Chapter 3  Inferences in Regression Analysis

3.1 Inferences concerning $\beta_1$

The reason for interest in testing whether or not $\beta_1 = 0$ is that $\beta_1 = 0$ indicates that there is no linear association between $y$ and $x$.

(i) Sampling distribution of $b_1$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sum_{i=1}^{n} k_i y_i$$

where $k_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$, $\sum k_i = \sum \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = 0$, $\sum k_i x_i = \sum \frac{(x_i - \bar{x})x_i}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} = 1$.

1. Normality

\[\therefore y_i s’ \text{ are independent normal distribution, and } b_1 \text{ is a linear combination of the } y_i.\]

\[\therefore \text{ The sampling distribution of } b_1 \text{ is normal distribution.}\]

2. Mean

$$E(b_1) = E\{\sum_{i} k_i y_i\} = \sum_{i} k_i E(y_i) = \sum_{i} k_i (\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \sum_{i} k_i + \beta_1 \sum_{i} k_i x_i = \beta_1$$
3. Variance
\[ \text{Var}(b_1) = \sigma^2 \{b_1\} = \text{Var}(\sum k_i y_i) = \sum k_i^2 \text{Var}(y_i) = \sum k_i^2 \sigma^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \]

4. Estimated variance
\[ s^2 \{b_1\} = \frac{MSE}{\sum (x_i - \bar{x})^2} = \frac{MSE}{\sum x_i^2 - (\sum x_i)^2 / n} \]

(ii) Sampling distribution of \((b_1 - \beta_1)/s\{b_1\}\)

**Theorem 3.1**: For the regression model (2.1), \(\text{SSE}/\sigma^2\) is distributed as \(\chi^2(n-2)\), and is independent of \(b_0\) and \(b_1\).

By theorem 3.1,
\[ \frac{b_1 - \beta_1}{s\{b_1\}} = \frac{\frac{b_1 - \beta_1}{\sigma\{b_1\}}}{\sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2} \cdot \frac{1}{\sigma\{b_1\}}}} \sim \frac{z}{\sqrt{\frac{\chi^2(n-2)}{n-2}}} = t_{n-2}, \]
where \(\frac{b_1 - \beta_1}{\sigma\{b_1\}} \sim N(0,1)\),

\[ s^2 \{b_1\} = \frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \hat{y}_i)^2} = \frac{MSE}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} = \frac{\sum (y_i - \hat{y}_i)^2}{\sigma^2(n-2)} \sim \chi^2(n-2). \]

(iii) Confidence interval for \(\beta_1\)

the \(1 - \alpha\) confidence interval for \(\beta_1\) is \(b_1 \pm t(1-\alpha/2; n-2)s\{b_1\}\)

(iv) Test for \(\beta_1\)

1. Two-sided test
\[ H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0, \text{ significant level } \alpha \]
Test statistic: \( t^* = \frac{b_1}{s\{b_1\}} \)

If \( |t^*| > t(1 - \alpha/2; n - 2) \), reject \( H_0 \).

2. One-sided test
   \( H_0 : \beta_1 \leq 0 \) vs. \( H_1 : \beta_1 > 0 \), significant level \( \alpha \)

Test statistic: \( t^* = \frac{b_1}{s\{b_1\}} \)

If \( t^* > t(1 - \alpha; n - 2) \), reject \( H_0 \).

### 3.2 Inferences concerning \( \beta_0 \)

(i) Sampling distribution of \( b_0 \)

\[
b_0 = \frac{1}{n} \left( \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i \right) = \frac{1}{n} \sum_{i=1}^{n} y_i - \bar{x} \sum_{i=1}^{n} k_i y_i = \sum_{i=1}^{n} \left( \frac{1}{n} - k_i \bar{x} \right) y_i
\]

1. Normality
   
   \( \therefore \) \( y_i \)'s are independent normal distribution, and \( b_0 \) is a linear combination of the \( y_i \).

   \( \therefore \) The sampling distribution of \( b_0 \) is normal distribution.

2. Mean

\[
E(b_0) = E\{\sum_{i=1}^{n} \left( \frac{1}{n} - k_i \bar{x} \right) y_i \} = \sum_{i=1}^{n} \left( \frac{1}{n} - k_i \bar{x} \right) E(y_i) = \sum_{i=1}^{n} \left( \frac{1}{n} - k_i \bar{x} \right) (\beta_0 + \beta_1 x_i)
\]

\[
= \beta_0 + \beta_1 \bar{x} - \beta_0 \bar{x} \sum_{i=1}^{n} k_i - \beta_1 \bar{x} \sum_{i=1}^{n} k_i x_i = \beta_0
\]

3. Variance

\[
Var(b_0) = \sigma^2 \{b_0\} = Var\{\sum_{i=1}^{n} \left( \frac{1}{n} - k_i \bar{x} \right) y_i \} = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]
\]

