Show that E[$\hat{\epsilon}$] = 0 and determine Cov($\hat{\epsilon}$) for the linear regression problem with $\hat{\beta} = (X^T X)^{-1} X^T y$

Solution:

We know,

$y = X\beta + \epsilon$ ……. (1)

From (1)

$$E[y] = E[X\beta + \epsilon]$$
$$= X\beta + E[\epsilon]$$
$$= X\beta \quad \text{……..} \quad (2)$$

From (1) we can write that,

$$\hat{\epsilon} = y - X\hat{\beta}$$
$$\hat{\epsilon} = y - X\,(X^T X)^{-1} X^T y \text{ [we know } \hat{\beta} = (X^T X)^{-1} X^T y \text{]}$$
$$\hat{\epsilon} = y - Hy \text{ [we know H = } X\,(X^T X)^{-1} X^T \text{] ……..…. (3)}$$

Now from (3),

$$E[\hat{\epsilon}] = E[y - Hy]$$
$$= E[y] - HE[y]$$
$$= X\beta - X\,(X^T X)^{-1} X^T\, X\beta \text{ [using (2)]}$$
$$= X\beta - (XX^T)\,(X^T X)^{-1}\, X\beta$$
$$= X\beta - X\beta$$
$$= 0$$

Comments:

- $E[\hat{\epsilon}] = 0$, i.e., the average residuals should be zero. The residuals, like the error terms, have an expectation of zero. The average value of the error term must equal zero for the model to be unbiased [2].

Now,

$$Cov(\beta) = E[((X^TX)^{-1}X^T \in)((X^TX)^{-1}X^T \in)^T]$$

$$= E[(X^TX)^{-1}X^TX \in\in^T (X^TX)^{-1}]$$

$$= (X^TX)^{-1}X^TX(X^TX)^{-1}E[\in\in^T]$$

$$= \underbrace{(X^TX)^{-1}X^T}_{A}\underbrace{X(X^TX)^{-1}}_{A^T}\sigma^2 \; [E[\in\in^T] = \sigma^2 I, here\ I\ is\ the\ identity\ m * m\ matrix]$$

$$= \sigma^2 A A^T \text{ [Taking A = } (X^TX)^{-1}X^T] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

Using (3) we can write that,

$$Cov(\hat{\in}) = Cov(y - Hy)$$

$$= Cov((I_n - H)y)$$

$$= \sigma^2\big((I_n - H)y\big)\big((I_n - H)y\big)^T \text{ [Using (4)]}$$

$$= \sigma^2\big((I_n - H)yy^T\big)(I_n - H)^T$$

$$= \sigma^2 I_n(I_n - H)(I_n - H)^T$$

$$= \sigma^2 (I_n - H)$$

[for symmetric $(I_n - H) = (I_n - H)^T$ and for idempotent $(I_n - H)(I_n - H) = (I_n - H)$] [1]

Comments:

❖ $Cov(\hat{\in}) = \sigma^2 (I_n - H) = \sigma^2 (I_n - X(X^TX)^{-1}X^T)$, So, the residuals are correlated. But in error terms $Cov(\in) = \sigma^2 I_n$ and it is uncorrelated.

Proof:

The equation of the regression line is:

$$\hat{Y} = \hat{a}X + \hat{b}$$

Here, $\hat{Y}$ is the regression prediction of Y based on X, $\hat{a}$ slop of the regression line, $\hat{a} = \frac{E(D_X D_Y)}{\sigma_X{}^2} = \frac{r\sigma_X\sigma_Y}{\sigma_X{}^2} = r\frac{\sigma_Y}{\sigma_X}$, here r is correlation between two variables.

The intercept of the regression line is $\hat{b} = \mu_Y - \hat{a}\mu_X$.

