# Simple linear regression: prediction

# Prediction

estimation: for parameter
prediction: for random variable

Consider the regression model:

$$\boxed{Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)}$$

independently for $i = 1,2,\dots,n$.

How do we predict for an additional independent random variable:

$$\boxed{\underline{Y_0} \sim N(\beta_0 + \beta_1 \underline{x_0}, \sigma^2)?}$$

If $\beta_0, \beta_1, \sigma^2$ known:

point prediction $E[Y_0 \mid x = x_0] = \beta_0 + \beta_1 x_0$

interval prediction $\beta_0 + \beta_1 x_0 \pm Z_{\frac{\alpha}{2}} \sigma$

If $\beta_0, \beta_1, \sigma^2$ are unknown and estimated:
$\hat{\beta}_0 + \hat{\beta}_1 x_0$ is an estimator of $\beta_0 + \beta_1 x_0$.

# Theorem 9

Suppose $Y_1, Y_2, \ldots, Y_n$ are independent with

$$E[Y_i] = \beta_0 + \beta_1 x_i \text{ and } var(Y_i) = \sigma^2$$

then

1. $E[\hat{\beta}_0 + \hat{\beta}_1 x_0] = \beta_0 + \beta_1 x_0$
2. $var[\hat{\beta}_0 + \hat{\beta}_1 x_0] = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$
3. If, furthermore, $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, then

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

independently of $S_e^2$.

# Proof of Theorem 9

① From Theorem 8, we have $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$.

$$E[\hat{\beta}_0 + \hat{\beta}_1 x_0] = E[\hat{\beta}_0] + E[\hat{\beta}_1] x_0 = \beta_0 + \beta_1 x_0$$

② From Theorem 8, $var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$, $var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$, $cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$.

$$\begin{aligned}
var(\hat{\beta}_0 + \hat{\beta}_1 x_0) &= var(\hat{\beta}_0) + x_0^2 \, var(\hat{\beta}_1) + 2x_0 \, cov(\hat{\beta}_0, \hat{\beta}_1) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) + x_0^2\left(\frac{\sigma^2}{S_{xx}}\right) + 2x_0\left(-\frac{\sigma^2 \bar{x}}{S_{xx}}\right) \\
&= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} + \frac{x_0^2}{S_{xx}} - \frac{2x_0\bar{x}}{S_{xx}}\right] \\
&= \sigma^2\left[\frac{1}{n} + \frac{1}{S_{xx}}\left(\bar{x}^2 - 2x_0\bar{x} + x_0^2\right)\right] \\
&= \sigma^2\left[\frac{1}{n} + \frac{1}{S_{xx}}(x_0 - \bar{x})^2\right]
\end{aligned}$$

# Proof of Theorem 9

③ $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of $Y_i \sim N$

$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$ and $\hat{\beta}_0 = \sum_{i=1}^{n} b_i Y_i$ where $a_i = \frac{x_i - \bar{x}}{S_{xx}}$, $b_i = \frac{1}{n} - a_i \bar{x}$.

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = \sum_{i=1}^{n} b_i Y_i + \sum_{i=1}^{n} a_i Y_i x_0$$

$$= \sum_{i=1}^{n} \underbrace{\boxed{(b_i + a_i x_0)}}_{c_i} Y_i$$

$$= \sum_{i=1}^{n} c_i Y_i \qquad \text{where } c_i = b_i + a_i x_0$$

By Lemma 1, $\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0,\ \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$.

# Corollary 9

If $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ independently for $i = 1,2,\ldots,n$, then

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{S_e\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0,\ \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right)$$

So $$Z = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \sim N(0,1).$$

Recall $V = \frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}$.

$$T = \frac{Z}{\sqrt{\frac{V}{n-2}}} = \frac{\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\cancel{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}}}{\sqrt{\frac{\cancel{(n-2)} S_e^2}{\cancel{(n-2)}\cancel{\sigma^2}}}} = \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{S_e\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

# Confidence interval for $\beta_0 + \beta_1 x_0$

$$\underbrace{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right)}_{\hat{y}_0} \pm t_{n-2,\alpha/2}\, S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

# Prediction interval for $Y_0$

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{n-2,\alpha/2}\, S_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

CI: based on error of estimation $(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)$

PI: based on error of prediction $Y_0 - \hat{Y}_0 = (\beta_0 + \beta_1 x_0 + \varepsilon) - (\hat{\beta}_0 + \hat{\beta}_1 x_0)$

$= (\beta_0 + \beta_1 x_0) - (\hat{\beta}_0 + \hat{\beta}_1 x_0) \boxed{+\varepsilon}$