4. Estimated variance
\[ s^2\{b_0\} = MSE\left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right] \]

(ii) Sampling distribution of \( (b_0 - \beta_0)/s\{b_0\} \)

\[ \frac{b_0 - \beta_0}{s\{b_0\}} \sim t_{n-2}. \]

(iii) Confidence interval for \( \beta_0 \)

the \( 1 - \alpha \) confidence interval for \( \beta_0 \) is \( b_0 \pm t(1 - \alpha/2; n - 2)s\{b_0\} \)

3.4 Interval estimation of \( E(y_h) \)

Level \( x_h \): may be value occurred in the sample, or some other value of the independent variable within the scope of the model.

\[ \Rightarrow \hat{y}_h = b_0 + b_1 x_h \]

(i) Sampling distribution of \( \hat{y}_h \)

1. Normality

\[ \therefore \hat{y}_h \text{ is a linear combination of the } y_i. \]

\[ \therefore \text{The sampling distribution of } \hat{y}_h \text{ is normal distribution.} \]

2. Mean

\[ E(\hat{y}_h) = E(b_0 + b_1 x_h) = E(b_0) + x_h E(b_1) = \beta_0 + \beta_1 x_h \]

\[ \Rightarrow \hat{y}_h \text{ is an unbiased estimator of } E(y_h) \]

3. Variance

To first show that \( b_1 \) and \( \bar{y} \) are uncorrelated,

\[ Cov(b_1, \bar{y}) = \sigma\{b_1, \bar{y}\} = \sum_n \frac{1}{n} k_i Var(y_i) = \frac{\sigma^2}{n} \sum k_i = 0, \text{ where} \]
\[ b_i = \sum_{i=1}^{n} k_i y_i, \quad \bar{y} = \frac{\sum y_i}{n}. \]

\[ \Rightarrow \text{Var}(\hat{y}_h) = \sigma^2 \{ \hat{y}_h \} = \text{Var}[\bar{y} + b_1(x_h - \bar{x})] \]

\[ = \text{Var}(\bar{y}) + (x_h - \bar{x})^2 \cdot \text{Var}(b_1) = \frac{\sigma^2}{n} + (x_h - \bar{x})^2 \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \]

\[ \Rightarrow s^2 \{ \hat{y}_h \} = \text{MSE} \left[ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right] \]

(ii) Sampling distribution of \((\hat{y}_h - E\{y_h\})/s\{\hat{y}_h\}\) and confidence interval for \(E(y_h)\)

\[ \frac{\hat{y}_h - E(y_h)}{s\{\hat{y}_h\}} \sim t_{n-2} \]

The \(1 - \alpha\) confidence interval for \(E(y_h)\) is \(\hat{y}_h \pm t(1-\alpha/2; n-2)s\{\hat{y}_h\}\)

### 3.8 Analysis of variance (ANOVA)

(i) Partitioning of total sum of squares

**Basic Notions:**

1. \(y_i - \bar{y}\): the deviation of the \(y_i\) around the mean \(\bar{y}\)

2. \(SSTO = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2\): total sum of square

3. \(SSE = \sum (y_i - \hat{y}_i)^2\): error sum of squares

4. \(SSR = \sum (\hat{y}_i - \bar{y})^2 = b_1 \left( \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \right) \]

\[ = b_1 [\sum (x_i - \bar{x}) (y_i - \bar{y})] = b_1^2 \sum (x_i - \bar{x})^2 \]: regression sum of squares

\[ \therefore \sum (y_i - \bar{y})^2 = \sum [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \]

\[ + 2 \sum (\hat{y}_i - \bar{y}) (y_i - \hat{y}_i) = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2, \]
where \[ 2 \sum (\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 2 \sum \hat{y}_i(y_i - \hat{y}_i) - 2 \overline{y} \sum (y_i - \hat{y}_i) = 2 \sum e_i \hat{y}_i = 0 \]

\[ \therefore SSTO = SSR + SSE \]

(ii) Analysis of variance table

<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 - \overline{y})^2 )</td>
<td>1</td>
<td>MSR = SSR/1</td>
<td>( \sigma^2 + \beta_1^2 \sum (x_i - \overline{x})^2 )</td>
</tr>
<tr>
<td>Error</td>
<td>SSE = ( \sum (y_i - \hat{y}_i)^2 )</td>
<td>( n - 2 )</td>
<td>MSE = SSE/( n - 2 )</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td>Total</td>
<td>SSTO = ( \sum (y_i - \overline{y})^2 )</td>
<td>( n - 1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correction for mean</td>
<td>SS (correction for mean) = ( n\overline{y}^2 )</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total, uncorrected</td>
<td>SSTOU = ( \sum y_i^2 )</td>
<td>( n )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note:

1. SSTOU = \( \sum y_i^2 \) : total uncorrected sum of squares

2. \( df \) of SSE: two degrees of freedom are lost because the two parameters \( \beta_0 \) and \( \beta_1 \) were estimated in obtaining the fitted values \( \hat{y}_i \).

