After fitting an Ordinary Least Squares (OLS) model, the residuals $r_i^{\text{OLS}} = Y_i - X_i\hat{\beta}$ can be plotted against explanatory variables or fitted values to get an idea of the model fit or to assess whether the constant variance assumption holds, for example. With the OLS residuals, it is possible to see evidence of nonconstant variance if, rather than looking like a random scatter, residuals tend to be larger for large values of the fitted value/explanatory variable for example, sometimes referred to as a funnel or megaphone shape, but the shape could also be like a bowtie or a bulge.

The variance-covariance of the OLS residuals is

$$\begin{aligned}\text{var}(r^{\text{OLS}}) &= (I - H^{\circ})\text{var}(Y)(I - H^{\circ})^T \\ &= \sigma^2(I - H^{\circ})\end{aligned}$$

where $H^{\circ} = X(X^TX)^{-1}X^T$ is the leverage matrix, so the variance of an individual residual is

$$\text{var}(r_i^{\text{OLS}}) = \sigma^2(1 - H_i^{\circ}).$$

For the OLS model, the observations are uncorrelated but the residuals are not uncorrelated.

The Leave-One-Out Cross Validation (LOOCV) residuals for the OLS model are

$$Y_i - X_i\hat{\beta}_{(i)} = \frac{r_i}{1 - H_i^{\circ}}$$

where $\hat{\beta}_{(i)}$ is the OLS estimator of β when case i is excluded from the data. The derivation of this result is not new, and it is outlined in the Appendix.

Another OLS residual, DFFIT_i, is

$$X_i\hat{\beta} - X_i\hat{\beta}_{(i)} = \frac{r_i H_i^{\circ}}{1 - H_i^{\circ}}$$

We note that $r_i + \text{DFFIT}_i = \text{LOOCV}_i$.

The LOOCV residuals are squared and summed to gauge model performance on unseen data.

For OLS, Cook's distance, a measure of case influence, is

$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T[\widehat{\text{var}(\hat{\beta})}]^{-1}(\hat{\beta} - \hat{\beta}_{(i)})}{\text{rank}X}$$

where

$$\begin{aligned}\widehat{\text{var}(\hat{\beta})} &= (X^TX)^{-1}X^T\text{var}(Y)X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}\end{aligned}$$

so

$$\begin{aligned} C_i &= \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T X^T X (\hat{\beta} - \hat{\beta}_{(i)})}{p\sigma^2} \\ &= \frac{r_i^2 H_i^\circ}{p\sigma^2 (1 - H_i^\circ)^2} \end{aligned}$$

Cook's distance divides by σ^2 if it is known, or an estimate of σ^2 otherwise. When plots are made of OLS residuals, or the sum of squares of LOOCV residuals is calculated, there is no need to think about dividing by the possibly unknown σ^2 because the effect on each residual is constant. Minimising the residual sum of squares is the same as minimising the residual sum of squares divided by a constant.

Now consider what happens when the OLS assumptions of uncorrelated observations and constant variance are dropped. For the OLS model, the variance of the observations Y is $\sigma^2 I$. Under generalised least squares (GLS), the assumption of uncorrelated errors is relaxed, as is the assumption that the error variance is constant; the variance of Y is now a nondiagonal matrix V, which has elements V_{ij}. As well as OLS and GLS, another possibility is that errors are uncorrelated but the variance is not constant, in which case the variance of Y is a diagonal matrix D with terms σ_i^2. This last situation of homoscedastic errors is called weighted least squares (WLS).

It is easy to see that under WLS, if Y has variance D, $D^{-\frac{1}{2}}Y$ has variance I. The OLS estimator for β from the model for $D^{-\frac{1}{2}}Y$ is of course the same as the WLS estimator for β from the model for Y: $(X^T D^{-1} X)^{-1} X^T D^{-1} Y$.

For WLS, residual plots can be carried out after model fitting. Here one would plot the scaled residuals, $D^{-\frac{1}{2}}(Y - X\hat{\beta})$, in order to avoid detecting the non constancy of variance which may have been the reason for fitting a WLS model in the first place. The variance of the residuals $r^{\mathrm{WLS}} = Y - X\hat{\beta}$ from the WLS model is given by

$$\begin{aligned} \mathrm{var}(r^{\mathrm{WLS}}) &= (I - HD^{-1})\mathrm{var}(Y)(I - HD^{-1})^T \\ &= (I - HD^{-1})D(I - D^{-1}H) \\ &= D - H \end{aligned}$$

where $H = X(X^T D^{-1} X)X^T$, whereas the variance of the scaled residuals $\tilde{r}^{\mathrm{WLS}} = D^{-\frac{1}{2}} r^{\mathrm{WLS}}$ is

$$\begin{aligned} \mathrm{var}(\tilde{r}^{\mathrm{WLS}}) &= D^{-\frac{1}{2}}(I - HD^{-1})\mathrm{var}(Y)(I - HD^{-1})^T D^{-\frac{1}{2}} \\ &= D^{-\frac{1}{2}}(D - H)D^{-\frac{1}{2}} \\ &= I - D^{-\frac{1}{2}} H D^{-\frac{1}{2}} \end{aligned}$$

