ECON 203, CHALLENGE QUIZ 7, PART TWO

CHARLES ANCEL

Consider the same setup as in the previous question. Show that

$$Y|X = x = \widehat{m}(x) + \mathsf{N}\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{ns_X^2}\right]\right).$$

Proof. Consider the linear regression model $Y_i = \beta_0 + \beta_1 X_i + U_i$ where $U_i \sim \mathsf{N}(0, \sigma^2)$. The estimated regression function at a point x is given by $\widehat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$.

We aim to show that the conditional distribution of Y given $X = x$ is normally distributed with mean $\widehat{m}(x)$ and variance $\sigma^2 \left[1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{ns_X^2}\right]$.

The key steps involve:

Step 1: Establishing the mean of the conditional distribution $Y|X = x$ as $\widehat{m}(x)$, leveraging the linearity and unbiasedness of the OLS estimators.

Step 2: The variance of Y_i can be decomposed as:

$$\begin{aligned} Var(Y_i) &= Var(\beta_0 + \beta_1 X_i + U_i) \\ &= Var(\beta_0) + Var(\beta_1 X_i) + Var(U_i) + 2Cov(\beta_0, \beta_1 X_i) + 2Cov(\beta_0, U_i) + 2Cov(\beta_1 X_i, U_i) \end{aligned}$$

Since β_0 and β_1 are constants, their variances are zero, and $Cov(\beta_0, U_i) = Cov(\beta_1 X_i, U_i) = 0$ due to the independence of errors from the regressors and the constants.

Step 3: We then detail the derivation of each variance component:

- $Var(U_i) = \sigma^2$ directly from the model assumption.
- Deriving $\frac{\sigma^2}{n}$ involves analyzing the variance introduced by estimating β_0 and β_1, which in turn depends on the sample size and the variability of X_i.
- The term $\frac{\sigma^2(x-\bar{X})^2}{ns_X^2}$ is derived by considering the additional variance introduced when predicting Y for a given $X = x$, not at the mean $\bar{X}$.

Step 4: Combining these, the total variance of $Y|X = x$ is $\sigma^2 \left[1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{ns_X^2}\right]$.

Step 5: Thus, the conditional distribution of Y given $X = x$ is expressed as $Y|X = x = \widehat{m}(x) + \mathsf{N}\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{ns_X^2}\right]\right)$, illustrating that Y given $X = x$ is normally distributed around the regression line with specified variance.

This proof shows the detailed derivation of the conditional distribution's variance, providing a comprehensive understanding of its components and their origins. $\square$