3. \( df \) of SSR: there are two parameters in SSR, but the deviations \( \hat{y}_i - \overline{y} \) must sum to zero; hence one degree of freedom is lost.

4. \( E\{MSR\} = E\{SSR\} = E\{b_1^2 \sum (x_i - \overline{x})^2\} = \sigma^2 + \beta_1^2 \sum (x_i - \overline{x})^2 \),

where \( E(b_1^2) = \text{Var}(b_1) + [E(b_1)]^2 = \frac{\sigma^2}{\sum (x_i - \overline{x})^2} + \beta_1^2 \).

(iii) \( F \) test of \( \beta_1 = 0 \) vs. \( \beta_1 \neq 0 \)

Cochran’s theorem: If \( y_1, \ldots, y_n \) 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 

$$\sum_{r=1}^k df_r = n - 1.$$  

$\Rightarrow$ If $SSTO = SS_1 + SS_2 + \ldots + SS_k$ and $n - 1 = df_1 + df_2 + \ldots + df_k$,

then $SS_r/\sigma^2 \overset{\text{indep.}}{\sim} \chi^2(df_r), \ r = 1 \ldots k$

Under $H_0$ (i.e., $\beta_1 = 0$), $y_1, \ldots, y_n$ 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 (by Cochran’s theorem).

$H_0 : \beta_1 = 0 \ vs. \ H_1 : \beta_1 \neq 0$, significant level $\alpha$

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

where $\frac{SSR/\sigma^2}{1} \sim \chi^2(1), \ \frac{SSE/\sigma^2}{n-2} \sim \chi^2(n-2)$.  

If $F^* > F(1 - \alpha; 1, n - 2)$, reject $H_0$.

(v) Equivalence of $F$ test and $t$ test

$H_0 : \beta_1 = 0 \ vs. \ H_1 : \beta_1 \neq 0$, significant level $\alpha$

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

$F$ test: $F^* = \frac{SSR/\sigma^2}{1} \div \frac{SSE/\sigma^2}{n-2} = \frac{b_1^2 \sum (x_i - \bar{x})^2}{MSE} = \left( \frac{b_1}{s\{b_1\}} \right)^2 = (t^*)^2$

where $s^2\{b_1\} = MSE/\sum (x_i - \bar{x})^2$.

$\Rightarrow F(1 - \alpha; 1, n - 2) = [t(1 - \alpha/2; n - 2)]^2$
3.10 Descriptive measures of association between \( x \) and \( y \) in regression model

(i) Coefficient of determination

\( SSTO \): measures the uncertainty in predicting \( y \) when \( x \) is not considered.

\( SSE \): measures the variation in \( y_i \) when \( x_i \) is employed.

A measure of the effect of \( x \) in reducing the variation in \( y \) is

The coefficient of determination:

\[
 r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}
\]

\( \therefore 0 \leq SSE \leq SSTO \) \( \therefore 0 \leq r^2 \leq 1 \)

The interpretation: the proportionate reducing of total variation associated with the use of the independent variable \( x \).

The limiting values of \( r^2 \):

1. If all observations fall on the fitted regression line

\[ y_i = \hat{y}_i \Rightarrow SSE = 0 \Rightarrow r^2 = 1 \]

\( \Rightarrow \) the independent variable \( x \) accounts for all variation in the observation \( y \).

2. If the slope of the fitted regression line is \( b_1 = 0 \)

\[ \hat{y}_i = \bar{y} \Rightarrow SSE = SSTO \Rightarrow r^2 = 0 \]

\( \Rightarrow \) there is no linear association between \( x \) and \( y \).

(ii) Coefficient of correlation

\[ r = \pm \sqrt{r^2}, -1 \leq r \leq 1, \]

where a plus or minus sign is according to the sign of the slope of the fitted regression line.
Note:

The relation between $b_1$ and $r$:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\left[ \sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2 \right]^{1/2}} = \frac{\sum x_i y_i - \frac{\sum x_i y_i}{n}}{\left[ \left( \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right) \left( \sum y_i^2 - \frac{(\sum y_i)^2}{n} \right) \right]^{1/2}}$$

$$\Rightarrow b_1 = \left[ \frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2} \right]^{1/2} \cdot r = \left( \frac{s_y}{s_x} \right) r \quad \text{where} \quad s_y = \left[ \sum (y_i - \bar{y})^2 / n - 1 \right]^{1/2} ,$$

$$s_x = \left[ \sum (x_i - \bar{x})^2 / n - 1 \right]^{1/2}$$

Note: when $b_1 = 0$, $r = 0$ (implies a horizontal fitted regression line)