Now, $\hat{Y} = \hat{a}X + \mu_Y - \hat{a}\mu_X = \hat{a}(X - \mu_X) + \mu_Y$

Residual is defined as: $D = Y - \hat{Y}$

Now, $D = Y - (\hat{a}(X - \mu_X) + \mu_Y) = (Y - \mu_Y) - \hat{a}(X - \mu_X) = D_Y - \hat{a}D_X$

By the definition of correlation:

$$r(D,X) = E\left(\left(\frac{D - \mu_D}{\sigma_D}\right)\left(\frac{X - \mu_X}{\sigma_X}\right)\right)$$
$$= \frac{1}{\sigma_D \sigma_X} E(DD_X)$$

Because, $\mu_D = 0$. Therefore to show $r(D,X) = 0$, we just have to show that $E(DD_X) = 0$.

Now,

$$E(DD_X) = E\big((D_Y - \hat{a}D_X)D_X\big)$$
$$= E(D_X D_Y) - \hat{a}E\big({D_X}^2\big)$$
$$= r\sigma_X\sigma_Y - r\frac{\sigma_Y}{\sigma_X}{\sigma_X}^2 \; [E(D_X D_Y) = E((X - \mu_X)(Y - \mu_Y)) = r\sigma_X\sigma_Y\,]$$
$$= 0$$

As $E(DD_X) = 0$; so, we can say that residuals are correlated.

On the other hand,

We know, error terms $\epsilon = y - X\beta$

So,

$$r(\epsilon, Z) = E\left(\left(\frac{\epsilon - \mu_\epsilon}{\sigma_\epsilon}\right)\left(\frac{Z - \mu_Z}{\sigma_Z}\right)\right)$$
$$= \frac{1}{\sigma_\epsilon \sigma_Z} E(\epsilon Z)$$

$$E(\epsilon Z) = E((y - \beta X)Z)$$
$$= E(yZ) - \beta E(XZ)$$
$$= r\sigma_y\sigma_Z - r^2\sigma_X\sigma_Z$$
$$= r(\sigma_y - r\sigma_X)\sigma_Z$$

As $E(\epsilon Z) \neq 0$; so, we can say that error terms are uncorrelated.

❖ One of the assumptions of classical linear regression is that the error terms conditional on different X values all have the same variance, i.e., $\sigma_i^2 = \sigma_j^2$ for any X_i and X_j. This assumption, known as homoscedasticity, may or may not be met for a particular model applied to a particular population. Before drawing conclusions from ordinary least squares (OLS) regression it is good practice to apply appropriate tests to assess whether this assumption is met. Where the assumption is met, we are justified in using a common symbol, usually σ^2, for the common variance of the error terms [4].

The variance of residuals is:

$$Var(D) = E(D^2)$$

$$= E((D_Y - \hat{a}D_X)^2)$$

$$= E(D_Y{}^2) - 2\hat{a}E(D_X D_Y) + \hat{a}E(D_X{}^2)$$

$$= \sigma_Y{}^2 - 2r\frac{\sigma_Y}{\sigma_X} r\sigma_X\sigma_Y + r^2\frac{\sigma_Y{}^2}{\sigma_X{}^2}\sigma_X{}^2 \; [E(D_X D_Y) = r\sigma_X\sigma_Y \; and \; \hat{a} = r\frac{\sigma_Y}{\sigma_X}]$$

$$= \sigma_Y{}^2 - 2r^2\sigma_Y{}^2 + r^2\sigma_Y{}^2$$

$$= \sigma_Y{}^2 - r^2\sigma_Y{}^2$$

$$= (1 - r^2)\sigma_Y{}^2$$

i.e., the residuals have heteroscedastic variances (in contrast to the error terms ϵ_i) [3].

❖ If the residuals are normally distributed, we are able to derive the distribution of the residuals:

$$\hat{\epsilon} \sim N(0, \sigma^2(I_n - H))$$

But for error term, $\epsilon \sim N(0, \sigma^2 I_n)$

Reference:

[1]. https://www.stat.purdue.edu/~boli/stat512/lectures/topic3.pdf, Page 8.
[2]. https://statisticsbyjim.com/regression/ols-linear-regression-assumptions/
[3]. http://stat88.org/textbook/notebooks/Chapter_11/05_The_Error_in_Regression.html
[4]. https://stats.stackexchange.com/questions/48553/linear-regression-variance-error-term