- Residuals (Definition)

$$e_i = y_i - \hat{y}_i$$

Hence $\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_N - \hat{y}_N \end{bmatrix} = y - X\hat{\beta}$

For assumptions Gaussian iid random errors, $\varepsilon_i, i = 1, 2, \ldots N$, we have

$$\hat{\beta} = \hat{\beta}_{LS} = (X^T X)^{-1} X^T y .$$

As a result, $\mathbf{e} = y - X\hat{\beta}$ $\qquad \hat{y} = Hy$

$$= y - X(X^T X)^{-1} X^T y$$

$$= \left(I - X(X^T X)^{-1} X^T\right) y = \underbrace{(I - H) y}$$

By definition, $H \triangleq X(X^T X)^{-1} X^T$ is called "hat matrix".

We can further derive:

$$e = (I - H) y$$

$$= (I - H)(X\beta + \boldsymbol{\varepsilon}), \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

$$= IX\beta - HX\beta + (I - H)\boldsymbol{\varepsilon}$$

$$= X\beta - \underbrace{X(X^T X)^{-1} X^T X\beta}_{= X\beta} + (I - H)\boldsymbol{\varepsilon}$$

$$= (I - H)\boldsymbol{\varepsilon}$$

● Nice properties of the hat matrix $H = X(X^TX)^{-1}X^T$

① H is idempotent., symmetric. ✓ Shown before

② $(I-H)$ is idempotent, symmetric. ✓ Shown before

③ The i-th diagonal entry of H, i.e. h_{ii} satisfies $0 \le h_{ii} \le 1$.

④ Suppose $\text{rank}(X) = p$, then $\text{rank}(X(X^TX)^{-1}X^T) = p$, and the eigenvalue of H consists of p ones and $N-p$ zeros.

⑤ $Hy = H\hat{y}$, $\rightarrow \sum_{j=1}^{N} e_j h_{ij} = 0, \forall i = 1, 2, \ldots, N$

▸ We give the proof below:

①
$$H^T = X\left[(X^TX)^{-1}\right]^T X^T = X\left[(X^TX)^T\right]^{-1} X^T = X(X^TX)^{-1}X^T = H$$

Therefore, H is a symmetrix matrix.

$$H \cdot H = X(X^TX)^{-1}\underbrace{X^T X(X^TX)^{-1}}_{\text{merged to be } I_p}X^T = X(X^TX)^{-1}X^T$$

②
$$(I-H)^T = I - H^T = I - H \quad \text{(using the result above)}$$

Therefore $I-H$ is a symmetric matrix.

$$(I-H)(I-H) = I - H \underbrace{- H + H \cdot H}_{=0} \quad (\text{due to } HH = H)$$
$$= I - H$$

③ Due to the fact that $HH = H$

Suppose $H = \begin{bmatrix} h_{11}, & h_{12}, & \cdots & h_{1N} \\ h_{21}, & h_{22}, & \cdots & h_{2N} \\ \vdots & & & \\ h_{N1}, & h_{N2} & \cdots & h_{NN} \end{bmatrix}$

h_{ii} = the ith row of H $\overset{\text{(multiply)}}{\times}$ the ith column of H (i.e. the ith row of H transp

$$= \sum_{j=1}^{N} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i}^{N} h_{ij}^2$$

This yields

$$h_{ii} - h_{ii}^2 = \sum_{j \neq i}^{N} h_{ij}^2 \geq 0$$

$$\Rightarrow \quad h_{ii}(1 - h_{ii}) \geq 0$$

$$\Rightarrow \quad 0 \leq h_{ii} \leq 1$$

④

Suppose v and λ are respectively the eigenvector and eigenvalue of H, then we have $Hv = \lambda v$.

As we know: $H^n v = \lambda^n v$, $\Rightarrow \underline{H^2 v = \lambda^2 v}$

Since $H^2 = HH = H$ due to the idempotent property of H, we final have $H^2 v = \underline{\lambda^2 v = \lambda v} = Hv \Rightarrow \lambda = 0$, or 1.

