# Confidence and Prediction Intervals

- **Normal Error Regression Model.** That is,

$$Y_i = (\beta_0 + \beta_1 x_i) + \epsilon_i. \qquad i = 1, 2, \ldots, n.$$

where:

**1.** $\beta_0$ and $\beta_1$ are parameters.

**2.** $x_i$ is a known constant.

**3.** $\epsilon_i$ are independent $N(0, \sigma^2)$.

- **Equation of the Least-Squares Regression Line.** The equation of the least-squares regression line of $y$ on $x$ is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

with *slope*

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y)}{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} \tag{1}$$

and *intercept*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2}$$

- **Mean Response of $Y$ at a specified value $x_h$, $E(Y_h) = \mu_{Y|x_h}$.**

**1. Point Estimate.** For a specific value $x_h$, the estimate of the **mean** value of $Y$ is given by

$$\hat{\mu}_{Y|x_h} = b_0 + b_1 x_h$$

Note: $\dfrac{\hat{\mu}_{Y|x_h} - \mu_{Y|x_h}}{SE_{\hat{\mu}}} \sim t_{df=n-2}$, where, $SE^2_{\hat{\mu}} = MSE\left[\dfrac{1}{n} + \dfrac{(x_h - \bar{x})^2}{SS_{xx}}\right]$

**2. Confidence Interval.** For a specific value $x*$, the $(1-\alpha)100\%$ confidence interval for $\mu_{Y|x_h}$ is given by

$$\hat{\mu}_{Y|x_h} \pm t_{\alpha/2;(n-2)} SE_{\hat{\mu}}$$

- **Prediction of New Observation, $Y_h$ at a specified value $x_h$.**

**1. Point Estimate.** For a specific value $x_h$, the predicted value of $Y$ is given by

$$\hat{y} = b_0 + b_1 x_h$$

Note: $\dfrac{Y_{h(new)} - \hat{Y}_h}{SE_{\hat{y}}} \sim t_{df=n-2}$, where, $SE^2_{\hat{y}} = MSE\left[1 + \dfrac{1}{n} + \dfrac{(x_h - \bar{x})^2}{SS_{xx}}\right]$

**2. Prediction Interval.** For a specific value $x_h$, the $(1-\alpha)100\%$ prediction interval is given by

$$\hat{y} \pm t_{\alpha/2;(n-2)} SE_{\hat{y}}$$

- **Prediction of Mean of $m$ new observations at a specified value $x_h$.**

**1. Point Estimate.** For a specific value $x_h$, the predicted value of $Y$ is given by

$$\hat{\mu}_{Y_m|x_h} = b_0 + b_1 x_h$$

2. **Prediction Interval.** For a specific value $x_h$, the $(1-\alpha)100\%$ prediction interval is given by

$$\hat{\mu}_{Y_m|x_h} \pm t_{\alpha/2;(n-2)} SE_{\hat{\mu}_{Y_m|x_h}}$$

where, $SE_{\hat{\mu}_{Y_m|x_h}} = MSE\left[\frac{1}{m} + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{SS_{xx}}\right]$

- **Confidence Band for Regression Line.** The Working-Hotelling $1-\alpha$ confidence band for the regression line has the following two boundary values at any level $x_h$:

$$\hat{\mu}_{Y|x_h} \pm W * SE_{\hat{\mu}}, \text{ where, } W^2 = 2 * F(1-\alpha; 2, n-2)$$

- **Production Run.** Consider the following data of 10 production runs of a certain manufacturing company.

| Production run | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Lot size ($x$) | 30 | 20 | 60 | 80 | 40 | 50 | 60 | 30 | 70 | 60 |
| Man-Hours ($y$) | 73 | 50 | 128 | 170 | 87 | 108 | 135 | 69 | 148 | 132 |

1. Determine the regression line.

2. Find an estimate for the **mean** number of man-hours ($\hat{\mu}_{Y|x=100}$) required to produce a lot size 100.

3. Construct a 95% confidence interval for the **mean** number of man-hours ($\mu_{Y|x=100}$) required to produce a lot size 100.

4. Predict the number of man-hours ($\hat{y}$) required to produce a lot size 100.

5. Construct a 95% prediction interval for the number of man-hours ($\hat{y}$) required to produce a lot size 100.

**6.** Construct a 95% prediction interval for the mean number of work hours in three new production runs, each for $X_h = 100$ units.

**7.** Construct the 95% confidence band for the regression line.

- R commands

```
size=c(30,20,60,80,40,50,60,30,70,60)
hours=c(73,50,128,170,87,108,135,69,148,132)

lm.size.hours=lm(hours~size)
coef(lm.size.hours)
plot(size,hours)
abline(lm.size.hours)

# Confidence intervals for the mean response
new=data.frame(size=c(size,100))
predict(lm.size.hours,newdata=new)      # the predicted value of a new observation

# This will give the mean C.I. for the new data
predict(lm.size.hours,newdata=new,interval="confidence")
new2=data.frame(size=c(90,100))   # new2 contains only the 2 new observations

# This will give the mean C.I. for the 2 new data
predict(lm.size.hours,newdata=new2,interval="confidence")

# Prediction interval for future values
predict(new2=data.frame(size=c(90,100))     # new2 contains only the 2 new observations
predict(lm.size.hours,newdata=new2,interval="prediction",level=.99) # The default level is 0.95
```

```
# Confidence Band
CI=predict(lm.size.hours,se.fit=TRUE)     # se.fit=SE(mean)
W=sqrt(2*qf(0.95,2,8))
band.lower=CI$fit - W*CI$se.fit
band.upper=CI$fit + W*CI$se.fit

plot(size,hours,xlab="Production Size",ylab="Work Hours",main="Confidence Band")
abline(lm.size.hours)

points(sort(size), sort(band.lower), type="l", lty=2)
points(sort(size), sort(band.upper), type="l", lty=2)

# If the regression slope is negative, you need to sort in reverse order
points(sort(size), sort(band.lower), decreasing=TRUE), type="l", lty=2)
points(sort(size), sort(band.upper), decreasing=TRUE), type="l", lty=2)

# Or alternatively, you can use:
# source('Math445_Fall2016/confidence.band.R')
confidence.band(lm.size.hours)
```

- **Tree Circumference and Height.** Listed below are the circumferences (in feet) and the heights (in feet) of trees in Marshall, Minnesota (based on data from "Tree Measurements" by Stanley Rice, *American Biology Teacher*.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x$ (circ) | 1.8 | 1.9 | 1.8 | 2.4 | 5.1 | 3.1 | 5.5 |
| $y$ (height) | 21.0 | 33.5 | 24.6 | 40.7 | 73.2 | 24.9 | 40.4 |
| $x$ (circ) | 5.1 | 8.3 | 13.7 | 5.3 | 4.9 | 3.7 | 3.8 |
| $y$ (height) | 45.3 | 53.5 | 93.8 | 64.0 | 62.7 | 47.2 | 44.3 |

  1. Determine the regression line.
  2. Find an estimate for the **mean** height of trees with circumference of 5 ft.
  3. Construct a 95% confidence interval for the **mean** height of trees with circumference of 5 ft.
  4. Predict the height ($\hat{y}$) of a tree with circumference of 5 ft.
  5. Construct a 95% prediction interval for the height ($\hat{y}$) of a tree with circumference of 5 ft.
  6. Construct a 95% prediction interval for the mean height of five new trees, each with circumference of about 5 ft.
  7. Construct the 95% confidence band for the regression line.

- **Recommended problems:** Answer the following questions:
  On pages 91-92, #13 and 16.