$\hat{\beta}_1$

THEOREM 2.10

$$t_{calc} = \underbrace{\bar{y}_1}_{} - \bar{y}_2 - \Delta_0 \Big/ \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$$

Note: $MSE = s_p^2$, $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_2$, $\Delta_0 = \beta_{10}$ (the hypothesized value frequently zero).

Example 6

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{1}{S_{xx}}$$

To investigate the maternal behavior of laboratory rats, we move the rat pup a fixed distance from the mother and record the time (in seconds) required for the mother to retrieve the pup to the nest. We run the study with 5- and 20-day old pups. Note: These are not the same pups being measured twice.

G_1 5 Days	G_2 20 Days
15	30
10	15
25	20
15	25
20	23
18	20

$\bar{y}_1 = 17.1667$

$\bar{y}_2 = 22.1667$

$n_1 = n_2 = 6$

Use regression techniques with a dummy variable to test if the retrieval time differs per group.

	Response (Y)	Covariate (x)	$\hat{Y}$
1	15	1	17.1667
2	10	1	17.1667
3	25	1	.
4	15	1	.
5	20	1	.
6	18	1	.
7	30	0	22.1667
8	15	0	22.1667
.	20	0	.
.	25	0	.
	23	0	.
12	20	0	.

$$X = \begin{cases} 1 & 5 \text{ day} \\ 0 & 20 \text{ day} \end{cases}$$

$$\hat{y} = \underbrace{\bar{y}_2}_{\hat{\beta}_0} + \underbrace{(\bar{y}_1 - \bar{y}_2)}_{\hat{\beta}_1} \cdot x$$

$$= 22.1667 + (17.1667 - 22.1667) \cdot x$$

$$= 22.1667 - 5 \cdot x$$

The R code follows:

```
# two groups
y1 <- c(15,10,25,15,20,18)
y2 <- c(30,15,20,25,23,20)
# response
y <- c(y1,y2)
# dummy variable
x <- c(rep(1,6),rep(0,6))
# multiple boxplot
boxplot(y~x)
# test
summary(lm(y~x))
t.test(y1,y2,var.equal = TRUE)
```

Stack

n = 12

The R output follows:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   22.167      2.088  10.615 9.18e-07 ***
x             -5.000      2.953  -1.693    0.121
---

Residual standard error: 5.115 on 10 degrees of freedom
Multiple R-squared:  0.2228,Adjusted R-squared:  0.145
F-statistic: 2.866 on 1 and 10 DF,  p-value: 0.1213
```

$\sqrt{MSE}$

The R output follows:

