
<td>$MSR = \frac{SSR}{1}$</td>
</tr>
<tr>
<td>Error</td>
<td>$SSE = \sum (Y_i - \hat{Y}_i)^2$</td>
<td>$n-2$</td>
<td>$MSE = \frac{SSE}{n-2}$</td>
</tr>
<tr>
<td>Total</td>
<td>$SSTO = \sum (Y_i - \bar{Y})^2$</td>
<td>$n-1$</td>
<td></td>
</tr>
<tr>
<td>Correction for mean</td>
<td>$SS(\text{correction for mean}) = n \bar{Y}^2$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total, uncorrected</td>
<td>$SSTOU = \sum Y_i^2$</td>
<td>$n$</td>
<td></td>
</tr>
</tbody>
</table>
Expected Mean Squares

\[
E\{MSE\} = \sigma^2 \\
E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2
\]

- $MSE$ is an unbiased estimator of $\sigma^2$.

Implication:

- The mean of the sampling distribution of $MSE$ is $\sigma^2$;
- The mean of the sampling distribution of $MSR$ is $\sigma^2$ when $\beta_1 = 0$;
- When $\beta_1 \neq 0$, $E\{MSR\} > E\{MSE\} = \sigma^2$.
  ($\because \beta_1^2 \sum (X_i - \bar{X})^2 > 0$)
F test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$

- The analysis of variance approach provides us with a battery highly useful tests for regression models.
- For the simple linear regression case, the ANOVA provides us with a test:

  $H_0 : \beta_1 = 0$
  $H_a : \beta_1 \neq 0$

Test Statistics:

$$F^* = \frac{MSR}{MSE}$$

- large values of $F^* \Rightarrow H_a$;
- values of $F^*$ near 1 $\Rightarrow H_0$;
F test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$ (cont.)

Cochran’s theorem

If all $n$ observations $Y_i$ come from the same normal distribution with mean $\mu$ and variance $\sigma^2$, and $SSTO$ is decomposed into $k$ sums of squares $SS_r$, each with degrees of freedom $df_r$, then the $SS_r/\sigma^2$ terms are independent $\chi^2$ variables with $df_r$ degrees of freedom if

$$\sum_{r=1}^{k} df_r = n - 1$$
**F test of** $\beta_1 = 0$ **vs.** $\beta_1 \neq 0$ (cont.)

**Property**

If $\beta_1 = 0$ so that all $Y_i$ have the same mean $\mu = \beta_0$ and the same variance $\sigma^2$, $SSE/\sigma^2$ and $SSR/\sigma^2$ are independent $\chi^2$ variables.

- **When** $H_0$ holds:

  $$F^* = \frac{SSR}{\sigma^2} \div \frac{SSE}{\sigma^2} = \frac{MSR}{1} \div \frac{MSE}{n-2} \sim \frac{\chi^2(1)}{1} \div \frac{\chi^2(n-2)}{n-2} \sim F(1, n-2)$$

- **When** $H_a$ holds, $F^*$ follows the noncentral $F$ distribution.
Construction of Decision Rule

- $F^* \sim F(1, n - 2)$
- The decision rule: $\alpha =$ Type I error

**Decision**

- If $F^* \leq F(1 - \alpha; 1, n - 2)$, conclude $H_0$;
- If $F^* > F(1 - \alpha; 1, n - 2)$, conclude $H_a$;

where $F(1 - \alpha; 1, n - 2)$ is the $(1 - \alpha)100$ percentile of the approximate $F$ distribution.
Example

- Toluca Company
- Earlier test on $\beta_1$: (the $t$ test, p46)

**Two-Sided Test**

$$H_0 : \beta_1 = \beta_{10}$$

If $|t^*| = \frac{b_1 - \beta_{10}}{s\{b_1\}} \leq t(1 - \alpha/2; n - 2)$, conclude $H_0$

If $|t^*| \frac{b_1 - \beta_{10}}{s\{b_1\}} > t(1 - \alpha/2; n - 2)$, conclude $H_a$
**Example**

- **Using the $F$ test**

  $H_0 : \beta_1 = \beta_{10} = 0$

  $H_a : \beta_1 \neq \beta_{10} = 0$

- $\alpha = 0.05$; $n = 26$; $F(0.95; 1, 23) = 4.28$

  If $F^* \leq 4.28$, conclude $H_0$

  We have

  $$F^* = \frac{MSR}{MSE} = \frac{252,378}{2,384} = 105.9$$

- **What is the conclusion?**
Equivalence of \( F \) test and \( t \) Test

For a given \( \alpha \) level, the \( F \) test of

\[
H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0
\]

is equivalence algebraically to the two-tailed \( t \) test.

