Residuals and Model Diagnosis (§ 9. Applied linear Regression book)

① Linear regression Regression (p predictors)

Review

✓ For $i = 1, \cdots, n$.

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + e_i$$

$$= x_i^T \beta + e_i$$

where $\underset{(p+1)\times 1}{x_i} = \begin{pmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,p} \end{pmatrix}$, $\underset{(p+1)\times 1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$

$$\Rightarrow \quad Y = X\beta + e$$

where $\underset{n\times 1}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\underset{n\times(p+1)}{X} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$, $\underset{n\times 1}{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$

✓ Assumptions on statistical error (e)

(1) $E[\underset{n\times 1}{e} \mid X] = 0$

(2) $Var(e \mid X) = \sigma^2 I_n$

(Equivalently, for $i = 1, \cdots, n$: $E[e_i \mid X] = 0$, $Var(e_i \mid X) = \sigma^2$, and e_i's are uncorrelated (independent))

Note that Assumption (1) implies that for $i = 1, \cdots, n$

$$E[e_i x_i] = 0 \quad \text{and} \quad Cov(e_i, x_i) = 0$$

This is because

✓ (1) $\Rightarrow E[e_i \mid X] = 0 \Rightarrow E[e_i x_i \mid X] = 0$

$\Rightarrow E[e_i x_i] = 0$ (since $E[E(e_i x_i \mid X)]$)

✓ $Cov(e_i, x_i) = \underbrace{E(e_i x_i)}_{=0} - \underbrace{E e_i \cdot E x_i}_{=0} = 0.$

✓ For OLS estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$. (Assuming $(X^T X)^{-1}$ exists)

From previous lecture, we have

$$E(\hat{\beta} \mid X) = \beta, \qquad \underbrace{Var(\hat{\beta} \mid X)}_{(p+1)\times(p+1) \text{ matrix}} = \sigma^2 (X^T X)^{-1}.$$

($\hat{\beta}$ is an unbiased estimator of β)

(2.) Properties of Residuals $\underset{n\times 1}{\hat{e}} = \begin{pmatrix} \hat{e}_1 \\ \vdots \\ \hat{e}_n \end{pmatrix} = \underset{n\times 1}{y} - \underset{n\times 1}{\hat{y}} = \begin{pmatrix} y_1 - x_1^T\hat{\beta} \\ \vdots \\ y_n - x_n^T\hat{\beta} \end{pmatrix}$

Prop:

(2.1) sample version. ✓ Sample mean of residuals $= 0$ (from earlier lecture)

$$\bar{\hat{e}} \triangleq \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i = 0$$ (can proved using normal equations)

✓ Normal equations: $X^T\hat{e} = 0_{(p+1)\times 1}$

$$\Leftrightarrow \underbrace{(x_1\ x_2\ \cdots\ x_n)}_{\downarrow}\begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \vdots \\ \hat{e}_n \end{pmatrix} = 0$$

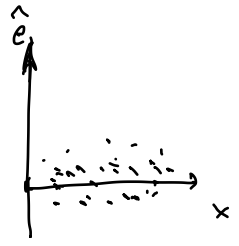
$$\Leftrightarrow \sum_{i=1}^{n}\hat{e}_i x_i = 0_{(p+1)\times 1} \quad (*)$$

- Prop: The Sample covariance between predictors $x_1, \ldots, x_n$ and residuals $\hat{e}_1, \ldots, \hat{e}_n$ $= 0$.

proof: Sample covariance $(x, \hat{e})$ $= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(\hat{e}_i - \bar{\hat{e}})$

$\bar{x}$: Sample mean of x $= \frac{1}{n}\sum_{i=1}^{n}x_i$

$\bar{\hat{e}}$: Sample mean of $\hat{e}$ $= \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i = 0$



if linear reg. Assumptions hold

$$= \frac{1}{n-1}\left[\underbrace{\sum_{i=1}^{n}x_i\hat{e}_i}_{=0 \text{ by } (*)} - \underbrace{\sum_{i=1}^{n}\bar{x}\hat{e}_i}_{\bar{x}\sum_{i=1}^{n}\hat{e}_i = 0}\right]$$

$$= 0.$$

- Prop: The sample covariance between fitted values $\hat{y}_1 \ldots \hat{y}_n$ and residuals $\hat{e}_1 \ldots \hat{e}_n$

if linear reg Assumptions hold.

