The Wrong Model

How to spot it

J N S Matthews

Biostatistics Research Group, Newcastle University

Genesis of approach

- Most statistical models make assumptions and the usual approach to checking them is to fit the model and then examine various *diagnostic* quantities.
- Approach depends on software to make fitting a model straightforward and quick.
- Software became available in the early 1980s and led to many new diagnostic quantities being defined, largely focussed on continuous outcomes
- Several books from that era.
  1. Belsley, Kuh & Welsch, 1980
  2. Cook & Weisberg, 1982
  3. Atkinson, 1985
- We will cover some aspects of these now-familiar quantities

Regression equation and familiar results

- Outcomes are in $\boldsymbol{y}$, an n-dim vector
- Design matrix is $\boldsymbol{X}$, an $n \times p$ matrix
- $\boldsymbol{\beta}$ is a p-dim vector of parameters
- $\boldsymbol{\epsilon}$ is an n-dim vector of residuals

and

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Elements of vectors will be denoted in usual way - i.e. y_i is ith element of $\boldsymbol{y}$,

Some assumptions

- We will assume $n > p$ and that $\boldsymbol{X}$ is of rank p
- We will assume that the elements of $\boldsymbol{\epsilon}$ are independent, and
- have zero mean and constant variance σ^2, i.e. $\text{var}(\epsilon_i) = \sigma^2$ or, equivalently, $\text{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$.
- Quite often we assume the ϵ_i follow a Normal distribution, i.e. $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

Basic results

- We have $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ - the least squares estimator
- Fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{y}$, with $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
- Estimated residuals are $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$
- $\boldsymbol{P}$ is projection onto column space of $\boldsymbol{X}$, so $\boldsymbol{P}^2 = \boldsymbol{P}$ - and hence $(\boldsymbol{I} - \boldsymbol{P})^2 = \boldsymbol{I} - \boldsymbol{P}$
- We say $\boldsymbol{P}$ and $\boldsymbol{I} - \boldsymbol{P}$ are *idempotent*.

- Note also $\mathrm{E}[\sum \hat{\epsilon}_i^2] = \mathrm{E}[\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}]$ and this is $\mathrm{E}[\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}] = \mathrm{tr}(\mathrm{E}[(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]) = \sigma^2\mathrm{tr}(\boldsymbol{I} - \boldsymbol{P})$ and this is $(n-p)\sigma^2$, which leads to $\hat{\sigma}^2 = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}/(n-p)$

Note use of $\mathrm{tr}(ABC) = \mathrm{tr}(CAB) = \mathrm{tr}(BCA)$ where $\mathrm{tr}(\cdot)$ is the trace

Do we need Normality?

- If $\boldsymbol{\epsilon}$ are assumed to be Normal, $\hat{\boldsymbol{\beta}}$ is the *maximum likelihood estimator*, and this is known to have optimal properties.
- What if $\text{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$, i.e. constant variance but not Normal?
- Remember the *Gauss-Markov theorem*.
- Among all unbiased estimators of $\boldsymbol{\beta}$ that are linear in $\boldsymbol{y}$, the least squares estimator minimises the variance.
- Consider $\boldsymbol{Ay}$: if unbiased then $\boldsymbol{AX} = \boldsymbol{I}$ and variance is $\sigma^2 \boldsymbol{AA}^T$. Also

$$\boldsymbol{AA}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \boldsymbol{AA}^T - \boldsymbol{AX}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}^T = \boldsymbol{A}(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{A}^T$$