Decision

\[
F^* = \frac{b_1^2}{s^2\{b_1\}} = \left( \frac{b_1}{s\{b_1\}} \right)^2 = (t^*)^2
\]

\[
[t(1 - \alpha/2; n - 2)]^2 = F(1 - \alpha; 1, n - 2)
\]

\( t \) test: two-tailed; \( F \) test: one-tailed;
Steps for General Linear Test Approach

1. Fit the full model and $SSE(F)$
2. Fit the reduced model under $H_0$ and $SSE(R)$
3. Use test statistic and decision rule
1. Full Model

- The full or unrestricted model:

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

- \( SSE(F) \):

\[
SSE(F) = \sum \left[ Y_i - (b_0 + b_1 X_i) \right]^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE
\]
2. Reduced Model

- Hypothesis:
  \[ H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0 \]

- The reduced or restricted model when \( H_0 \) holds:
  \[ Y_i = \beta_0 + \varepsilon_i \]

- \( SSE(R) \):
  \[ SSE(R) = \sum \left[ Y_i - (b_0) \right]^2 = \sum (Y_i - \bar{Y})^2 = SSTO \]
3. Test Statistic

\[ SSE(F) \leq SSE(R) \]

- The more parameters are in the model, the better one can fit the data and the smaller are the deviations around the fitted regression function.
When $SSE(F)$ is not much less than $SSE(R)$, using the full model does not account for much more of the variability of the $Y_i$ than does the reduced model.

⇒ Suggest that the reduced model is adequate i.e., $H_0$ holds.

When $SSE(F)$ is close to $SSE(R)$, the variation of the observations around the fitted regression function for the full model is almost as great as the variation around the fitted regression function for the reduced model.
3. Test Statistic (cont.)

- A small difference \( SSE(R) - SSE(F) \) suggests that \( H_0 \) holds. ⇔ A large difference suggests that \( H_a \) holds.
- Test Statistic: a function of \( SSE(R) - SSE(F) \):

\[
F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \sim F \text{distribution}
\]

when \( H_0 \) holds.
- Decision rule:

If \( F^* \leq F(1 - \alpha; df_R - df_F, df_F) \), conclude \( H_0 \)

If \( F^* > F(1 - \alpha; df_R - df_F, df_F) \), conclude \( H_a \)
For testing whether or not $\beta_1 = 0$, we have

\[ SSE(R) = SSTO \quad SSE(F) = SSE \]
\[ df_R = n - 1 \quad df_F = n - 2 \]
\[ \Rightarrow F^* = \frac{MSR}{MSE} \]
The usefulness of estimates or predictions depends upon the width of the interval.

The user’s needs for precision which vary from one application to another.

No single descriptive measure of the “degree of linear association” can capture the essential information as to whether a given regression relation is useful in any particular application.
Coefficient of Determination

- $SSTO$ is a measure of the uncertainty in predicting $Y$ when $X$ is not considered.
- $SSE$ measures the variation in the $Y_i$ when a regression model utilizing the predictor variable $X$ is employed.
- A natural measure of the effect of $X$ in reducing the variation in $Y$ is to express the reduction in variation ($SSTO - SSE = SSR$) as a proportion of the total variation:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$
Coefficient of Determination (cont.)

- \( R^2 \): the coefficient of determination; 判定係數；決定係數
- \( 0 \leq SSE \leq SSTO \)

\[ \Rightarrow 0 \leq R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \leq 1 \]

\( SSE = 0 \Rightarrow R^2 = 1 \) (If all \( Y_i = \hat{Y}_i \))
\( SSR = 0 \Rightarrow R^2 = 0 \) (If all \( \hat{Y}_i = \bar{Y} \) (\( b_1 = 0 \Rightarrow X \) 與 \( Y \) 無直線關係))
Coefficient of Determination (cont.)

- $R^2 = 0 \Rightarrow X$ 與 $Y$ 無直線關係
  There is no linear association between $X$ and $Y$ in the sample data, and the predictor variable $X$ is of no help in reducing the variation in $Y_i$ with linear regression.