As we also know, $tr(H) = tr(X(X^T X)^{-1} X^T) = tr(I_p)$

$= p = \sum_{i=1}^{N} \lambda_i \Rightarrow p$ eigenvalues $= 1$

⑤ Since $e = y - \hat{y} = (I - H) y = y - Hy$

$\Rightarrow \underline{\underline{\hat{y} = Hy}} \Rightarrow H\hat{y} = H \cdot H y = Hy$ (due to $HH = H$)

$Hy = H\hat{y} = H(y - \hat{y}) = H\mathbf{e} = \mathbf{0}$

Suppose $H = \begin{bmatrix} h_{11} & h_{12}, & \cdots, & h_{1N} \\ h_{21} & h_{22}, & \cdots, & h_{2N} \\ \vdots & & & \\ h_{N1} & h_{N2} & \cdots & h_{NN} \end{bmatrix} = \begin{bmatrix} h_1^t \\ h_2^t \\ \vdots \\ h_N^t \end{bmatrix}$

$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$

Then $H\mathbf{e} = 0 \Rightarrow \forall$ row i, $h_i^t e = \sum_{j=1}^{N} h_{ij} e_j = 0$ //

- For the standardized Residuals;

$Var(e_i) = var(\varepsilon_i) = \sigma^2$, this assumption holds if $\hat{\beta}_{LS} = \beta$, i.e.

The true parameter is precisely estimated. This is in general impossible, but in case $\hat{\beta}_{LS}$ is a very good estimator of β, we can treat approximately $var(e_i) \approx var(\varepsilon_i) = \sigma^2$. Since σ^2 is unknown, $MS_{Res} = \frac{SS_{Res}}{n-p}$ is used as an estimator of $Var(e_i)$.

- Studentized Residuals.

Essentially, we can compute the covariance matrix of $\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$ precisely.

$$\begin{aligned}
Cov(\mathbf{e}) &\triangleq E\{(\mathbf{e} - E(\mathbf{e}))(\mathbf{e} - E(\mathbf{e}))^T\} \\
&= E\{((I-H)\varepsilon - 0)((I-H)\varepsilon - 0)^T\} && [\text{due to } \mathbf{e} = (I-H)\varepsilon,\ E(\varepsilon) = \vec{0}] \\
&= (I-H)\, E\{\varepsilon\varepsilon^T\}\, (I-H)^T \\
&= (I-H) \cdot Cov(\varepsilon)\, (I-H)^T && [\text{due to } Cov(\varepsilon) \triangleq E\{\varepsilon\varepsilon^T\}] \\
&= \sigma^2 \cdot (I-H) && [\text{due to } I-H \text{ is symmetric and idempodent}]
\end{aligned}$$

We ~~can~~ expand Cov (**e**), which is $N \times N$ matrix of the following form:

$$\mathrm{Cov}(\mathbf{e}) = \begin{bmatrix} \mathrm{Var}(e_1) & \mathrm{Cov}(e_1,e_2), & \mathrm{Cov}(e_1,e_3) & \cdots & \mathrm{Cov}(e_1 e_N) \\ \mathrm{Cov}(e_2,e_1) & \mathrm{Var}(e_2), & \mathrm{Cov}(e_2,e_3) & \cdots & \mathrm{Cov}(e_2 e_N) \\ \vdots & \vdots & \mathrm{Var}(e_3) & & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(e_N,e_1) & \mathrm{Cov}(e_N,e_2) & \mathrm{Cov}(e_N,e_3) & \cdots & \mathrm{Var}(e_N) \end{bmatrix}$$

Then, it is not hard to see

$$\mathrm{var}(e_i) = \sigma^2 \cdot (1 - h_{ii})$$

$$\mathrm{cov}(e_i, e_j) = -\sigma^2 h_{ij}$$

Note: h_{ii} is actually a function of the input!

where h_{ii} is the ith diagonal entry of H
h_{ij} the the (i,j)th element of H.

Furthermore, we should have $e_i \sim N(0, \sigma^2(1-h_{ii}))$, then

$$\frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}} \sim N(0,1).$$

Suppose MS_{Res} is a good estimator of σ^2, we can approximately regard:

$$\frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}} \mathrel{\dot\sim} N(0,1)$$

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

Derivation of h_{ii} for Simple Linear Regress

$$h_{ii} = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{S_{xx}}, \quad i = 1, 2, \cdots, N$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & \vdots \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}_{N\times 2}$$

$$X^T = \begin{bmatrix} 1, & 1, & \cdots & 1 \\ x_1, & x_2, & \cdots & x_N \end{bmatrix}_{2\times N}$$

$$(X^TX) = \begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

$$(X^TX)^{-1} = \frac{1}{N\cdot\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} \cdot \begin{bmatrix} \sum_{i=1}^{N} x_i^2 & -\sum_{i=1}^{N} x_i \\ -\sum_{i=1}^{N} x_i & N \end{bmatrix}$$

$$= \frac{1}{N\cdot(S_{xx})} \cdot \begin{bmatrix} \sum_{i=1}^{N} x_i^2 & -\sum_{i=1}^{N} x_i \\ -\sum_{i=1}^{N} x_i & N \end{bmatrix}_{2\times 2}$$

