STAT 231 – Mar. 21, 2025

CONFIDENCE INTERVAL FOR $\mu(x) = \alpha + \beta x$

$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta} x$ (The invariance property of MLE)

$= \bar{Y} - \tilde{\beta}\bar{x} + \tilde{\beta} x$

$= \frac{1}{n}\sum_{i=1}^{n} Y_i + \tilde{\beta}(x - \bar{x})$

$= \frac{1}{n}\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i (x - \bar{x})$

$= \sum_{i=1}^{n} \left\{ \frac{1}{n} + \frac{(x_i - \bar{x})(x - \bar{x})}{S_{xx}} \right\} Y_i$

CONFIDENCE INTERVAL FOR $\mu(x) = \alpha + \beta x$

$$\Rightarrow \quad \tilde{\mu}(x) = \sum_{i=1}^{n} b_i Y_i$$

$\tilde{\mu}(x)$ is a combination of independent Gaussian random variables. So, $\tilde{\mu}(x)$ is also Gaussian distributed.

Confidence Interval for $\mu(x) = \alpha + \beta x$

$$\mathbb{E}(\hat{\mu}(x)) = \mathbb{E}\left(\sum_{i=1}^{n} b_i Y_i\right)$$

Recall $Y_i = \alpha + \beta x_i + R_i$,

$R_i \overset{iid}{\sim} G(0, \sigma)$

$$= \sum_{i=1}^{n} b_i \mathbb{E}(Y_i)$$

$$= \sum_{i=1}^{n} b_i (\alpha + \beta x_i)$$

$$= \alpha \sum_{i=1}^{n} b_i + \beta \sum_{i=1}^{n} b_i x_i$$

$$= \alpha + \beta x$$ (unbiased)

Exercise:

Show that

(1) $\sum_{i=1}^{n} b_i = 1$

(2) $\sum_{i=1}^{n} b_i x_i = x$

Confidence Interval for $\mu(x) = \alpha + \beta x$

$$Var(\tilde{\mu}(x)) = Var\left[\sum_{i=1}^{n} b_i Y_i\right]$$

$$= \sum_{i=1}^{n} b_i^2 Var(Y_i) + \cancel{\sum_{i \neq j} b_i b_j Cov(Y_i, Y_j)} \xrightarrow{} 0$$

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

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

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]$$

Exercise:

Show that

$$\sum_{i=1}^{n} b_i^2 = \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}$$

CONFIDENCE INTERVAL FOR $\mu(x) = \alpha + \beta x$

In Summary:

$$\tilde{\mu}(x) \sim G\left(\overbrace{\alpha+\beta x}^{\mu(x)},\ \sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}\right).$$

(Exact Distribution)

Pivotal Quantity:

$$\frac{\tilde{\mu}(x) - \mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2).$$

$$\frac{\tilde{\mu}(x) - \mu(x)}{\sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1)$$

$$\boxed{S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \tilde{\mu}_i)^2}$$

$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2)$$

Then,

Using theorem from Ch.4

$$T = \frac{\tilde{\mu}(x) - \mu(x)}{\sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \Bigg/ \sqrt{\frac{S_e^2}{\sigma^2}} \sim t(n-2).$$

Confidence Interval for $\mu(x) = \alpha + \beta x$

A $100p$ % CI for $\mu(x)$ is

$$\hat{\mu}(x) \pm a \cdot Se\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

where $P(T \le a) = \frac{1+p}{2}$, $T \sim t(n-2)$.

Prediction Interval for Future Response

To predict a future response Y that is independent of our sample (x_i, Y_i), $i = 1, \cdots, n$.

Model: $Y = \alpha + \beta x + R$, $R \overset{iid}{\sim} G(0, \sigma)$.

$Y \perp Y_i$, $i = 1, \cdots, n$.

We predict Y using $\tilde{\mu}(x)$.

Prediction Interval for Future Response

We consider $Y - \tilde{\mu}(x)$

$$E(Y - \tilde{\mu}(x)) = E(Y) - E(\tilde{\mu}(x))$$

$$= \alpha + \beta x - (\alpha + \beta x)$$

$$= 0$$

$$\mathrm{Var}(Y - \tilde{\mu}(x)) = \mathrm{Var}(Y) + \mathrm{Var}(\tilde{\mu}(x)) - \cancelto{0}{2\,\mathrm{Cov}(Y, \tilde{\mu}(x))}$$

$$= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right].$$

Prediction Interval for Future Response

Note $Y - \tilde{\mu}(x)$ is Gaussian distributed.

$$\frac{Y - \tilde{\mu}(x)}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1)$$

$$\frac{Y - \tilde{\mu}(x)}{se\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2).$$

PREDICTION INTERVAL FOR FUTURE RESPONSE

A 100p% Prediction Interval for Y

$$\hat{\mu}(x) \pm a \cdot se\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}},$$

where $\mathbb{P}(T \le a) = \frac{1+p}{2}$, $T \sim t(n-2)$.

6.3 Model Checking - Assumptions to be Checked

Homoscedasticity

There are two main assumptions for Gaussian linear response models:

(1) Y_i (given covariates x_i) is Gaussian with standard deviation σ which does not depend on the covariates.

$Y_i = \alpha + \beta x_i + R_i$, $R_i \overset{ind}{\sim} G(0, \sigma)$.

(2) $E(Y_i) = \mu(x_i)$ is a linear combination of observed covariates with unknown coefficients.

$E(Y_i) = \mu(x_i) = \alpha + \beta x_i$, $i = 1, \cdots, n$.

MODEL ASSUMPTIONS SHOULD ALWAYS BE CHECKED!!!

We will examine three graphical methods to check these assumptions.

Method I - Scatterplot of Data and Fitted Regression Line

In simple linear regression, a scatterplot of the data with the fitted line $y = \widehat{\alpha} + \widehat{\beta}x$ superimposed shows how well the model fits. If there are any obvious departures from the fitted line then these departures might suggest a model which would fit the data better.