Chapter 3 Predicted Mean Value versus Predicted Value

1. Predicted mean value of *Y*

Model: $Y = \beta_0 + \beta_1 X + \varepsilon \Leftrightarrow E(Y) = \beta_0 + \beta_1 X$.

For a given X value, say X_0, $Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0$. Let $\mu_0 = E(Y_0) = \beta_0 + \beta_1 X_0$. A point estimator of μ_0 is $\hat{\mu}_0 = b_0 + b_1 X_0$. — Predicted mean value at $X = X_0$.

Q: What is the distribution of $\hat{\mu}_0 = b_0 + b_1 X_0$

Distribution results 1: If $\varepsilon_i, i = 1, \cdots, n$, are $iid\ N(0, \sigma^2)$, then

$$\hat{\mu}_0 = b_0 + b_1 X_0 \sim N(\beta_0 + \beta_1 X_0, \sigma^2 [\tfrac{1}{n} + \tfrac{(X_0 - \bar{X})^2}{S_{XX}}]).$$

Notes: (1) $E(\hat{\mu}_0) = E(b_0 + b_1 X_0) = E(b_0) + E(b_1) X_0 = \beta_0 + \beta_1 X_0 = \mu_0$. Thus, $\hat{\mu}_0$ is an unbiased estimator of μ_0.

(2) The standard error of $\hat{\mu}_0$ is: $se(\hat{\mu}_0) = s\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}}$, which attains the minimum at $X_0 = \bar{X}$ and increases as X_0 is moved away from $\bar{X}$ in either direction. — Pay price for extrapolation.

A 100(1-α)% confidence interval (CI) for $\mu_0 = \beta_0 + \beta_1 X_0$ is

$$\hat{\mu}_0 \pm t_{n-2}\left(1 - \tfrac{\alpha}{2}\right) se(\hat{\mu}_0) = (b_0 + b_1 X_0) \pm t_{n-2}\left(1 - \tfrac{\alpha}{2}\right) s\sqrt{\tfrac{1}{n} + \tfrac{(X_0 - \bar{X})^2}{S_{XX}}}.$$

Considering X_0 as a variable, we obtain 100(1-α)% *confidence bands* for the regression line $E(Y) = \beta_0 + \beta_1 X$, which are hyperbolas.

2. Predicted value of *Y*

Want to predict the Y value at a given X value, say X_0. Let $Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0$.

Ex. 3.1 Observed gas price on each day of the last 30 days, want to predict tomorrow's gas price.

How to estimate $Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0$?

Estimate $\beta_0 + \beta_1 X_0$ by $b_0 + b_1 X_0$.
Estimate ε_0 by 0 sine $E(\varepsilon_0) = 0$.

$\Rightarrow$ A point estimate of $Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0$ is $\hat{Y}_0 = b_0 + b_1 X_0$.
— The predicted Y value at $X = X_0$.

However, variance of $\hat{Y}_0$ must incorporate both $Var(b_0 + b_1X_0)$ and $Var(\varepsilon_0)$, i.e.,

$$Var(\hat{Y}_0) = Var(b_0 + b_1X_0) + Var(\varepsilon_0) = \sigma^2[\tfrac{1}{n}+\tfrac{(X_0-\bar{X})^2}{S_{XX}}]] + \sigma^2 = \sigma^2[1 + \tfrac{1}{n}+\tfrac{(X_0-\bar{X})^2}{S_{XX}}].$$

The standard error of $\hat{Y}$ is: $se(\hat{Y}_0) = s\sqrt{1 + \frac{1}{n}+\frac{(X_0-\bar{X})^2}{S_{XX}}}$, which is always larger than $se(\hat{\mu}_0) = s\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{S_{XX}}}$.

A 100(1-α)% *prediction interval* (PI) for $Y_0 = \beta_0 + \beta_1X_0 + \varepsilon_0$ is

$$\hat{Y}_0 \pm t_{n-2}\left(1-\tfrac{\alpha}{2}\right)se(\hat{Y}_0) = (b_0 + b_1X_0) \pm t_{n-2}\left(1-\tfrac{\alpha}{2}\right)s\sqrt{1 + \tfrac{1}{n}+\tfrac{(X_0-\bar{X})^2}{S_{XX}}}.$$

Notes: (1) We cannot say that we are constructing a confidence interval for Y_0 since Y_0 is not a parameter.
(2) Considering X_0 as a variable, we obtain 100(1-α)% *prediction bands*.

Ex. 3.2 The National Institute of Health is studying the relationship between number of cigarettes smoked per day and birthweight of babies born to mothers who smoke cigarettes. The following data are observed.

No. of cigarettes per day(X)	21	12	28	10	24	5
Birthweight (Y)	6.0	8.0	5.6	7.5	6.2	8.5

(1) Obtain the equation of the least squares line and use it to predict the birthweight of a baby whose mother smokes 25 cigarettes per day. In addition, construct a 90% confidence interval for the true mean value of Y and a 90% prediction interval for Y when X =25.
(2) Construct a scatter plot for this data set and draw the fitted line on the plot. In addition, superimpose the 90% confidence bands for the regression line and the 90% prediction bands onto the scatter plot.