Proof: Sample cov of $\hat{y}$ and $\hat{e}$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(\hat{e}_i - \bar{\hat{e}}), \quad \bar{\hat{e}} = 0$$

$\bar{\hat{y}}$ = sample mean of $\hat{y}_1 \cdots \hat{y}_n$ $= \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$

Note that $\bar{\hat{y}} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

because $\bar{\hat{y}} - \bar{y} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) = -\hat{e}_i = 0$

$\Rightarrow \bar{\hat{y}} = \bar{y}$

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

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}\hat{y}_i\hat{e}_i - \sum_{i=1}^{n}\bar{y}\hat{e}_i\right] = 0.$$

$\sum_{i=1}^{n}\hat{y}_i\hat{e}_i = \hat{y}^T\hat{e}$

$\bar{y}\sum_{i=1}^{n}\hat{e}_i = 0$

$\hat{y} = X\hat{\beta} \Rightarrow \hat{\beta}^T X^T \hat{e} = 0$ by normal equations.

(2.2) **population version** Next we will study the properties of $\hat{e}$ by treating it as random variables.

$$\hat{e}_{n\times 1} = y_{n\times 1} - \hat{y}_{n\times 1} = y - X(X^TX)^{-1}X^Ty = \left(I_n - X(X^TX)^{-1}X^T\right)y.$$

$$= X\beta + e - X\hat{\beta}$$

$(X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T(X\beta + e) = \beta + (X^TX)^{-1}X^Te.$

$$= X\beta + e - X\beta - X(X^TX)^{-1}X^Te$$

$X\beta + e = I_n e$

$$= \left(I_n - X(X^TX)^{-1}X^T\right)e$$

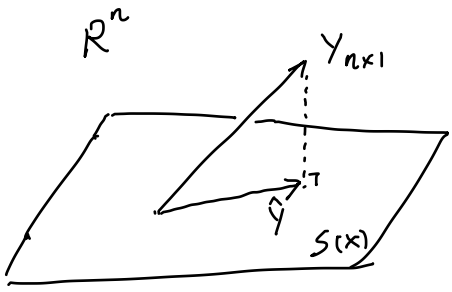
$H \triangleq X(X^TX)^{-1}X^T$ is often called the "Hat Matrix".

Note from the above derivation $\hat{y} = Hy$

Some properties of H: $\cdot\ HX = X$

because $HX = X(X^TX)^{-1}X^TX = X$.

- $HH = H$ because $HH = X(X^TX)^{-1}X^T X(X^TX)^{-1}X^T = H$.
- $H^T = H$



- From the geometry interpretation of OLS. $(n > p+1)$

  $\hat{Y}_{n\times 1}$ is the closest point in the subspace $S(X)$, expanded by the column vectors of $X_{n\times(p+1)}$ matrix, to our response vector $Y_{n\times 1}$, and the projection matrix is H.

  $(\hat{Y} = HY)$.

✓ In summary, $\hat{e} = (I_n - H)Y = (I_n - H)e$.

Next, we will calculate $E[\hat{e}|X]$ and $Var(\hat{e}|X)$

based on $E(e|X) = 0$, $Var(e|X) = \sigma^2 I$

linear reg. model assumptions.

Prop 1: $\underset{n\times 1}{E(\hat{e}|X)} = 0$.

$\Rightarrow Cov(\hat{e}_i, x_i|X) = E[\hat{e}_i x_i|X] - E[\hat{e}_i|X]\,E[x_i|X]$, with $E[\hat{e}_i|X] = 0$.

Similarly $Cov(\hat{e}_i, x_j|X) = 0$.

Proof: $E[\hat{e}|X] = E\left((I_n - H)e \,\middle|\, X\right) = 0$.

Prop 2: $\underset{n\times n}{Var(\hat{e}|X)} = \sigma^2(I_n - H)$.

Proof: $Var(\hat{e}|X) = Var\left((I_n - H)e \,|\, X\right)$

For A only depending on X.

$Var(Ae|X) = A\,Var(e|X)A^T$ $= (I_n - H)\,Var(e|X)\,(I_n - H)^T$

$Var(e|X) = \sigma^2 I_n$

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

$= \sigma^2 (I_n - H - H + \underbrace{HH}_{=H})$

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

- Let $H = (h_{ij})_{n\times n}$ with h_{ij} being the (i,j) element.

Then h_{ii} corresponds to the i^{th} diagonal element of H.

By prop 2. $\quad Var(\hat{e}_i|X) = \sigma^2(1-h_{ii})$ for $i=1,\dots,n$.

