Biostat 250B HW1

Elvis Cui

January 15, 2021

1 Commentary

Recall the classical linear regression model:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \mathbf{W})$$

As practitioners, we would like to estimate β based on observed data and address the goodness-of-fit based on some measurements such as **coefficient of determinants**.

There are 2 common ways to estimate β:

- Ordinary Least Squares (OLS): $\widehat{\beta}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
- Weighted Least Squares (WLS): $\widehat{\beta}_{WLS} = (\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}\mathbf{y}$.

For the second one, usually we write it in the **transformed** version, i.e.,

$$\widehat{\beta}_{WLS} = (\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_*^T\mathbf{y}_*$$

where

$$\mathbf{X}_* = \mathbf{W}^{-1/2}\mathbf{X},\ \mathbf{y}_* = \mathbf{W}^{-1/2}\mathbf{y}$$

Then, the coefficient of determination can be calcualted in three ways:

- $$R^2_{OLS} = 1 - \frac{\|\mathbf{y} - \mathbf{X}\widehat{\beta}_{OLS}\|_2^2}{\mathbf{y^T y} - n\bar{y}^2}$$
- $$R^2_{WLS} = 1 - \frac{\|\mathbf{y}_* - \mathbf{X}_*\widehat{\beta}_{WLS}\|_2^2}{\mathbf{y}_*^{\mathbf{T}}\mathbf{y}_* - n\bar{y}_*^2}$$
- $$\text{pseudo } R^2_{WLS} = 1 - \frac{\|\mathbf{y} - \mathbf{X}\widehat{\beta}_{WLS}\|_2^2}{\mathbf{y^T y} - n\bar{y}^2}$$

The first two are more intuitive than the last one. While usually the second one (R^2_{WLS}) is higher than the first one (R^2_{OLS}), the third one (pseudo R^2_{WLS}) is always less than the first one. Therefore, "sole reliance on the **coefficient of determination** may fail to reveal important data characteristics and model inadequacies".