Variance of Residuals

C. Durso

Distributions of Regression Quantities

In the notation of Regression I slides,

- $\hat{m}$ distributed as $M = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$
- $\hat{b}$ distributed as $B = \bar{Y} - M\bar{X} = m\bar{X} + b + \frac{1}{n}\sum \varepsilon_i - M\bar{X}$

Summary of Distributions

Theorem

Given n observations from $Y_i = mX_i + b + \varepsilon_i$ where $\varepsilon_i \sim normal(0, \sigma^2)$:

- $\hat{m} \sim normal\left(m, \frac{\sigma^2}{\sum (X_i - \bar{X})^2}\right)$
- $\hat{b} \sim normal\left(b, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right)\right)$
- $\hat{Y}_h \sim normal\left(mX_h + b, \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right)\right)$
- $\hat{Y}_h new \sim normal\left(mX_h + b, \sigma^2\left(\frac{n+1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right)\right)$
    - Estimation of σ by s results in Student's t distributions with $n-2$ degrees of freedom

Residual Formulas

The i^{th} residual, $e_i = y_i - \hat{y}_i$ is distributed as $Y_i - \hat{Y}_i$.

- The variance is $Var\left[Y_i - \hat{Y}_i\right]$
- $Var[Y_i] = \sigma^2$
- $Var\left[\hat{Y}_i\right] = \sigma^2 \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum\left(X_j - \bar{X}\right)^2}\right)$
- $\hat{Y}_i = (MX_i + \bar{Y} - M\bar{X})$

Covariance of Y_i and $\widehat{Y}_i$

$$Cov\left[Y_i, \widehat{Y}_i\right] = Cov[Y_i, MX_i + \bar{Y} - M\bar{X}]$$

$$Cov[Y_i, M(X_i - \bar{X})] + Cov[Y_i, \bar{Y}\ ]$$

$$= (X_i - \bar{X})Cov[Y_i, M] + \frac{\sigma^2}{n}$$

$$= (X_i - \bar{X})Cov\left[Y_i, \frac{\sum\left(X_j - \bar{X}\right)\left(Y_j - \bar{Y}\right)}{\sum\left(X_j - \bar{X}\right)^2}\right] + \frac{\sigma^2}{n}$$

$$= (X_i - \bar{X})Cov\left[Y_i, \frac{\sum\left(X_j - \bar{X}\right)Y_j}{\sum\left(X_j - \bar{X}\right)^2}\right] + \frac{\sigma^2}{n}$$

$$= \frac{(X_i - \bar{X})^2\sigma^2}{\sum\left(X_j - \bar{X}\right)^2} + \frac{\sigma^2}{n}$$

Variance of e_i

$$Var\left[Y_i - \hat{Y}_i\right] = Var[Y_i] + Var\left[\hat{Y}_i\right] - 2Cov\left[Y_i, \hat{Y}_i\right]$$

$$= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum\left(X_j - \bar{X}\right)^2}\right) - 2\left(\frac{\sigma^2}{n} + \frac{(X_i - \bar{X})^2\sigma^2}{\sum\left(X_j - \bar{X}\right)^2} + \frac{\sigma^2}{n}\right)$$

$$= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum\left(X_j - \bar{X}\right)^2}\right) - 2\sigma^2\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum\left(X_j - \bar{X}\right)^2}\right)$$

$$= \sigma^2 - \sigma^2\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum\left(X_j - \bar{X}\right)^2}\right)$$

Studentized Residual

Dividing e_i by $\sqrt{Var\left[Y_i - \hat{Y}_i\right]}$ with σ^2 approximated by s^2 results in a random variable with a Student's t distribution with $n - 2$ degrees of freedom, the *studentized residual*

Relation to Leverage

Definition

The leverage of the i^{th} case in regression with a single explanatory variable X equals $\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_j - \bar{X})^2}$

Conclude that the denominator $\sqrt{Var[Y_i - \hat{Y}_i]}$ used to standardize or studentize e_i is equal to $\sqrt{\sigma^2(1 - leverage(i))}$.

Cook’s Distance

C. Durso

Cook's Distance, a Formula

Definition

Let $\boldsymbol{y} = (y_1, y_2, \dots y_n)$ and $\boldsymbol{x} = (x_1, x_2, \dots x_n)$. From the least squares regression $\boldsymbol{y} = \hat{m}\boldsymbol{x} + \hat{b}$, define $s^2 = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n-1}$ where $\hat{y}_j = \hat{m}x_j + \hat{b}$. Cook's Distance D_i for the i^{th} observation (x_i, y_i) equals $\frac{1}{2s^2}\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2$ where $\hat{y}_{j(i)}$ equals the predicted value of y_j in a least squares regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ omitting the observation (x_i, y_i).