Recall:

In the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1,2,...,n$$
$$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2 \quad and \quad Cov(\varepsilon_i, \varepsilon_j) = 0 \ for\ all\ i \neq j.$$

Then

$$E(Y_i) = \beta_0 + \beta_1 X_i \qquad and \qquad Var(Y_i) = \sigma^2.$$

The point estimates of β_0, β_1 are

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = b_0 = \bar{Y} - b_1\bar{X}$$

$$\hat{\beta}_1 = b_1 = \sum_{i=1}^{n} K_i Y_i, \quad K_i = \frac{\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}, \ \hat{\beta}_0 = \sum_{i=1}^{n} L_i Y_i, \quad L_i = \frac{1}{n} - \bar{X} K_i$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}, \ Var(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right],$$

The unbiased estimate of σ^2 is

$$s^2 = MSE = \widehat{\sigma^2} = \frac{SSE}{n-2}, \quad SSE = \sum_{i=1}^{n} e_i^2, \quad e_i = Y_i - \widehat{Y}_i, \quad i = 1,2,...$$

Interval Estimation of the mean response when $X = X_h$

- The mean response when $X = X_h$ is denoted $E(Y_h)$
- The point estimation of the response mean is given by

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

- Sampling distribution of $\hat{Y}_h$:

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

  $b_0 \sim N\left(\beta_0, Var(b_0)\right)$, $b_1 \sim N\left(\beta_1, Var(b_1)\right)$, then $\hat{Y}_h$ is normally distributed with mean

$$E(\hat{Y}_h) = E(b_0) + E(b_1 X_h) = E(Y_h) = \beta_0 + \beta_1 X_h$$

and

$$\begin{aligned} Var(\hat{Y}_h) &= Var(b_0 + b_1 X_h) = Var(\bar{Y} + b_1(X_h - \bar{X})) \\ &= Var(\bar{Y}) + (X_h - \bar{X})^2 Var(b_1) + 2(X_h - \bar{X}) Cov(\bar{Y}, b_1) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2 (X_h - \bar{X})^2}{S_{xx}} + 0 \end{aligned}$$

because

$$Cov(\bar{Y}, b_1) = Cov\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} k_i y_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} k_i Cov(y_i, y_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} k_i Var(y_i) = 0$$

$$= \frac{1}{n}\sum_{i=1}^{n} k_i \sigma^2$$

Then

$$Var(\hat{Y}_h) = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}}\right) = MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}}\right)$$

$$S.E(\hat{Y}_h) = \sqrt{Var(\hat{Y}_h)}$$

Lemma

The statistics $\frac{\hat{Y}_h - E(Y_h)}{S.E(\hat{Y}_h)}$ had t distribution with (n-2) degrees of freedom. For more details, see the book (page 54).

This lemma enables us to construct $100(1-\alpha)\%$ confidence interval about the response mean Y_h as follows:

$$\hat{Y}_h \pm t_{1-\alpha/2, n-2}\ S.E.(\hat{Y}_h)$$

Example

Consider the Toluca Company example, find 90% confidence interval for the response mean $E(Y_h)$ when $X_h = 65$.

From the data, we have

$$\hat{Y}_h = b_0 + b_1 X_h = 62.37 + 3.5702(65) = 294.4$$

$$Var(\hat{Y}_h) = MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}}\right) = 2\,384\,[\frac{1}{25} + \frac{(65 - 70.0)^2}{19800}] = 98.37$$

$$S.E(\hat{Y}_h) = \sqrt{Var(\hat{Y}_h)} = 9.918$$

For a 90 percent confidence coefficient, we require *t(.95;* 23) = 1.714.

Hence, our confidence interval with confidence coefficient .90 is by

$$\hat{Y}_h \pm t_{1-\alpha/2,n-2}\ S.E.(\hat{Y}_h)$$

$$294.4 - 1.714(9.918) < E(Y_h) < 294.4 + 1.714(9.918)$$

$$277.4 < E(Y_h) < 311.4$$

We conclude with confidence coefficient .90 that the mean number of work hours required when lots of 65 units are produced is somewhere between 277.4 and 311.4 hours. We see that our estimate of the mean number of work hours is moderately precise.

