Computing Variance of Predictions in Linear Regression

Suppose we have independent variables

$$\{x_1, \dots, x_n\}$$

and dependent variables

$$\{y_1, \dots, y_n\}$$

Recall that linear regression uses predictions

$$\hat{y} = \overline{y} + r\frac{SD(y)}{SD(x)}(x - \overline{x})$$

Here

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \,, \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$SD(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2, \quad SD(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{SD(x)\cdot SD(y)}$$

Remember that we can use may expressions of the form

$$\hat{y} = a + bx$$

However we choose parameters a for the intercept and b for the slope with the average loss

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

We want to minimize the average loss by choosing some $\hat{a}$ and $\hat{b}$. We can use the derivative to determine where the average loss attains the minimum.

Rate of change with respect to a: $$-\frac{2}{n}\sum_{i=1}^{n} y_i - (a + bx_i) = 0$$

Rate of change with respect to b: $$-\frac{2}{n}\sum_{i=1}^{n}(y_i - (a + bx_i))\,x_i = 0$$

Remember that we call the errors

$$y_i - \hat{y}_i = y_i - (a + bx_i)$$

between observed values and predicted values the residuals. We can interpret these equations

Rate of change with respect to a: The average of the residuals $= 0$

Rate of change with respect to b: The covariance of the residuals and independent variables $= 0$

We want both expressions to be 0 because the rate of change should be 0 at the minimum. By setting both expressions equal to 0, we have two equations and two unknowns.

These equations allow us to solve for $\hat{a}$ and $\hat{b}$. Additionally these equations help us to relate the variance of $\{y_1, \dots, y_n\}$ to the variance of $\{\hat{y}_1, \dots, \hat{y}_n\}$. Note

$$\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$$

$$= \frac{1}{n}\sum_{i=1}^{n}\overline{y} + \frac{r\,SD(y)}{SD(x)}(x_i - \overline{x})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\overline{y} + \frac{1}{n}\sum\frac{r\,SD(y)}{SD(x)}(x_i - \overline{x})$$

$$= \overline{y} + 0$$

So we can calculate

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$$

use $\overline{y} = \overline{\hat{y}}$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{\hat{y}})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$$

We can simplify

$$\sum_{i=1}^{n}2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum_{i=1}^{n}2(y_i - \hat{y}_i)\left[\frac{r\,SD(y)}{SD(x)}(x_i - \overline{x})\right]$$

$$= \frac{2r\,SD(y)}{SD(x)}\sum_{i=1}^{n}(y_i - \hat{y}_i)(x_i - \overline{x})$$

with

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(x_i - \overline{x}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)x_i - \overline{x}\sum_{i=1}^{n} y_i - \hat{y}_i$$

$$= \underbrace{0}_{\text{use sum of residuals is 0}} + \underbrace{0}_{\text{use covariance of residuals and independent variables is 0}}$$

Therefore

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{\hat{y}})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{\hat{y}})^2$$

We conclude that

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{Variance of observed values}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}_{\text{mean square error}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}_{\text{Variance of predicted values}}$$

Note that the variance of predicted values is less than the variance of observed values. This is called regression to the mean.