and as this is non-negative definite, the result follows

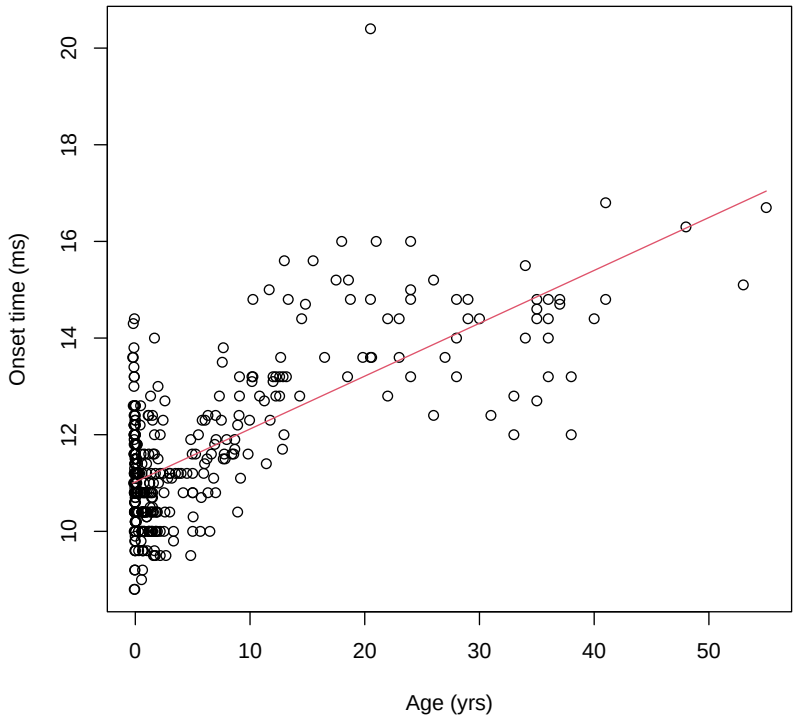
Some consequences for model checking

- $\mathrm{E}[\hat{\boldsymbol{\epsilon}}] = (\boldsymbol{I} - \boldsymbol{P})\mathrm{E}[\boldsymbol{y}] = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X}\boldsymbol{\beta} = 0$
- Also, if the model is right, $\mathrm{cov}[\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{y}}] = \mathrm{E}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{y}}^T] = (\boldsymbol{I} - \boldsymbol{P})\mathrm{E}[\boldsymbol{y}\boldsymbol{y}^T]\boldsymbol{P} = 0$, i.e. the fitted values and residuals are uncorrelated.

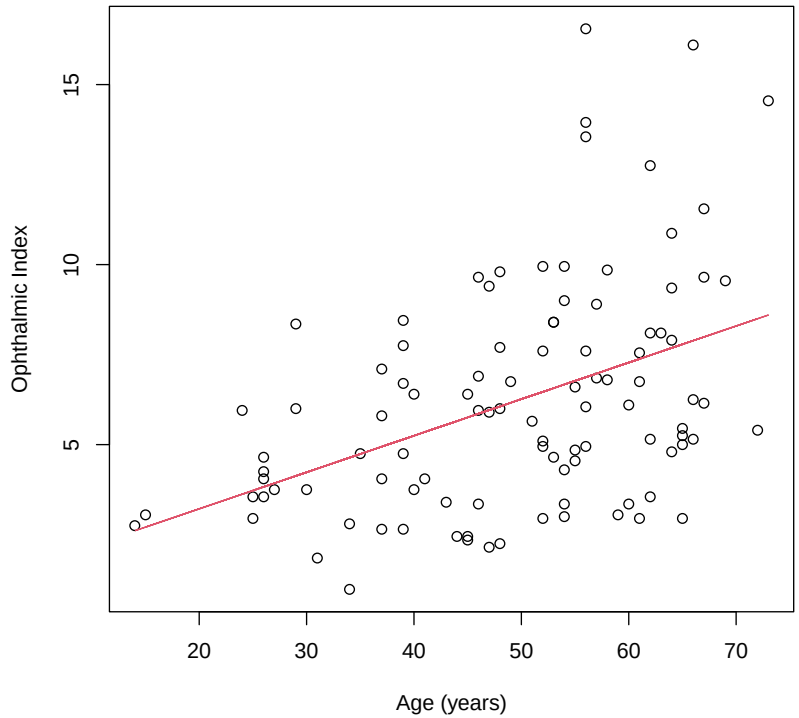
- If $\boldsymbol{x}_j^C$ is the jth column of $\boldsymbol{X}$, then $\boldsymbol{x}_j^{CT}\hat{\boldsymbol{\epsilon}} = \boldsymbol{x}_j^{CT}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = 0$, as $\boldsymbol{x}_j^C$ is trivially in column space of $\boldsymbol{X}$
- In particular $\sum \hat{\epsilon}_i = \mathbf{1}^T\hat{\boldsymbol{\epsilon}} = 0$, *provided* there is an intercept in the model [or something equivalent]. So the *sample* mean of residuals is identically zero if there is an intercept.