- The h_{ii} is called the leverage of the i^{th} observation.

from $H = X(X^TX)^{-1}X^T$, we know $\quad X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$

(1) $\quad h_{ii} = x_i^T(X^TX)^{-1}x_i$ recall $x_i = \begin{pmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,p} \end{pmatrix}_{(p+1)\times 1}$

here $x_i = (1, x_{i,1}, \dots x_{i,p}) \in \mathbb{R}^{p+1}$.

more generally

$$h_{ij} = x_i^T(X^TX)^{-1}x_j.$$

- Some properties of h_{ii}

① $\quad 0 \le h_{ii} \le 1$

② observations with ~~a~~ larger h_{ii} (smaller) will have smaller (larger) $Var(\hat{e}_i|X)$

$Var(\hat{e}_i|X) = \sigma^2(1-h_{ii})$ → σ^2 if $h_{ii} \to 0$; → 0 if $h_{ii} \to 1$

If $h_{ii}=1$, then $\hat{e}_i = 0$ and $\hat{y}_i = y_i$

that is, (x_i, y_i) is on the fitted OLS regression line/plane.

③ Consider the Simple linear Regression $(y_i \in \mathbb{R}, x_i \in \mathbb{R})_{i=1,\dots,n}$

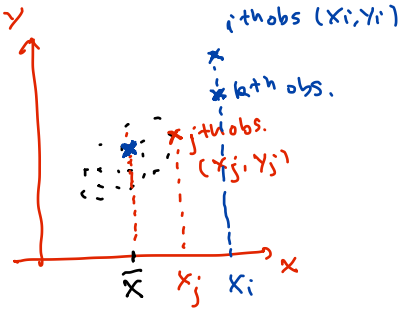
from equation (1) above, we have

$$h_{ii} = (1, x_i)(X^TX)^{-1}\begin{pmatrix} 1 \\ x_i \end{pmatrix} \quad \text{where } X_{n\times 2} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n\times 2}$$

[practice problem] $= \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{j=1}^n (x_j-\bar{x})^2}$ → fixed for all $i=1,\dots,n$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

- This implies $1 \ge h_{ii} \ge \frac{1}{n}$ if $x_i = \bar{x}$ then $h_{ii} = \frac{1}{n}$.



$h_{jj} < h_{ii} = h_{kk}$.

Prop 3: $\underset{n\times 1}{\hat{e}}, \underset{n\times 1}{\hat{Y}} \mid X) = 0$.

Proof: $\mathrm{Cov}(\hat{e}, \hat{Y} \mid X)$

$= \mathrm{Cov}((I_n - H)Y, HY \mid X)$

$= (I_n - H)\,\underbrace{\mathrm{Var}(Y|X)}_{\sigma^2 I_n}\,\underbrace{H^T}_{H}$

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

$= \sigma^2 (H - H^2) = 0$. #.

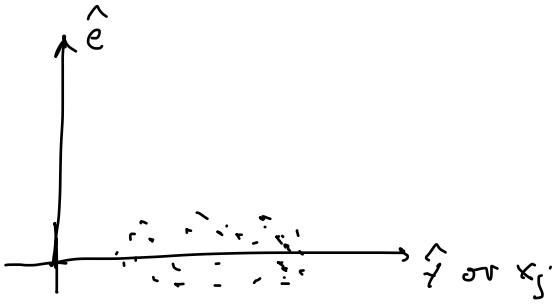
For matrix A, B only depending on X, we have

$\mathrm{Cov}(AY, BY \mid X) = A \cdot \mathrm{Var}(Y|X) \cdot B^T$

☆ Application of Residuals (§9.1 Applied linear Regression)

- Residuals are generally used in scatterplots of $\hat{e}$ against the fitted value $\hat{Y}$ or some predictors x_j, $j = 1, \ldots, p$.

[Residual Plot]

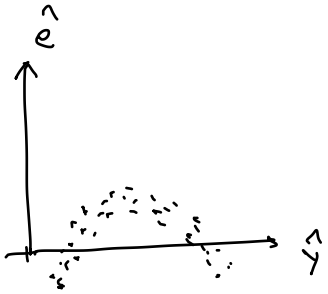


- If our regression model is true (meaning our mean and variance assumptions both hold)

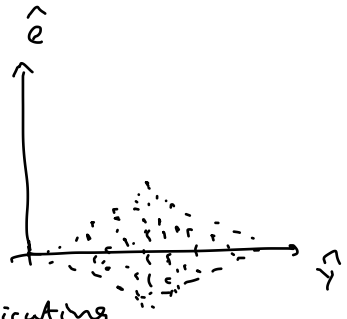
Prop 1 & 3 $\Rightarrow$ the scatterplot of $\hat{e}$ vs any x or $\hat{Y}$ should have a constant mean function $= 0$.

Prop 2 $\Rightarrow$ Residual should appear to have "Approximately" constant variance

E.g.:



indicating the failure of our linear mean function assumption



indicating that constant variance assumption is problematic.