In the OLS case, the variance for the residuals r^{OLS} was $\sigma^2(I - H^\circ)$. A scaled version of r^{OLS}, $\tilde{r}^{\mathrm{OLS}} = (\sigma^2)^{-1/2} r^{\mathrm{OLS}}$, has variance $I - H^\circ$. For WLS, the residuals $\tilde{r}^{\mathrm{WLS}}$ have variance $I - D^{-\frac{1}{2}} H D^{-\frac{1}{2}}$, where again the first matrix, I, which is the variance of $D^{-\frac{1}{2}}Y$, has the constant value, 1, on the diagonal.

For GLS, the estimate of β is given by $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$. I propose a sensible residual plot would be of the residuals given by $\tilde{r}^{\text{GLS}} = S^{\frac{1}{2}} V^{-1} r^{\text{GLS}}$ where S is the diagonal matrix with ith element $\frac{1}{V^{ii}}$, where V^{ii} is the ith diagonal element of V^{-1}.

The quantities $\tilde{X} = S^{\frac{1}{2}} V^{-1} X$, $\tilde{Y} = S^{\frac{1}{2}} V^{-1} Y$ and $\tilde{r} = S^{\frac{1}{2}} V^{-1} r$ arise when finding an expression for $\hat{\beta}_{(i)}$ for GLS. This is outlined in the Appendix and a complete derivation is in Baade 1998.

For WLS, $V = S = D$, so $S^{\frac{1}{2}} V^{-1} = D^{-\frac{1}{2}}$. For OLS, $V = S = \sigma^2 I$ so $S^{\frac{1}{2}} V^{-1} = (\sigma^2)^{-\frac{1}{2}} I$.

The variance of the residuals $r^{\text{GLS}} = Y - X\hat{\beta}$ from the GLS model is given by

$$\begin{aligned} \text{var}(r^{\text{GLS}}) &= (I - HV^{-1})\text{var}(Y)(I - HV^{-1})^T \\ &= (I - HV^{-1})V(I - V^{-1}H) \\ &= V - H \end{aligned}$$

where $H = X(X^T V^{-1} X)^{-1} X^T$, whereas the variance of the transformed residuals $\tilde{r}^{\text{GLS}} = S^{\frac{1}{2}} V^{-1} r^{\text{GLS}}$ is

$$\begin{aligned} \text{var}(\tilde{r}^{\text{GLS}}) &= S^{\frac{1}{2}} V^{-1} (I - HV^{-1}) \text{var}(Y) (I - HV^{-1})^T V^{-1} S^{\frac{1}{2}} \\ &= S^{\frac{1}{2}} V^{-1} (V - H) V^{-1} S^{\frac{1}{2}} \\ &= S^{\frac{1}{2}} V^{-1} S^{\frac{1}{2}} - S^{\frac{1}{2}} V^{-1} H V^{-1} S^{\frac{1}{2}} \end{aligned}$$

The matrix $S^{\frac{1}{2}} V^{-1} S^{\frac{1}{2}}$ has 1s on the diagonal. It is known as the partial correlation matrix. For OLS and WLS, $S^{\frac{1}{2}} V^{-1} S^{\frac{1}{2}} = I$. For GLS, $S^{\frac{1}{2}} V^{-1} S^{\frac{1}{2}}$ is not diagonal, but there is no reason for that to be a concern.

From now, I refer to $\tilde{r}$ as the transformed residuals coming from any of the three models, H as the matrix $X(X^T V^{-1} X)^{-1} X^T$ and $\tilde{H}$ as $S^{\frac{1}{2}} V^{-1} H V^{-1} S^{\frac{1}{2}}$. For OLS, the matrix $H^\circ = X(X^T X)^{-1} X^T$, is actually $\tilde{H}$.

If we are interested in defining some LOOCV residuals and DFFIT residuals for GLS, this transformation, is handy. In the appendix we show that

$$\hat{\beta}_{(i)} = \hat{\beta} - (X^T V^{-1} X)^{-1} \tilde{X}_i^T (1 - \tilde{H}_i)^{-1} \tilde{r}_i$$

for the GLS model.