Practice I

- Most methods stem from idea that $\hat{\epsilon}$ will be 'like' ϵ
- Inadequacies in $\boldsymbol{X}\boldsymbol{\beta}$ and in assumptions of constant σ^2 are expected to feed into $\hat{\epsilon}$



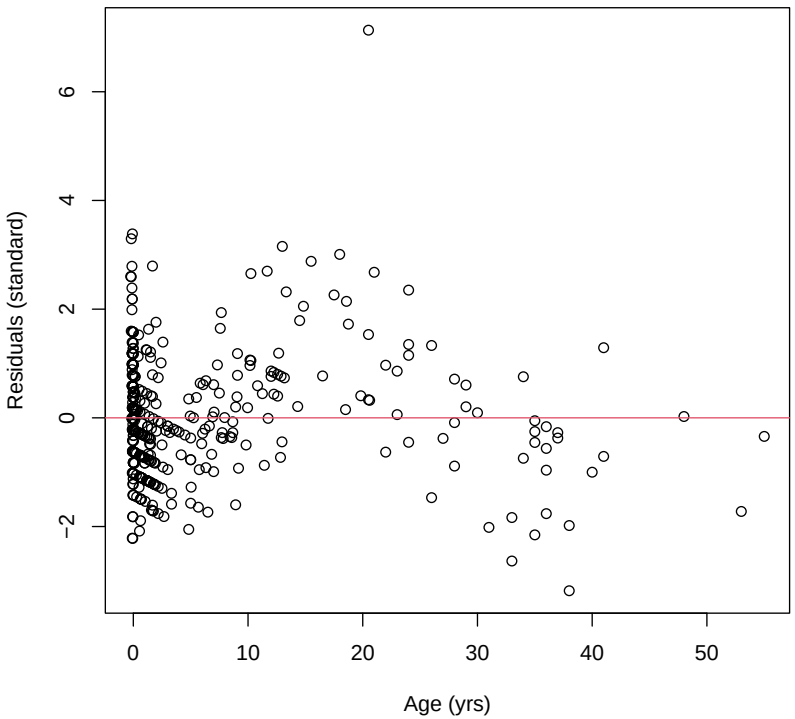
O'Sullivan *et al.*



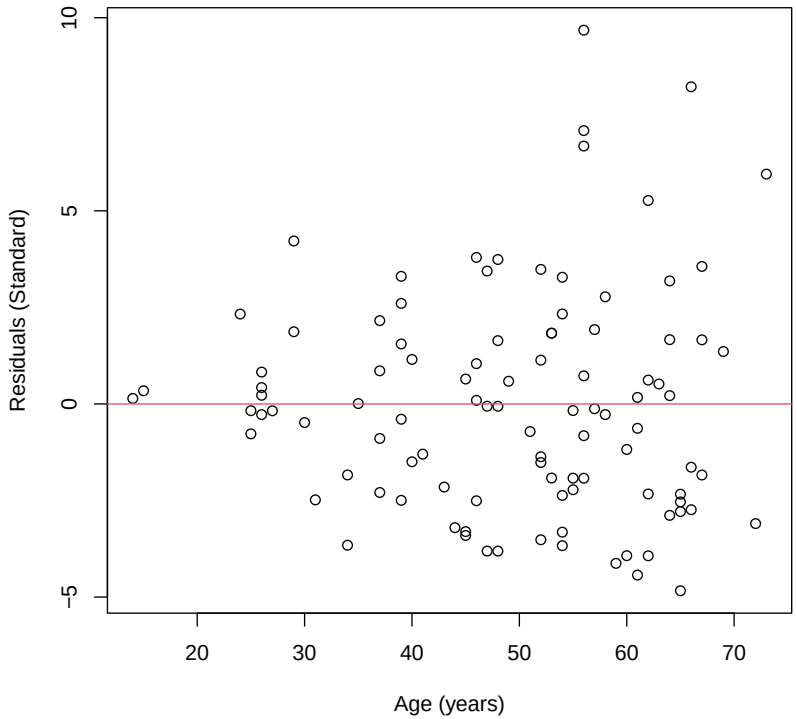
Perros *et al.*

Practice II

- If model is OK, ϵ is unrelated to covariates and has constant variance
- Proceed, at least initially, assuming this also applies to $\hat{\epsilon}$