- $R^2 \rightarrow 1 \Rightarrow X$ 軸與 $Y$ 軸變項間的直線關係越強(具有線性變化)
  The closer it is to 1, the greater is said to be the degree of linear association between $X$ and $Y$. 
The Toluca Company:

```r
fitreg <- lm(Hrs ~ Size, data = toluca)
anova(fitreg)
summary(fitreg)
```

```
> anova(fitreg)
Analysis of Variance Table

  Response: Hrs
          Df Sum Sq Mean Sq F value  Pr(>F)
Size      1 252378 252378 105.888 4.449e-10 ***
Residuals 23 54825   2384
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
The variation in work hours is reduced by 82.2% when lot size is considered.
Descriptive Measures of Linear Association between \( X \) and \( Y \)

**Coefficient of Determination (cont.)**

<table>
<thead>
<tr>
<th>TABLE 2.1</th>
<th>( n = 25 )</th>
<th>( \bar{X} = 70.00 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results for Toluca Company Example Obtained in Chapter 1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( b_0 = 62.37 )</td>
<td>( b_1 = 3.5702 )</td>
<td></td>
</tr>
<tr>
<td>( \hat{Y} = 62.37 + 3.5702X )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \sum(X_i - \bar{X})^2 = 19,800 )</td>
<td>( \text{SSE} = 54,825 )</td>
<td>( \text{MSE} = 2,384 )</td>
</tr>
<tr>
<td>( \sum(Y_i - \bar{Y})^2 = 307,203 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 2.2**

The regression equation is

\[ Y = 62.4 + 3.57\; X \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>62.37</td>
<td>26.18</td>
<td>2.38</td>
<td>0.026</td>
</tr>
<tr>
<td>( X )</td>
<td>3.5702</td>
<td>0.3470</td>
<td>10.29</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 48.82 \quad \text{R-sq} = 82.2\% \quad \text{R-sq(adj)} = 81.4\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>( F )</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>252378</td>
<td>252378</td>
<td>105.88</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>23</td>
<td>54825</td>
<td>2384</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>307203</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Descriptive Measures of Linear Association between $X$ and $Y$

**Coefficient of Determination (cont.)**

- $R^2$ is widely used for describing the usefulness of a regression model.

- **Serious misunderstanding:**
  - A high $R^2$ indicates that useful predictions can be made. (Not necessarily correct. Ex: $X_h = 100$)
  - A high $R^2$ indicated that the estimated regression line is a good fit. (Not necessarily correct. Curvilinear)
  - A $R^2$ near 0 indicated that $X$ and $Y$ are not related. (Not necessarily correct. Curvilinear)
Coefﬁcient of Determination (cont.)

(a) Scatter Plot with $R^2 = .69$
Linear regression is not a good ﬁt

(b) Scatter Plot with $R^2 = .02$
Strong relation between $X$ and $Y$
Coefficient of Correlation (相關係數)

- A measure of **linear association** between $Y$ and $X$ when $Y$ and $X$ are random is the **coefficient of correlation**.

$$ r = \pm \sqrt{R^2} $$

- A plus or minus sign is attached to this measure according to whether the **slope** of the fitted regression line is positive or negative.

$-1 \leq r \leq 1$

- The wider the $X_i$ are spaced, the higher $R^2$ will tend to be.

- **SSR**: the “expected variation” in $Y$

- **SSE**: the “unexplained variation”

- $R^2$ is interpreted in terms of the proportion of $SSTO$ in $Y$ which has been “explained” by $X$. 
Normal Correlation Models

- Assume that the $X$ values are known constants?
- The confidence coefficients and risks of errors refer to repeated sampling when $X$ values are kept the same from sample to sample.

- Frequently, it may not be appropriate to consider the $X$ values as known constants.
  - cannot control daily temperatures
  - “height of person” vs. “weight of person”: using correlation model