LOOCV residuals for the GLS model could be given by:

$$\begin{aligned} \text{LOOCV}_i &= \tilde{Y}_i - \tilde{X}_i \hat{\beta}_{(i)} \\ &= \tilde{Y}_i - \tilde{X}_i \hat{\beta} + \tilde{H}_i (1 - \tilde{H}_i)^{-1} \tilde{r}_i \\ &= \tilde{r}_i + \tilde{H}_i (1 - \tilde{H}_i)^{-1} \tilde{r}_i \\ &= (1 - \tilde{H}_i)^{-1} \tilde{r}_i \end{aligned}$$

and DFFIT residuals by:

$$\begin{aligned}\text{DFFIT}_i &= \tilde{X}_i\hat{\beta} - \tilde{X}_i\hat{\beta}_{(i)} \\ &= \tilde{X}_i\hat{\beta} - \tilde{X}_i\hat{\beta} + \tilde{H}_i(1-\tilde{H}_i)^{-1}\tilde{r}_i \\ &= \tilde{H}_i(1-\tilde{H}_i)^{-1}\tilde{r}_i\end{aligned}$$

Cook's distance for case i is

$$C_i = \frac{\tilde{r}_i^2\tilde{H}_i}{p(1-\tilde{H}_i)^2}$$

The Leave-M-Out Cross Validation residual would be

$$\begin{aligned}\text{LMOCV}_M &= \tilde{Y}_M - \tilde{X}_M\hat{\beta}_{(M)} \\ &= \tilde{Y}_M - \tilde{X}_M\hat{\beta} + \tilde{H}_M(S_M^{\frac{1}{2}}V^{MM}S_M^{\frac{1}{2}} - \tilde{H}_M)^{-1}\tilde{r}_M \\ &= \tilde{r}_M + \tilde{H}_M(S_M^{\frac{1}{2}}V^{MM}S_M^{\frac{1}{2}} - \tilde{H}_M)^{-1}\tilde{r}_M \\ &= S_M^{\frac{1}{2}}V^{MM}S_M^{\frac{1}{2}}(S_M^{\frac{1}{2}}V^{MM}S_M^{\frac{1}{2}} - \tilde{H}_M)^{-1}\tilde{r}_M\end{aligned}$$

No avoiding inverting m by m matrices. This would be useful for multiple observations per person.

DFFIT and Cook's Distance compare fitted values at two values of the estimate of β. My thesis worked through the algebra for Cook's Distance for cases M after cases G had been removed from the data. ie $\hat{\beta}_{(M,G)} - \hat{\beta}_{(G)}$ where $\text{var}(\hat{\beta})$ is replaced by $\text{var}(\hat{\beta}_{(G)})$. I'm grateful to Tony Lawrance who suggested this work.

Bibliography

Baade Ingrid, PhD thesis, QUT, 1998, Survival Analysis Diagnostics

multiple and conditional deletion diagnostics for general linear models Ingrid A. Baade & Anthony N. Pettitt Communications in statistics, 2000, vol 29, p1899-1910

Plan: try these out on a GLS model. Maybe the one in Harrell, RMS, which is AR(1). Would like to do LOOCV residuals on an elastic net model under GLS, use LOOCV or LMOCV to tune hyperparameter/s.

Appendix

Ordinary Least Squares

Once $\hat{\beta}_{(i)} = \hat{\beta} - (X^TX)^{-1}X_i^T(1-H_i^\circ)^{-1}r_i$ is derived, writing LOOCV_i and DFFIT_i for the OLS model is easy.

$$
\begin{aligned}
\text{LOOCV}_i &= Y_i - X_i\hat{\beta}_{(i)} \\
&= Y_i - X_i\hat{\beta} + H_i^{\circ}(1 - H_i^{\circ})^{-1} r_i \\
&= r_i + H_i^{\circ}(1 - H_i^{\circ})^{-1} r_i \\
&= (1 - H_i^{\circ})^{-1} r_i \\
\text{DFFIT}_i &= X_i\hat{\beta} - X_i\hat{\beta}_{(i)} \\
&= X_i\hat{\beta} - X_i\hat{\beta} + H_i^{\circ}(1 - H_i^{\circ})^{-1} r_i \\
&= H_i^{\circ}(1 - H_i^{\circ})^{-1} r_i
\end{aligned}
$$

By definition, $\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}$. $X_{(i)}^T X_{(i)}$ and $X_{(i)}^T Y_{(i)}$ can be partitioned as $X_{(i)}^T X_{(i)} = X^T X - X_i^T X_i$ and $X_{(i)}^T Y_{(i)} = X^T Y - X_i^T Y_i$. We show

$$
(X_{(i)}^T X_{(i)})^{-1} = (X^T X)^{-1} + (X^T X)^{-1} X_i^T (1 - H_i^{\circ})^{-1} X_i (X^T X)^{-1} \tag{1}
$$

and then substitute this into the definition of $\hat{\beta}_{(i)}$ so the result comes out. Henderson and Searle (1981) suggest the authorship of this result could be attributed to Schur (1917):

$$
(A - UD^{-1}V)^{-1} = A^{-1} + A^{-1}U(D - VA^{-1}U)^{-1}VA^{-1}
$$

Using $A = X^T X$, $D = 1$, $U = X_i^T$ and $V = X_i$ we have equation (1).