$$X(X^TX)^{-1} = \frac{1}{N S_{xx}} \begin{bmatrix} \sum_{i=1}^{N} x_i^2 - x_1\sum_{i=1}^{N} x_i, & -\sum_{i=1}^{N} x_i + Nx_1 \\ \sum_{i=1}^{N} x_i^2 - x_2\sum_{i=1}^{N} x_i, & -\sum_{i=1}^{N} x_i + Nx_2 \\ \vdots & \\ \sum_{i=1}^{N} x_i^2 - x_N\sum_{i=1}^{N} x_i, & -\sum_{i=1}^{N} x_i + N\cdot x_N \end{bmatrix}_{N\times 2}$$

Next, we compute $X(X^TX)^{-1}X^T$, using the results of $X(X^TX)^{-1}$ and X^T.

It is easy to derive that:

$$\left(X(X^TX)^{-1}X\right)_{ii} = h_{ii}$$

$$= \frac{1}{N \cdot S_{xx}} \cdot \left[\sum_{j=1}^{N} x_j^2 - x_i \sum_{j=1}^{N} x_j - x_i \sum_{j=1}^{N} x_j + N \cdot x_i^2 \right]$$

$$= \frac{1}{S_{xx}} \cdot \left[\frac{1}{N} \sum_{j=1}^{N} x_j^2 - \frac{2}{N} x_i \sum_{j=1}^{N} x_j + x_i^2 \right]$$

$$= \frac{1}{S_{xx}} \left[\frac{1}{N} \sum_{j=1}^{N} x_j^2 - 2 x_i \cdot \bar{x} + x_i^2 \right]$$

$$= \frac{1}{S_{xx}} \left[\underbrace{\frac{1}{N} \sum_{j=1}^{N} x_j^2 - (\bar{x})^2}_{= \frac{1}{N}\left(\sum_{j=1}^{N} x_j^2 - \frac{\left(\sum_{j=1}^{N} x_j\right)^2}{N}\right)} + \underbrace{(\bar{x})^2 - 2 x_i \bar{x} + x_i^2}_{= (\bar{x} - x_i)^2} \right]$$

$$= \frac{1}{S_{xx}} \left[\frac{1}{N} \underbrace{\left(\sum_{j=1}^{N} x_j^2 - \frac{\left(\sum_{j=1}^{N} x_j\right)^2}{N} \right)}_{= S_{xx} = \sum_{j=1}^{N} (x_j - \bar{x})^2} + (\bar{x} - x_i)^2 \right]$$

$$= \frac{1}{N} + \frac{(\bar{x} - x_i)^2}{S_{xx}}$$

- For big data, i.e., $N \to \infty$,

$$\frac{1}{N} \to 0$$

$$\frac{(\bar{x} - x_i)^2}{S_{xx}} = \frac{(\bar{x} - x_i)^2}{\sum_{j=1}^{N} (x_j - \bar{x})^2} \approx 0$$

PRESS residual:

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

$e_{(i)}$: scalar; y_i: scalar; $\hat{y}_{(i)}$: Scalar

herein, $\hat{y}_{(i)}$ is the fitted value of the ith output based on all datapoints except the ith one.

prove: $e_{(i)} = \dfrac{e_i}{1-h_{ii}}$, $i = 1, 2, \dots, N$

Let $\hat{\beta}_{(i)} = [X_{(i)}^T X_{(i)}]^{-1} X_{(i)}^T y_{(i)}$

$y_{(i)}$: Vector

The dimension of $\hat{\beta}_{(i)}$ and $\hat{\beta}$ the same! i.e. $(p \times 1)$

Note: $\hat{\beta}_{(i)}$ is the LS estimator of β based on N-1 data pair without the i-th one

Here, we assume

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}_{N \times p}, \quad \text{where } \mathbf{x}_i = \begin{bmatrix} 1 \\ x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,k} \end{bmatrix}_{p \times 1} \ (p = k+1)$$

$$X_{(i)} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_{i-1}^T \\ x_{i+1}^T \\ \vdots \\ x_N^T \end{bmatrix}_{(N-1) \times p} \qquad y_{(i)} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_N \end{bmatrix}_{(N-1) \times 1}$$

$$X^T = [\mathbf{x}_1, x_2, \cdots, x_N]_{p \times N} \qquad X_{(i)}^T = [\mathbf{x}_1, \mathbf{x}_2, \cdots, x_{i-1}, \mathbf{x}_{i+1}, \cdots, \underbrace{x_N}_{p \times 1}]$$