- the normal correlation model
Two variables $Y_1, Y_2$: bivariate normal distribution.

**Density Function**

The density function of the bivariate normal distribution:

$$f(Y_1, Y_2) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho_{12}}} \exp \left\{ - \frac{1}{2(1 - \rho_{12}^2)} \left[ \left( \frac{Y_1 - \mu_1}{\sigma_1} \right)^2 - 2 \rho_{12} \left( \frac{Y_1 - \mu_1}{\sigma_1} \right) \left( \frac{Y_2 - \mu_2}{\sigma_2} \right) + \left( \frac{Y_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}$$

Five parameters: $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}$
Bivariate Normal Distribution (cont.)
Distinction between Regression and Correlation Model

Bivariate Normal Distribution (cont.)

Marginal Distribution

\[ Y_1, Y_2 \sim N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}): \]

\[ Y_1 \sim N(\mu_1, \sigma_1^2) \Rightarrow f_1(Y_1) = \frac{1}{\sqrt{2\pi\sigma_1}} \exp \left[ -\frac{1}{2} \left( \frac{Y_1 - \mu_1}{\sigma_1} \right)^2 \right] \]

\[ Y_2 \sim N(\mu_2, \sigma_2^2) \Rightarrow f_2(Y_2) = \frac{1}{\sqrt{2\pi\sigma_2}} \exp \left[ -\frac{1}{2} \left( \frac{Y_2 - \mu_2}{\sigma_2} \right)^2 \right] \]

- When \( Y_1, Y_2 \sim N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}) \) \( \Rightarrow \)
  \[ Y_1 \sim N(\mu_1, \sigma_1^2); \ Y_2 \sim N(\mu_2, \sigma_2^2) \]

- The converse is not generally true.
Bivariate Normal Distribution (cont.)

- $\rho_{12}$: the coefficient of correlation between $Y_1$, $Y_2$

\[
\rho_{12} = \frac{\sigma\{Y_1, Y_2\}}{\sigma\{Y_1\}\sigma\{Y_2\}} = \frac{\sigma_{12}}{\sigma_1\sigma_2}
\]

- $Y_1 \perp Y_2 \Rightarrow \sigma_{12} = 0 \Rightarrow \rho_{12} = 0$

- If $Y_1$ and $Y_2$ are positively related $\Rightarrow \sigma_{12}$ and $\rho_{12}$ are positive.

- $-1 \leq \rho_{12} \leq 1$
**Conditional Probability Distribution of $Y_1$**

- $Y_1, Y_2 \sim N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12})$; and $Y_1 \sim N(\mu_1, \sigma_1^2)$; $Y_2 \sim N(\mu_2, \sigma_2^2)$;

- The density function of the conditional probability of $Y_1$ for given value of $Y_2$: ($Y_1|Y_2 \sim N(\alpha_{1|2} + \beta_{12} Y_2, \sigma_{1|2})$)

$$f(Y_1|Y_2) = \frac{f(Y_1, Y_2)}{f_2(Y_2)} = \frac{1}{\sqrt{2\pi\sigma_{1|2}}} \exp \left[ -\frac{1}{2} \left( \frac{Y_1 - \alpha_{1|2} - \beta_{12} Y_2}{\sigma_{1|2}} \right)^2 \right]$$

- $\alpha_{1|2} = \mu_1 - \mu_2 \rho_{12} \frac{\sigma_1}{\sigma_2}$

- $\beta_{12} = \rho_{12} \frac{\sigma_1}{\sigma_2}$

- $\sigma_{1|2}^2 = \sigma_1^2 (1 - \rho_{12}^2)$
Distinction between Regression and Correlation Model

Conditional Probability Distribution of $Y_1$ (cont.)

- $\alpha_{1|2}$: the intercept of the line of regression of $Y_1$ and $Y_2$
- $\beta_{12}$: the slope of this line
- The conditional distribution of $Y_1$, given $Y_2$, is equivalent to the normal error regression model (1.24).
Conditional Probability Distribution of $Y_1$ (cont.)

Three important characteristics of the conditional distributions of $Y_1$:

- **normal**: slice a bivariate normal distribution vertically; scaled its area;

![3D graph of a conditional probability distribution](image-url)
The means of the conditional probability distributions of $Y_1$ fall on a straight line:

$$E\{Y_1|Y_2\} = \alpha_{1|2} + \beta_{12} Y_2$$

All conditional probability distribution have the same standard deviation $\sigma_{1|2}$.

Equivalence to Normal Error Regression Model.
Can we still use regression model (2.1) if $Y_1$ and $Y_2$ are not bivariate normal?

1. The conditional distributions of the $Y_i$, given $X_i$, are normal and independent, with conditional means $\beta_0 + \beta_1 X_i$ and conditional variance $\sigma^2$.

