8 oct - LEC 3

PREDICTION OF THE RESPONSE VARIABLE

We observe (x_i, y_i) for $i = 1, \dots, n$.

Consider an additional unit observed at a value x_*. We want to make a prediction about the value of the response variable corresponding to x_*.

The model is $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, i.e. $\mathbb{E}[Y_i] = \mu_i = \beta_1 + \beta_2 x_i$

hence $Y_* = \beta_1 + \beta_2 x_* + \varepsilon_*$ with $\mu_* = \beta_1 + \beta_2 x_*$

The predicted value is $\hat{y}_* = \hat{\beta}_1 + \hat{\beta}_2 x_*$

The prediction $\hat{y}_*$ corresponds to the estimate of the mean μ_*.

If we consider the estimators $\hat{B}_1$ and $\hat{B}_2$, we obtain the corresponding estimator $\hat{M}_*$ of the mean of Y_*.

We can study the distribution of $\hat{M}_*$.

$$\hat{M}_* = \hat{B}_1 + \hat{B}_2 x_* = \bar{Y} - \hat{B}_2 \bar{x} + \hat{B}_2 x_* = \bar{Y} + \hat{B}_2 (x_* - \bar{x})$$

$$= \frac{1}{n}\sum_{i=1}^n Y_i + (x_* - \bar{x}) \sum_{i=1}^n w_i Y_i \qquad \text{since } \hat{B}_2 = \sum_{i=1}^n w_i Y_i \text{ with } w_i = \frac{(x_i - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2}$$

$$= \sum_{i=1}^n \left(\frac{1}{n} + (x_* - \bar{x}) w_i \right) Y_i$$

$\Rightarrow$ $\hat{M}_*$ is a linear combination of $Y_1, \dots, Y_n$

$\Rightarrow$ $\hat{M}_*$ has normal distribution $\quad \hat{M}_* \sim N(\dots, \dots)$ → we need to find the mean and variance

$$\mathbb{E}[\hat{M}_*] = \mathbb{E}[\hat{B}_1 + \hat{B}_2 x_*] \overset{\text{linearity}}{=} \beta_1 + \beta_2 x_* = \mu_* \qquad \text{unbiased}$$

homoscedastic $var(Y_i) = \sigma^2 \ \forall i$

$$var(\hat{M}_*) = var\left(\sum_{i=1}^n \left(\frac{1}{n} + (x_* - \bar{x}) w_i \right) Y_i \right) \overset{\perp\!\!\!\perp}{=} \sum_{i=1}^n \left(\frac{1}{n} + (x_* - \bar{x}) w_i \right)^2 \sigma^2 =$$

$$= \sum_{i=1}^n \left(\frac{1}{n^2} + w_i^2 (x_* - \bar{x})^2 + \frac{2}{n} w_i (x_* - \bar{x}) \right) \sigma^2 =$$

$$= \frac{1}{n}\sigma^2 + \sigma^2 (x_* - \bar{x})^2 \sum_{i=1}^n w_i^2 + \cancel{2\sigma^2 (x_* - \bar{x}) \underbrace{\sum_{i=1}^n w_i}_{=0}} =$$

$$= \sigma^2 \left(\frac{1}{n} + (x_* - \bar{x})^2 \underbrace{\frac{1}{\sum_{k=1}^n (x_k - \bar{x})^2}}_{\sum_{i=1}^n w_i^2} \right) =$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

$$\Rightarrow \hat{M}_* \sim N\left(\mu_* ; \underbrace{\sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}_{V(\hat{M}_*)} \right) = N(\mu_*, V(\hat{M}_*))$$

Let's derive a confidence interval for μ_*

We need a pivotal quantity

$$\frac{\hat{M}_* - \mu_*}{\sqrt{V(\hat{M}_*)}} \sim N(0,1)$$

since $V(\hat{M}_*)$ involves the unknown σ^2, similarly to what we have done for $\hat{B}_j$ we substitute $V(\hat{M}_*)$ with $\hat{V}(\hat{M}_*)$, obtaining

$$\frac{\hat{M}_* - \mu_*}{\sqrt{\hat{V}(\hat{M}_*)}} \sim t_{n-2} \qquad \text{where} \qquad \hat{V}(\hat{M}_*) = S^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

Thus, a confidence interval of level $1-\alpha$ for μ_* is obtained as

$$1-\alpha = \mathbb{P}\left(-t_{n-2; 1-\frac{\alpha}{2}} < \frac{\hat{M}_* - \mu_*}{\sqrt{\hat{V}(\hat{M}_*)}} < t_{n-2; 1-\frac{\alpha}{2}} \right)$$

$$1-\alpha = \mathbb{P}\left(\hat{M}_* - t_{n-2; 1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{M}_*)} < \mu_* < \hat{M}_* + t_{n-2; 1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{M}_*)} \right)$$

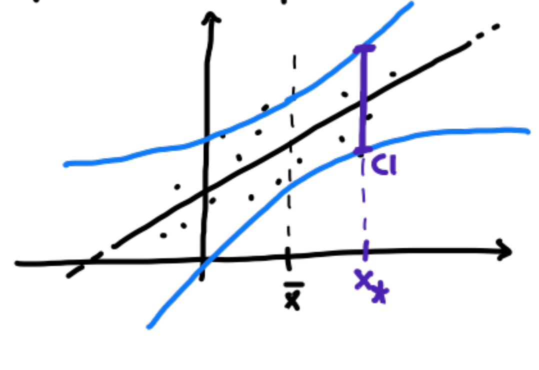
$$1-\alpha = \mathbb{P}\left(\hat{M}_* - t_{n-2; 1-\frac{\alpha}{2}} \sqrt{S^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)} < \mu_* < \hat{M}_* + t_{n-2; 1-\frac{\alpha}{2}} \sqrt{S^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)} \right)$$

conditioning now to the observed data: $\hat{y}_*$ estimate of μ_*, s^2 estimate of σ^2

$$\text{CI}: \quad \hat{y}_* \pm t_{n-2; 1-\frac{\alpha}{2}} \sqrt{s^2 \left(\frac{1}{n} + \underbrace{\frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}_{} \right)}$$

notice that the further x_* is from $\bar{x}$, the larger the CI will get

If I compute several pointwise CIs for varying x_*, I obtain "confidence bands"



(careful: the level $(1-\alpha)$ only holds pointwise)

→ why predicting outside of the range of the x_i is dangerous

These methods can be useful to formalize practical questions, for example:

- what is a reasonable set of values for Y if $x = \tilde{x}$? → compute CI for $\tilde{\mu}$
- is μ_0 a reasonable value for Y if I observe $x = \tilde{x}$? → test $H_0: \tilde{\mu} = \mu_0$

  $H_1: \tilde{\mu} \neq \mu_0$