The ith PRESS residual is written as:

$$\overset{\text{Scalar}}{e_{(i)}} = \overset{\text{Scalar}}{y_i} - \overset{\text{scalar}}{\hat{y}_{(i)}}$$

$$= y_i - \overset{\text{row vector}}{x_i^T} \underset{\text{Column vector}}{\hat{\beta}_{(i)}}$$

$$= y_i - x_i^T (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T \underset{\text{Column vector } (N-1)\times 1}{y_{(i)}}$$

Before we proceed further, we have:

$$X_{(i)}^T X_{(i)} = \sum_{j=1, j\neq i}^{N} x_j \cdot x_j^T$$

(outer-product View of Matrix Multiplication)

$$X^T X = \sum_{j=1}^{N} x_j x_j^T$$

$$X^T X - x_i x_i^T = X_{(i)}^T X_{(i)}$$

Hence

$$\left(X_{(i)}^T X_{(i)}\right)^{-1} = \left(X^T X - x_i x_i^T\right)^{-1}$$

Using Sherman–Morrison–Woodbury Theorem:

$$(A + UCV)^{-1} = A^{-1} - A^{-1} U (C^{-1} + V A^{-1} U)^{-1} V A^{-1}$$

where A, U, C, V are all of correct size, e.g.

$A_{n\times n}$ $U_{n\times k}$, $C_{k\times k}$, $V_{k\times n}$.

We can regard: $A = X^TX$ (size $p\times p$)

$U = x_i$ (size $p\times 1$)

$C = -1$ (size 1×1)

$V = x_i^T$ (size $1\times p$)

Hence:

$$(X^TX - x_i x_i^T)^{-1} = (X^TX)^{-1} - (X^TX)^{-1} x_i \underbrace{(-1 + x_i^T X^T X x_i)^{-1}}_{\text{Scalar}} x_i^T (X^T$$

$$= (X^Tx)^{-1} + \frac{(X^TX)^{-1} x_i x_i^T (X^TX)^{-1}}{\underbrace{1 - x_i^T X^T X x_i}_{= 1-h_{ii}}}$$

We use the above result in

$$e_{(i)} = y_i - x_i^T (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T y_{(i)}$$

$$= y_i - x_i^T (X^TX)^{-1} X_{(i)}^T y_{(i)} - \frac{\overbrace{x_i^T (X^Tx)^{-1} x_i}^{h_{ii}} x_i^T (X^TX)^{-1} X_{(i)}^T y_{(}}{1-h_{ii}}$$

$$= \frac{(1-h_{ii}) y_i - x_i^T (X^TX)^{-1} X_{(i)}^T y_{(i)}}{1-h_{ii}} \longrightarrow \text{annoying!}$$

$$X_{(i)}^T y_{(i)} = X^T y - \mathbf{x}_i y_i$$, using outer-product formulation of matrix multiplication.

$\mathbf{x}_i y_i$: Scalar

Then we have

$$e_{(i)} = \frac{(1-h_{ii}) y_i - x_i^T (X^T X)^{-1} X_{(i)}^T y_{(i)}}{1-h_{ii}}$$

y_i: Scalar; $y_{(i)}$: vector $= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_N \end{bmatrix}$

$$= \frac{(1-h_{ii}) y_i - \mathbf{x}_i^T (X^T X)^{-1} (X^T y - \mathbf{x}_i y_i)}{1-h_{ii}}$$

y_i: scalar

$$= \frac{(1-h_{ii}) y_i - \mathbf{x}_i^T \hat{\beta} + h_{ii} y_i}{1-h_{ii}}$$

$$= \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{1-h_{ii}} = \frac{e_i}{1-h_{ii}}$$

$\boxed{\text{Final Conclusion:}}$ amazingly, we don't need to compute $e_{(i)}$, N times and each time re-training the model with the ith data point unused (withheld). What we require is to fit (using LS) the model once with all datapoints and compute the ordinary residuals e_i, $i=1,2,\ldots,N$ and leverage score h_{ii}, $i=1,2,\ldots N$ as the scaling factor of e_i. Eventually, $e_{(i)} = \frac{e_i}{1-h_{ii}}$, $i=1,2,\ldots$

- Variance of PRESS residual:

$$Var(e_{(i)}) = var\left(\frac{e_i}{1-h_{ii}}\right) = \frac{1}{(1-h_{ii})^2} \cdot var(e_i) = \frac{1}{(1-h_{ii})^2} \cdot \sigma^2(1-h_{ii})$$

Note $1-h_{ii}$ is squared!

$$= \frac{\sigma^2}{1-h_{ii}}$$

- Standardized PRESS residual

$$\frac{e_{(i)}}{\sqrt{var(e_{(i)})}} = \frac{e_i/(1-h_{ii})}{\sqrt{\sigma^2/(1-h_{ii})}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

If σ^2 is replaced by MS_{Res}, then the <u>Standardized PRESS residual is equivalent to Studentized residual</u>.

