Residualer

$$\text{Residual } r(y) = y - \hat{y} = (I - H)y$$
$$r_i(y) = y_i - \hat{y}_i$$

$$\text{Kovariansmatrix } \mathrm{Var}(r(Y)) = \sigma^2(IH)$$
$$\mathrm{Var}(r_i(y)) = \sigma^2(1 - h_{ii})$$

Standardiserede residualer

$$r_i^{rs}(y) = \frac{r_i(y)}{\sqrt{\hat{\sigma}^2(y)(1 - h_{ii})}}$$

Disse har samme varians, dog ikke 1, da $\hat{\sigma}^2$ er brugt i stedet for σ^2.

Residualer bruges bl.a. til at identificere outliers. Outliers får $\hat{\sigma}^2$ til at blive overestimeret, hvilket kan betyde, at r_i^{rs} bliver mindre of ikke indikerer en outlier (selvom den burde). Derfor bruger vi variansen for modellen uden måling i.

Ny designmatrix:

$$\text{Uden måling } i \begin{cases} X_{-i} & \text{designmatrix} \\ H_{-i} & \text{projektion} \\ y_{-i} & data \end{cases}$$
$$\hat{\sigma}^2_{(i)} = \frac{\|y_{-i} - H_{-i}y_{-i}\|^2}{n - 1 - k}$$

Studentiserede residualer

$$r_i^{rt}(y) = \frac{r_i(y)}{\sqrt{\hat{\sigma}^2_{(i)}(1 - h_{ii})}}$$

Residualerne er t-fordelte

$$r_i^{rt} \sim t(n - 1 - k)$$

Bevis
Lad $i = 1$. Vi har tidligere vist, at

$$r_1(Y) \sim \mathcal{N}(0, \sigma^2(1 - h_{ii}))$$
$$\hat{\sigma}^2_{(i)}(Y_{-1}) \sim \sigma^2\chi^2(n - i - k)/(n - 1 - k)$$

Dvs., hvis $r_1(Y)$ og $\hat{\sigma}^2_{(1)}(Y_{-1})$ (eller $Y_{-1} - H_{-1}Y_{-1}$) er uafhængige, så

$$\begin{aligned}
r_1^{rt}(Y) &= \frac{r_1(Y)}{\sqrt{\hat{\sigma}^2_{(1)}(1-h_{11})}} \\
&= \frac{r_1(Y)/\sqrt{\sigma^2(1-h_{11})}}{\sqrt{\hat{\sigma}^2_{(1)}/\sigma^2}} \\
&\sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2(n-1-k)/(n-1-k)}} \\
&= t(n-i-k)
\end{aligned}$$

Da

$$\begin{aligned}
\begin{bmatrix} r_1(Y) \\ Y_{-1} - H_{-1}Y_{-1} \end{bmatrix} &= \begin{bmatrix} Y_1 - x_1(X^TX)^{-1}X^TY \\ Y_{-1} - X_{-1}(X_{-1}^TX_{-1})^TX_{-1}Y_{-1} \end{bmatrix} \\
&= \begin{bmatrix} 1 - x_1(X^TX)^{-1}x_1^T & -x_1(X^TX)^{-1}X_{-1}^T \\ 0 & I - X_{-1}(X_{-1}^TX_{-1})^{-1}X_{-1}^T \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_{-1} \end{bmatrix} \\
&= AY \sim \mathcal{N}(\cdots)
\end{aligned}$$

hvor

$$A = \begin{bmatrix} A_1^{1\times n} \\ A_2^{(n-1)\times N} \end{bmatrix}$$

Derfor er $r_1(Y) = A_1Y$ og $Y_{-1} - H_{-1}Y_{-1} = A_2Y$ uafhængige, hvis de er ukorrelerede.

$$\begin{aligned}
\mathrm{Cov}(A_1Y_1A_2Y) &= A_1\mathrm{Cov}(Y)A_2^T \\
&= A_1\sigma^2IA_2^T \\
&= \sigma^2A_1A_2^T \\
&= \sigma^2\begin{bmatrix} 1 - x_1(X^TX)^{-1}x_1^T & -x_1(X^TX)^{-1}X_{-1}^T \end{bmatrix} \begin{bmatrix} 0 \\ I - X_{-1}(X_{-1}^TX_{-1})^{-1}X_{-1}^T \end{bmatrix} \\
&= \sigma^2\left(-x_1(X^TX)^{-1}X_{-1}^T\left(I - X_{-1}(X_{-1}^TX_{-1})^{-1}X_{-1}^T\right)\right) \\
&= \sigma^2\left(-x_1(X^TX)^{-1}X_{-1}^T + x_1(X^TX)^{-1}X_{-1}^TX_{-1}(X_{-1}^TX_{-1})^{-1}X_{-1}^T\right) \\
&= 0
\end{aligned}$$

Bemærk: De standardiserede residualer er ikke t-fordelt, idet $r_i(Y)$ ikke er uafhængig af $\hat{\sigma}^2(Y)$ ($r_i(Y)$ indgår i $\hat{\sigma}^2$).

Praktiske anvendelser af residualer

- Outliers: store $r_i^{rt}(y)$ indikerer outliers.
- Fordeling: QQ-plots af $r_i^{rt}(y)$ – er de t-fordelt eller cirka normalfordelt for store n?
- Residualplots: Residualer mod tid/indeks, uafhængige variable, fittede værdier $\hat{y}$.

Udseenden af residualplots mod tid/indeks

- Konstant interval for fejl: fint.
- Trompetform: fejlen vokser som funktion af tid. Variansen stiger. Lav transformation, så dette undgås.
- Lineært voksende fejl: modellen mangler afhængighed af tid/indeks.
- Parabelformet fejl: modellen mangler afhængighed af tid^2/indeks^2.

Mod uafhængig variabel

- Fint.
- Voksende varians med x_i. Transformér.
- Fejl i udregning.
- Mangler x_i^2.

Mod fittede værdier

- Fint.
- Voksende varians med $\hat{y}$. Transformér.
- Mangler et konstantled.
- Dårlig model: afhængighed eller manglende x_i.

Indflydelse
Da $\hat{\mu}(y) = Hy$, så er
$$\frac{\partial \hat{\mu}(y)}{\partial y} = H.$$
Dvs, h_i fortæller hvor meget $\hat{mu}_i(y)$ ændrer sig når y_i ændrer sig.
Leverage
h_{ii} kaldes leverage for den i'te observation.