```
Two Sample t-test

data:  y1 and y2
t = -1.693, df = 10, p-value = 0.1213
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.580454   1.580454
sample estimates:
mean of x mean of y
 17.16667  22.16667
```

2.10 Prediction

In simple linear regression, there are two fundamental goals:

1. Test if there is a relationship between the response variable Y and covariate x. The first goal is accomplished by testing null hypothesis $H_0 : \beta_1 = \beta_{10}$.] Design

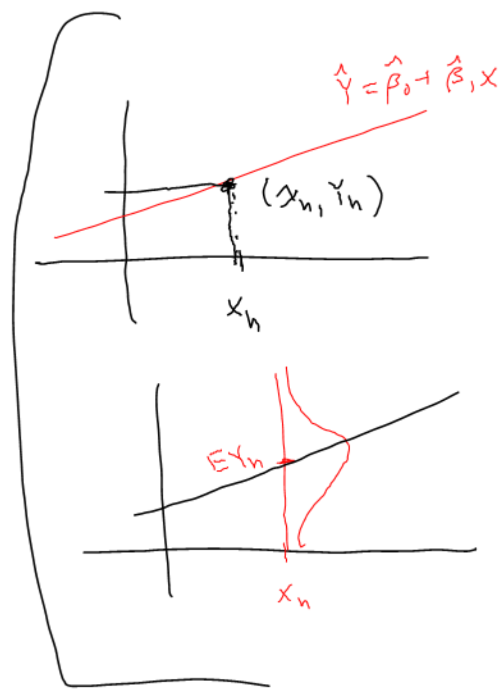
2. Predict the response Y given a fixed value of x. This section describes predictions and confidence intervals on predictions.] ML

<u>Inferences concerning $E[Y_h]$</u>

$E[Y_h]$ vs. Y_{new}

The parameter of interest is: $\theta = E[Y_h]$

μ_h



PROPOSITION 2.10 *Let*

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

where x_h is some fixed value of x. Then

$$E[\hat{Y}_h] = \beta_0 + \beta_1 x_h$$

and

$$\text{Var}[\hat{Y}_h] = \sigma^2 \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}} \right]$$

From the above proposition, the standardized score of $\hat{Y}_h$ is

$$Z = \frac{\overbrace{\hat{\beta}_0 + \hat{\beta}_1 x_h}^{\hat{Y}_h} - \overbrace{(\beta_0 + \beta_1 x_h)}^{E\hat{Y}_h}}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}} \right)}}.$$

$\sigma_{\hat{Y}_h} \rightarrow$

Since $\hat{Y}_h$ is a linear combination or response variable Y_i, the random variable Z has a standard normal distribution. The studentized score of $\hat{Y}_h$ is

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_h - (\beta_0 + \beta_1 x_h)}{\sqrt{MSE \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}} \right)}}.$$

The *appropriate* proposition (reference GR5204) implies T has a student's t-distribution with $n - 2$ degrees of freedom. Consequently, the confidence interval of interest follows.

Confidence interval for $E[Y_h]$

The $100(1-\alpha)\%$ confidence interval for $E[Y_h]$ when $x = x_h$ is

$$\boxed{\underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_h)}^{\hat{Y}_h} \pm t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)}}$$

Inferences concerning the prediction of future Y values ($Y_{h(new)}$)

Goal: For fixed $x = x_h$, we want to find a C.I. for a *single future value* $Y_{h(new)}$ as compared to a C.I for the *true average* $E[Y_h]$.

The **prediction error** for a single future response value is

$$W = Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h) = Y_{h(new)} - \hat{Y}_h \tag{2.18}$$

PROPOSITION 2.11 *The expected value and variance of the prediction error defined Equation (2.18) are respectively given by*

$$EW = E[Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h)] = 0$$

and

$$\text{Var}\,W = \text{Var}[Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h)] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]$$

Proof

$$E[Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h)] = EY_{h(new)} - E[\hat{\beta}_0 + \hat{\beta}_1 x_h]$$
$$= \beta_0 + \beta_1 x_h - (\beta_0 + \beta_1 x_h)$$
$$= 0$$

$$\text{Var}(\underbrace{Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h)}_{\text{Ind.}}] = \text{Var}(Y_{h(new)}) + \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_h)$$
$$= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)$$
$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]$$

In a similar manner as the confidence interval for $E[Y_h]$, the studentized score of prediction error (2.18) is given by

$$T = \frac{Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h) - 0}{\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)}}.$$

The *appropriate* proposition (reference GR5204) implies T has a student's t-distribution with $n-2$ degrees of freedom. Consequently, the interval of interest follows below.

Prediction interval for a single future value $Y_{h(new)}$

The $100(1-\alpha)\%$ prediction interval for a single future value of $Y_{h(new)}$ when $x = x_h$ is

$\hat{Y}_h$

$$(\hat{\beta}_0 + \hat{\beta}_1 x_h) \pm t_{\alpha/2,n-2} \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)}$$

Example 2 continued

C.I. i. Using the appropriate interval, estimate the **true average** energy expenditure for someone with a fat-free body mass of 65 [kg].

P.I. ii. Using the appropriate interval, estimate the energy expenditure for a **single respondent** having a fat-free body mass of 65 [kg].

Example 6 continued

i. Using the appropriate interval, estimate the **true average** retrieval time for both the five day and twenty day old pups.

Example 2 R code follows:

```
# data
x <- c(49.3,59.3,68.3,48.1,57.6,78.1,76.1)
y <- c(1894,2050,2353,1838,1948,2528,2568)
# model
model <- lm(y~x)
# dataframe
x_data <- data.frame(x=65)
# confidence interval
predict(model,newdata=x_data,interval="confidence")
# prediction interval
predict(model,newdata=x_data,interval="prediction")
```

Example 2 R output follows:

$\hat{Y}_h$

```
       fit      lwr     upr
1 2233.459 2168.777 2298.14
```

C.I.

$\hat{Y}_h$

```
       fit      lwr      upr
1 2233.459 2054.654 2412.264
```

P.I. ← wider

Example 6 R code follows:

```
# two groups
y1 <- c(15,10,25,15,20,18)
y2 <- c(30,15,20,25,23,20)
# response
y <- c(y1,y2)
# dummy variable
x <- c(rep(1,6),rep(0,6))
# prediction
x_data <- data.frame(x=c(1,0))
predict(lm(y~x),newdata=x_data,interval="confidence")
```

Example 6 R output follows:

```
        fit      lwr      upr
1 17.16667 12.51358 21.81975
2 22.16667 17.51358 26.81975
```

SKIP

<u>Further Inspection on confidence intervals and prediction intervals</u>

C.I.

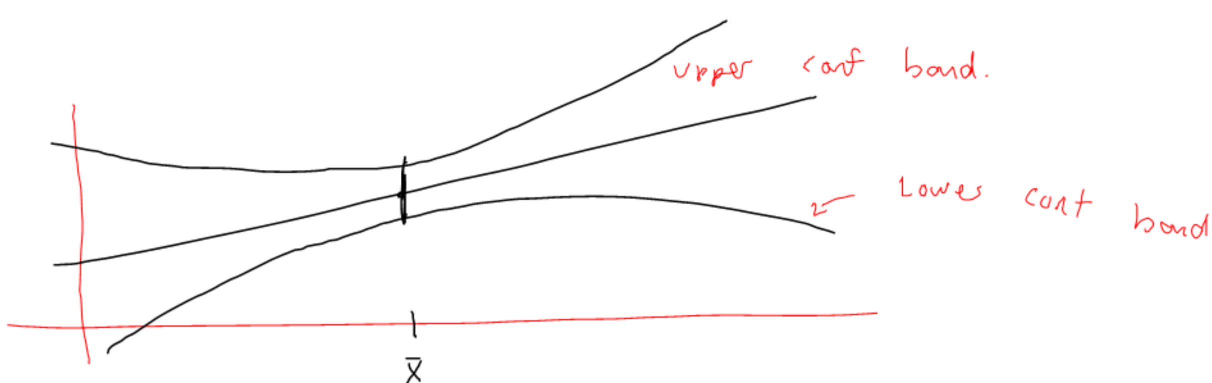
$$SE = \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)}$$

P.I.

$$SE = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)}$$

1) SE of C.I. < SE of P.E.

2)

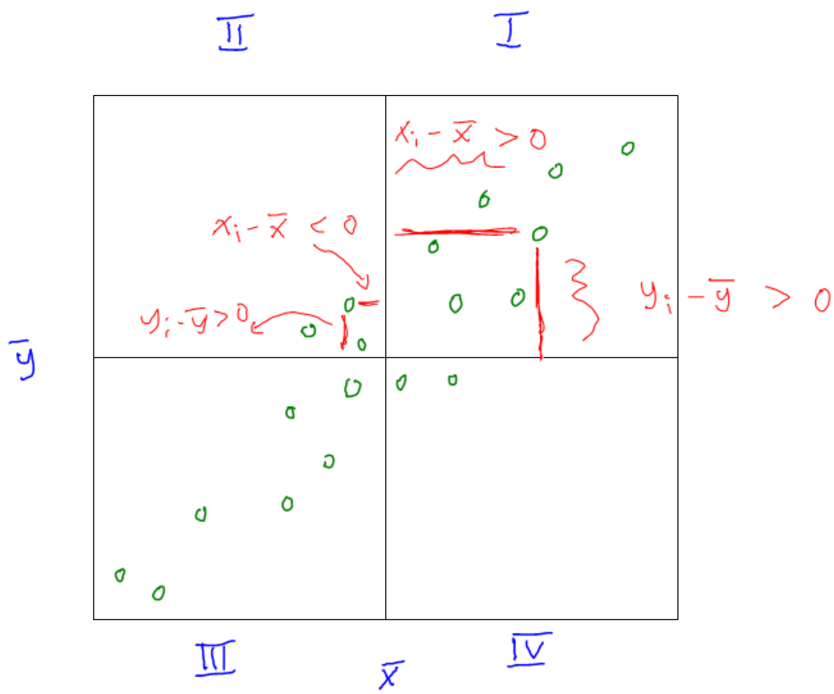


2.11 Linear Correlation

Linear correlation and covariance are both measures of the linear relationship between variables X and Y.

<u>Deviations</u>

- $y_i - \bar{y}$ = the **deviation** of each case y_i from the sample mean of the response variable $\bar{y}$.
- $x_i - \bar{x}$ = the **deviation** of each case x_i from the sample mean of the predictor variable $\bar{x}$.
- $(x_i - \bar{x})(y_i - \bar{y})$ = the product of the **deviations**.



Quadrant	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	+	+	+
2	+	−	−
3	−	−	+
4	−	+	−

#

Sample covariance

$$= \frac{1}{n-1} \Sigma (x_i - \bar{x})(y_i - \bar{y})$$