Prop 4 (Note that all props assume linear regression model assumptions hold)

p predictors.

$$E[\hat{\sigma}^2 | X] = \sigma^2$$

where $\hat{\sigma}^2$ is the OLS estimator $\hat{\sigma}^2 = \frac{RSS}{n-(p+1)}$, $RSS = \sum_{i=1}^{n} \hat{e}_i^2$.

That is, $\hat{\sigma}^2$ is an unbiased estimator of σ^2. ($\sigma^2 = Var(e_i|x)$)

Proof: We know $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-(p+1)} = \frac{\hat{e}^T \hat{e}}{n-(p+1)}$, $\underset{n\times 1}{\hat{e}} = \begin{pmatrix} \hat{e}_1 \\ \vdots \\ \hat{e}_n \end{pmatrix}$

$\hat{e} = (I_n - H)e$

$$= \frac{\left[(I_n - H)e\right]^T (I_n - H)e}{n-(p+1)}$$

$$= \frac{e^T (I_n - H)(I_n - H) e}{n-(p+1)}$$

Note that $(I_n - H)(I_n - H) = I_n - 2H + HH = I_n - H$ ($HH = H$)

$$= \frac{e^T (I_n - H) e}{n-(p+1)}$$

Then

$$E[\hat{\sigma}^2 | X] = \frac{1}{n-(p+1)} E\left[e^T (I_n - H) e \,|\, X \right]$$

$$= \frac{1}{n-(p+1)} E\left[Tr\left(e^T (I_n - H) e \right) | X \right]$$

For A, B.
$Tr(AB) = Tr(BA)$

$$= \frac{1}{n-(p+1)} E\left[Tr\left((I_n - H) e e^T \right) | X \right]$$

$$= \frac{1}{n-(p+1)} Tr\left[E\left[(I_n - H) e e^T | X \right] \right]$$

Note that $E[ee^T | X] = Var(e|X) = \sigma^2 I_n$

$\left(Var(e|X) = E\left[(e - E(e|X))(e^T - E(e^T|X)) | X \right] \right)$ from our variance assumption.

$$= \frac{1}{n-(p+1)} Tr\left[(I_n - H) \sigma^2 I_n \right]$$

$$= \sigma^2 \cdot \frac{Tr(I_n - H)}{n-(p+1)}.$$

From above, to show $E(\hat{\sigma}^2 | X) = \sigma^2$, we only need to show

$$(\star) \quad \cdots \quad Tr(I_n - H) = n - (p+1).$$

To show $(\star)$, Note that

$$Tr(I_n - H) = Tr(I_n) - Tr(H)$$

$$= n - Tr\left(X (X^T X)^{-1} X^T \right)$$

(X: $n \times (p+1)$; $X(X^TX)^{-1}X^T$: $n \times n$)

$$= n - Tr\left((X^T X)^{-1} X^T X \right)$$

(X^TX: $I_{(p+1)\times(p+1)}$)

$$= n - (p+1). \quad \star.$$

Note that from the above proof, we have

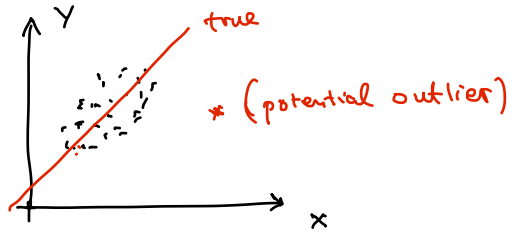
$$Tr(H) = p+1, \qquad E[RSS | X] = \sigma^2 (n-(p+1)).$$

$$\Rightarrow \sum_{i=1}^{n} h_{ii} = p+1.$$

(3) Detect Outliers (§9.4 Applied linear Reg.)

(Linear Reg. model)

- Outliers : observations that do not follow the same model as the majority and often called "Outliers"



- Cases with large residuals are candidates for outliers
  - Not all large residuals are outliers
  - Outliers may not be bad.
  - Some outliers will have larger influence on regression estimates than others ↑ (will be discussed in a later lecture)

- We use the Studentized Residual to detect outliers.

For the ith observation, is $t_i = \dfrac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$ ← Recall $\sqrt{\mathrm{var}(\hat{e}_i \mid X)} = \sqrt{\sigma^2(1-h_{ii})}$

where $\hat{\sigma}_{(i)}$ is the estimated σ without using the ith observation.

Fit regression $(Y_j, X_j)_{j=1,\cdots,i-1,i,\cdots n}$ (OLS) $\Rightarrow \hat{\sigma}^2_{(i)} = \dfrac{RSS_{(i)}}{(n-1)-(p+1)}$

If the ith observation is not an outlier (the rest also are not outliers) and we assume normal statistical error,

then

$$t_i \sim t\text{-distribution}$$

with d.f. $= (n-1)-(p+1)$

$= n-p-2.$