2. The $X_1$ are independent r.v. whose probability distribution does not involve the parameter $\beta_0, \beta_1, \sigma^2$. 
Inferences on Correlation Coefficients

- To study the relationship between two variables: $\rho_{12}$
- MLE of $\rho_{12}$:

$$ r_{12} = \frac{\sum(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\sqrt{\sum(Y_{i1} - \bar{Y}_1)^2 \sum(Y_{i2} - \bar{Y}_2)^2}} $$

- $r_{12}$ is a biased estimator of $\rho_{12}$.
- $-1 \leq r_{12} \leq 1$
Distinction between Regression and Correlation Model

Inferences on Correlation Coefficients (cont.)

Test

The population is bivariate normal

\[ H_0 : \rho_{12} = 0 \quad \text{vs.} \quad H_a : \rho_{12} \neq 0 \]

\((\rho_{12} = 0 \Rightarrow Y_1 \perp Y_2)\)

\[ \iff H_0 : \beta_{12} = 0 \quad \text{vs.} \quad H_a : \beta_{12} \neq 0 \]

\[ \iff H_0 : \beta_{21} = 0 \quad \text{vs.} \quad H_a : \beta_{21} \neq 0 \]

Test statistics:

\[ t^* = \frac{r_{12} \sqrt{n - 2}}{\sqrt{1 - r_{12}^2}} \sim t(n - 2) \]
This test statistics is identical to the regression $t^*$ test statistics $\frac{b_1}{s\{b_1\}}$. ($\therefore r = \sqrt{\frac{SSR}{SSTO}} = b_1 \left( \frac{\sum (X_i - \bar{X})^2}{SSTO} \right)^{1/2}$)

The appropriate decision rule to control the Type I error $\alpha$:

- If $|t^*| \leq t(1 - \alpha/2; n-2)$, conclude $H_0$
- If $|t^*| > t(1 - \alpha/2; n-2)$, conclude $H_a$
Inferences on Correlation Coefficients (cont.)

Inferences on Correlation Coefficients

- A national oil company: service station gasoline sales vs. sales of auxiliary product
- 23 of its service stations
- Average monthly sales data: $Y_1 =$ gasoline sales vs. $Y_2 =$ auxiliary products and services
- $r_{12} = 0.52$; $\alpha = 0.05$
- To test whether or not the association was positive

$$H_0 : \rho_{12} \leq 0 \quad \text{vs.} \quad H_a : \rho_{12} > 0$$

$$t^* = 2.79 > t(0.95; 21) = 1.721; \quad (P\text{-value} = 0.006)$$
Interval Estimation of $\rho_{12}$

- The *Fisher z transformation*:
  \[
  z' = \frac{1}{2} \ln \left( \frac{1 + r_{12}}{1 - r_{12}} \right)
  \]

- When $n$ is large ($n \geq 25$), the sampling distribution of $z'$ is
  \[
  \approx N(E\{z'\}, \sigma^2\{z'\}).
  \]

  \[
  E\{z'\} = \varsigma = \frac{1}{2} \ln \left( \frac{1 + \rho_{12}}{1 - \rho_{12}} \right) \tag{2.90}
  \]

  \[
  \sigma^2\{z'\} = \frac{1}{n - 3} \quad \text{(只與$n$有關)} \tag{2.91}
  \]

- $(z', \varsigma)$: Table B.8
Interval Estimation of $\rho_{12}$ (cont.)

- Interval estimate: ($n \geq 25$)

$$\frac{z' - \varsigma}{\sigma\{z'\}} \sim \mathcal{N}(0, 1)$$

$$\Rightarrow z' \pm z(1 - \alpha/2)\sigma\{z'\}$$

- The $1 - \alpha$ C.I. for $\rho_{12}$ are obtained by transforming the limits on $\varsigma$ by (2.90).
A C.I. for $\rho_{12}$ can be employed to test whether or not $\rho_{12}$ has a specified value. (ex. 0.5)

$0 \leq \rho_{12}^2 \leq 1$: measures the relative reduction in the variability of $Y_2$ associated with the use of variable $Y_1$.

$$\rho_{12}^2 = \frac{\sigma_1^2 - \sigma_{1|2}^2}{\sigma_1^2}$$

$$\rho_{12}^2 = \frac{\sigma_2^2 - \sigma_{2|1}^2}{\sigma_2^2}$$
Spearman Rand Correlation Coefficient

- When **no appropriate transformations** can be found, a nonparametric *rank correlation* procedure may be useful for making inferences about the association between \(Y_1\) and \(Y_2\).
- The ordinal Pearson product-moment correlation coefficient:

\[
r_s = \frac{\sum (R_{i1} - \bar{R}_1)(R_{i2} - \bar{R}_2)}{\left[\sum (R_{i1} - \bar{R}_1)^2 \sum (R_{i2} - \bar{R}_2)^2\right]^{1/2}}
\]

- Test: (two-sided)

\[
H_0 : \text{There is no association between } Y_1 \text{ and } Y_2
\]

\[
H_a : \text{There is an association between } Y_1 \text{ and } Y_2
\]

\[
\Rightarrow t^* = \frac{r_s \sqrt{n - 2}}{1 - r_s^2} \sim t(n - 2)
\]