Generalised Least Squares

For the GLS we have to invert $X_{(i)}^T V_{(i)}^{-1} X_{(i)}$. We would like to partition this like we did for $X_{(i)}^T X_{(i)}$ for the OLS case, but this time we have an additional term $V_{(i)}^{-1}$. We include the derivation for deletion of M cases at the time. Consider

$$
\begin{aligned}
V^{-1} - V^{M\text{cols}}(V^{MM})^{-1} V^{M\text{rows}} &= \begin{bmatrix} V^{MM} - V^{MM}(V^{MM})^{-1}V^{MM} & V^{M(M)} - V^{MM}(V^{MM})^{-1}V^{M(M)} \\ V^{(M)M} - V^{(M)M}(V^{MM})^{-1}V^{MM} & V^{(M)} - V^{(M)M}(V^{MM})^{-1}V^{M(M)} \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 \\ 0 & V^{(M)} - V^{(M)M}(V^{MM})^{-1}V^{M(M)} \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 \\ 0 & (V_{(M)})^{-1} \end{bmatrix} \qquad (2)
\end{aligned}
$$

Premultiply equation (2) by X^T and postmultiply by X:

$$
\begin{aligned}
X^T(V^{-1} - V^{M\text{cols}}(V^{MM})^{-1}V^{M\text{rows}})X &= X^T \begin{bmatrix} 0 & 0 \\ 0 & (V_{(M)})^{-1} \end{bmatrix} X \\
&= X_{(M)}^T (V_{(M)})^{-1} X_{(M)}
\end{aligned}
$$

$V^{M\text{rows}}X$ is $S_M^{-\frac{1}{2}}\tilde{X}_M$ so

$$\begin{aligned} X_{(M)}^T(V_{(M)})^{-1}X_{(M)} &= X^TV^{-1}X - \tilde{X}_M^T S_M^{-\frac{1}{2}}(V^{MM})^{-1}S_M^{-\frac{1}{2}}\tilde{X}_M \\ &= X^TV^{-1}X - \tilde{X}_M^T(S_M^{\frac{1}{2}}V^{MM}S_M^{\frac{1}{2}})^{-1}\tilde{X}_M \end{aligned}$$

and similarly

$$X_{(M)}^T(V_{(M)})^{-1}Y_{(M)} = X^TV^{-1}Y - \tilde{X}_M^T(S_M^{\frac{1}{2}}V^{MM}S_M^{\frac{1}{2}})^{-1}\tilde{Y}_M$$

We can use the formula $(A-UD^{-1}V)^{-1} = A^{-1} + A^{-1}U(D-VA^{-1}U)^{-1}VA^{-1}$ to invert $X_{(M)}^T(V_{(M)})^{-1}X_{(M)}$, with $A = X^TV^{-1}X$, $U = \tilde{X}_M^T$, $V = \tilde{X}_M$ and $B = S_M^{\frac{1}{2}}V^{MM}S_M^{\frac{1}{2}}$.

This gives

$$(X_{(M)}^T(V_{(M)})^{-1}X_{(M)})^{-1} = (X^TV^{-1}X)^{-1} + (X^TV^{-1}X)^{-1}\tilde{X}_M^T(S_M^{\frac{1}{2}}V^{MM}S_M^{\frac{1}{2}} - \tilde{H}_M)^{-1}\tilde{X}_M(X^TV^{-1}X)^{-1}$$

where $\tilde{H}_M = \tilde{X}_M(X^TV^{-1}X)^{-1}\tilde{X}_M^T$.

After some algebra,

$$\hat{\beta}_{(M)} = \hat{\beta} - (X^TV^{-1}X)^{-1}\tilde{X}_M^T(S_M^{\frac{1}{2}}V^{MM}S_M^{\frac{1}{2}} - \tilde{H}_M)^{-1}\tilde{r}_M$$

and specifically for a single case i,

$$\hat{\beta}_{(i)} = \hat{\beta} - (X^TV^{-1}X)^{-1}\tilde{X}_i^T(1-\tilde{H}_i)^{-1}\tilde{r}_i$$