$Y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$

$\hat{Y}_0 \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$ $(\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0)$

Since $Y_0$ is independent of $Y_1, Y_2, \ldots, Y_n$, it is also independent of $\hat{Y}_0$

$Y_0 - \hat{Y}_0 \sim N(E(Y_0) - E(\hat{Y}_0), var(Y_0) + var(\hat{Y}_0))$

$= N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$

# Prediction interval for $Y_0$

$$Z = \frac{(Y_0 - \hat{Y}_0) - 0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} = \frac{Y_0 - \hat{Y}_0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0,1).$$

We estimate $\sigma^2$ with $S_e^2$. Recall $V = \frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}$.

Both $Y_0$ and $\hat{Y}_0$ are independent of $S_e^2$.

$$T = \frac{Z}{\sqrt{\frac{V}{n-2}}} = \frac{Y_0 - \hat{Y}_0}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

Using $T$ as a pivotal quantity for $Y_0$, we can write

$$\begin{aligned} 1-\alpha &= P\left(-t_{n-2,\frac{\alpha}{2}} \leq T \leq t_{n-2,\frac{\alpha}{2}}\right) \\ &= P\left(-t_{n-2,\frac{\alpha}{2}} \leq \frac{Y_0 - \hat{Y}_0}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \leq t_{n-2,\frac{\alpha}{2}}\right) \\ &= P\left(\hat{Y}_0 - t_{n-2,\frac{\alpha}{2}} S_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \leq Y_0 \leq \hat{Y}_0 + t_{n-2,\frac{\alpha}{2}} S_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right) \end{aligned}$$

# Confidence vs Prediction intervals

Although the confidence interval for $\beta_0 + \beta_1 x_0$ and the prediction interval for $Y_0$ are very similar in form, they are logically quite different.

- The confidence interval must be thought of as a randomly constructed interval that specifies the $100(1-\alpha)\%$ confidence limits for the fixed but unknown parameter.

- On the other hand, the prediction interval specifies the $100(1-\alpha)\%$ probability limits for the random variable $Y_0$

e.g. $Y$ = student mark in a test, $x$ = number of hours of study

At $x = 40$ hours, CI = (67, 80)

($\alpha = 0.05$) PI = (56, 92)

CI: We are 95% confident that the **mean test mark** for all students who studied 40 hours is between 67% and 80%.

PI: We are 95% confident that a **randomly chosen student** who studied 40 hours will score a mark between 56% and 92%.

# Example 3.2

Consider the Example 3.1 again. The fitted SLR model is

$$\hat{y} = \underset{\hat{\beta}_0}{-3.1011} + \underset{\hat{\beta}_1}{2.0266}\, x.$$

Other useful information are

$$\bar{x} = 3.8, S_{xx} = 263.6, S_e^2 = 0.9768, \; t_{8,0.025} \approx 2.306$$

a) Obtain the 95% confidence interval for $x = 5$.

b) Obtain the 95% prediction interval for $x = 5$.

$x_0 = 5$

$\hat{y}_0 = -3.1011 + 2.0266(5) \approx 7.6319$

# Example 3.2 Solution

a) $CI = \hat{y}_0 \pm t_{8,0.025}\, S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$

$\approx 7.6319 \pm 2.306 \sqrt{0.9768\left(\frac{1}{10} + \frac{(5-3.8)^2}{263.6}\right)}$

$\approx (6.2917,\ 7.7720)$

b) $PI = \hat{y}_0 \pm t_{8,0.025}\, S_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$

$\approx 7.6319 \pm 2.306 \sqrt{0.9768\left(1 + \frac{1}{10} + \frac{(5-3.8)^2}{263.6}\right)}$

$\approx (4.6356,\ 9.4281)$

# Example 3.2 Solution (using R)

```
#Example 3.1
x <- c(-1, 0, 2, -2, 5, 6, 8, 11, 12, -3)
y <- c(-5, -4, 2, -7, 6, 9, 13, 21, 20, -9)
lm1 <- lm(y~x)                #fit SLR
summary(lm1)                  #information about the fit
confint(lm1)                  #CI for model parameters (β0, β1)
                                (can also use confint.lm(lm1))

#Example 3.2
x0 <- data.frame(x=5)
predict(lm1, newdata = x0, interval = "confidence", level
= 0.95) #CI for x0

predict(lm1, newdata = x0, interval = "prediction", level
= 0.95) #prediction interval for x0
```

confint.default(lm1) uses normal approximation (ie. z-critical values)