Example 2 Suppose the Toluca Company wishes to estimate $E\{Y_h\}$ for lots with $X_h = 100$ units with a 90 percent confidence interval. We require:

$$\hat{Y}_h = 62.37 + 3.5702(100) = 419.4$$
$$s^2\{\hat{Y}_h\} = 2{,}384\left[\frac{1}{25} + \frac{(100 - 70.00)^2}{19{,}800}\right] = 203.72$$
$$s\{\hat{Y}_h\} = 14.27$$
$$t(.95; 23) = 1.714$$

Hence, the 90 percent confidence interval is:

$$419.4 - 1.714(14.27) \le E\{Y_h\} \le 419.4 + 1.714(14.27)$$
$$394.9 \le E\{Y_h\} \le 443.9$$

Note that this confidence interval is somewhat wider than that for Example 1, since the X_h level here ($X_h = 100$) is substantially farther from the mean $\bar{X} = 70.0$ than the X_h level for Example 1 ($X_h = 65$).

One may use R commands:

```
newx = data.frame(x=100)
predict(model, newx, level=0.90,interval="confidence")
```

Predicting a Future Observation When *X* is Known

We consider now the prediction of a new observation Y corresponding to a given level X of the predictor variable.

If β_0, β_1, σ were known, we'd know that the distribution of responses when $X=X_h$ is normal with mean $\beta_0 + \beta_1 X_h$ and standard deviation σ. Thus, making use of the normal distribution (and equivalently, the empirical rule) we know that if we took a sample item from this distribution, it is very likely that the value fall within 2 standard deviations of the mean. That is, we would know that the probability that the sampled item lies within the range $(\beta_0 + \beta_1 X_h - \sigma, \beta_0 + \beta_1 X_h + \sigma)$ is approximately 0.95.

In practice, we don't know the mean $\beta_0 + \beta_1 X_h$ or the standard deviation σ. However, we just constructed a (1-α)100% Confidence Interval for $E\{Y_h\}$,

The prediction error is for the new observation is the difference between the observed value and its predicted value: $Y_{h(new)} - \hat{Y}_h$. Since the data are assumed to be independent, the new (future) value is independent of its predicted value, since it wasn't used in the regression analysis. The variance of the prediction error can be obtained as follows:

$$Var\{pred\} = Var\{Y_{h(new)} - \hat{Y}_h\} = Var\{Y_{h(new)}\} + Var\{\hat{Y}_h\} = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$

$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$

and an unbiased estimator is:

$$Var\{pred\} = MSE\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$

$$S.E\{pred\} = \sqrt{Var(pred)}$$

A (1-α)100% Prediction Interval for New Observation When $X=X_h$

$$\hat{Y}_h \pm t(\alpha/2; n-2)\sqrt{MSE\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}.$$

Example (book: page.59)

The Toluca Company studied the relationship between lot size and work hours primarily to obtain information on the mean work hours required for different lot sizes for use in determining the optimum lot size. The company was also interested, however, to see whether the regression relationship is useful for predicting the required work hours for individual lots. Suppose that the next lot to be produced consists of $X_h = 100$ units and that a 90 percent prediction interval is desired. We require $t(.95; 23) = 1.714$. From earlier work, we have:

In this example

$$\hat{Y}_h = b_0 + b_1 X_h = 62.37 + 3.5702(100) = 419.4$$

$$Var\{pred\} = MSE\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right] = 2587.72$$

$$S.E\{pred\} = \sqrt{Var(pred)} = 50.87$$

For a 90 percent confidence coefficient, we require $t(.95;\ 23) = 1.714$.

Hence, our confidence interval with confidence coefficient .90 is by

$$\hat{Y}_h \pm t_{1-\alpha/2,n-2}\ S.E.(\text{Pred})$$

$$419.4 - 1.714(50.87) < Y_{h(new)} < 419.4 + 1.714(50.87)$$

$$332.2 < Y_{h(new)} < 506.6$$

With confidence coefficient .90, we predict that the number of work hours for the next production run of 100 units will be somewhere between 332 and 507 hours

R commands:

```
newx = data.frame(x=100)
predict(model, newx, level=0.90,interval="predict")
```