- R-student residual

$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2(1-h_{ii})}}$$

, where $S_{(i)}^2$ is a robust estimator of σ^2 defined by:

$$S_{(i)}^2 \triangleq \frac{\sum_{j\neq i}^{N}(y_j - x_j^T\hat{\beta}_{(i)})^2}{N-p-1} \overset{\text{C.8 (Textbook) Appendix}}{=\!=} \frac{(N-p)MS_{Res} - e_i^2/(1-h_{ii})}{N-p-1}$$

$$S_{(i)}^2 = \sum_{j\neq i, j=1}^{N} (y_j - x_j^T \hat{\beta}_{(i)})^2 / N-p-1$$

$$= \Big(\underset{\text{Column vector } (N-1)\times 1}{y_{(i)}} - X_{(i)}(X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T y_{(i)}\Big)^T \Big(y_{(i)} - X_{(i)}(X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T y_{(i)}\Big) / N-p-1$$

$$= \frac{y_{(i)}^T (I - H_{(i)}) y_{(i)}}{N-p-1}$$

Suppose $U = y_{(i)}^T \boxed{(I - H_{(i)})} y_{(i)}$ → Similar to $I-H$, it is symmetric and idempotent.

We know $y_{(i)} \sim N(X_{(i)}\beta, \sigma^2 I_{N-1})$

$$\frac{U}{\sigma^2} \sim \chi^{2\prime}_{p', \lambda}$$

Since $p' = \text{rank}(I - H_{(i)}) = N-1-p$ (Similar to $\text{rank}(I-H)$)

$$\lambda = E^T(y_{(i)})(I - H_{(i)}) E(y_{(i)}) / \sigma^2 \leftarrow \text{Easy!}$$

$$= 0 .$$

Therefore, $\frac{U}{\sigma^2} \sim \chi^2_{N-1-p}$ i.e. $\frac{(N-p-1) S_{(i)}^2}{\sigma^2} \sim \chi^2_{N-p-1}$

Since $e_i \sim N(0, \sigma^2(1-h_{ii})) \Longrightarrow \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}} \sim N(0,1)$

and $\frac{S_{(i)}^2 (N-p-1)}{\sigma^2} \sim \chi^2_{N-1-p}$

Hence

$$t_i = \frac{\frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}}{\sqrt{\frac{S_{(i)}^2 (N-p-1)}{\sigma^2} \Big/ (N-p-1)}} = \frac{e_i}{\sqrt{S_{(i)}^2 (1-h_{ii})}} \sim t_{n-p-1}$$

☆ Additionally, we have to verify if $\underline{e_i}$ and $y_{(i)} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_N \end{bmatrix}$ are independent

Proof: $e_{(i)}$ and $S_{(i)}^2$ are independent, $\forall i$.

$$e_{(i)} = y_i - \hat{y}_{(i)} = y_i - x_i^T \hat{\beta}_{(i)} = \overset{\text{term 1}}{y_i} - \overset{\text{term 2}}{\boxed{x_i^T (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T y_{(i)}}}$$

scalar

row vector

vector

$$S_{(i)}^2 = \frac{y_{(i)}^T (I - H_{(i)}) y_{(i)}}{N-p-1} = \frac{y_{(i)}^T (I - X_{(i)}(X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T) y_{(i)}}{N-p-1}$$

Since $\overset{\text{the first term in } e_{(i)}}{y_i}$ is independent of $y_{(i)} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_N \end{bmatrix}$, we only need to check

if the second term in $e_{(i)}$, i.e. $x_i^T (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T y_{(i)}$ is independent with $y_{(i)}^T (I - H_{(i)}) y_{(i)}$. To do this, we use the result given in Appendix C.2.4. bullet 3 of the textbook. If we follow the notations (See also Theorem 4 in lecture 2) therein, we have

$$W = \underbrace{x_i^T (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T}_{B} y_{(i)}$$

$$U = y_{(i)}^T \underbrace{(I - H_{(i)})}_{A} y_{(i)}$$

$$y_{(i)} \sim N(X_{(i)} \beta, \underbrace{\sigma^2 I}_{V})$$

notice this matrix is of size $(N-1)\times(N-1)$!!

Then we only need to check if $BVA = 0$?

In fact: $BVA = \sigma^2 \cdot \left(x_i^T (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T - x_i^T (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T X_{(i)} (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T \right) = 0$