Non-random pattern



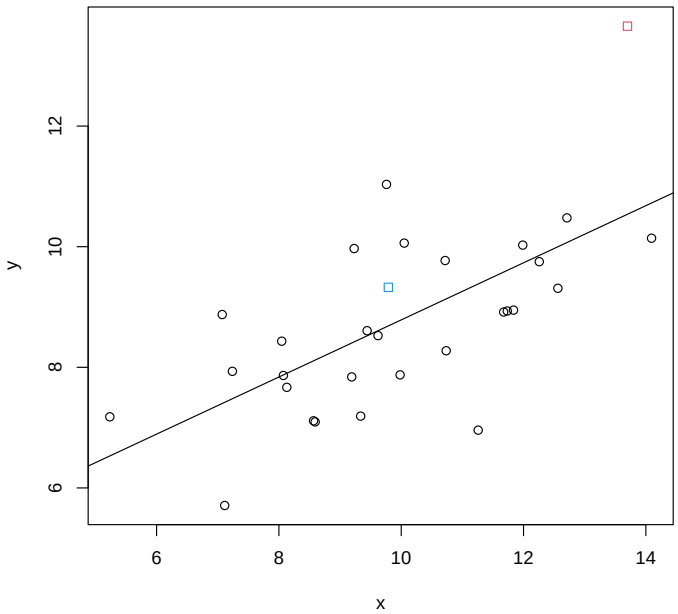
Spread increases with age

Variance of $\hat{\epsilon}$

- Consider the OI data. Spread seems to increase with age
- Assumed spread of ϵ_i does not change with age
- So plot of $\hat{\epsilon}_i$ should reveal this?
- Is the spread of $\hat{\epsilon}_i$ constant? Better check.
- $\text{var}[\hat{\boldsymbol{\epsilon}}] = \text{var}[(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}] = \sigma^2(\boldsymbol{I} - \boldsymbol{P})$
- So $\text{var}(\hat{\epsilon}_i) = \sigma^2(1 - h_i)$, with h_i being ith diagonal element of $\boldsymbol{P}$
- So constant variance of ϵ_i does not, in general, imply constant variance of $\hat{\epsilon}_i$
- Depends on properties of the h_i - the *leverages*
- These feature a lot in the following, so worthy of some exploration

Leverages I

- Leverages depend only on covariates, not on $\boldsymbol{y}$
- They reflect the potential an observation at $\boldsymbol{x}_i$ has to affect $\hat{\boldsymbol{\beta}}$
- Consider fictitious data with outcome y, single covariate x and $n = 30$



- Red square $h_i = 0.139$
- Blue square $h_i = 0.034$

Leverages: $0 \leq h_i \leq 1$

- Any symmetric, idempotent matrix, $\boldsymbol{A}$, is non-negative definite, viz.

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^2\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{A}\boldsymbol{x}) \geq 0$$

- Apply this to $\boldsymbol{A} = \boldsymbol{P}$ and $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{P}$, with $\boldsymbol{x} = \boldsymbol{e}_i$, the vector with a 1 in the ith place and 0 elsewhere: this gives $h_i \geq 0$ and $1 - h_i \geq 0$
- These limits can be achieved - consider $y_i = \gamma x_i + \epsilon_i$, then

$$h_i = \frac{x_i^2}{\sum_k x_k^2}$$

and this can be 0 and can approach 1 arbitrarily closely

Leverages: $\frac{1}{n} \leq h_i \leq 1$

- If the model contains an intercept, then the lower bound increases to $\frac{1}{n}$
- To see this define $\boldsymbol{f}_i = \boldsymbol{e}_i - \frac{1}{n}\mathbf{1}$ and note that $\boldsymbol{P}\mathbf{1} = \mathbf{1}$ as there is an intercept in the model
- Note that $\boldsymbol{f}_i^T \boldsymbol{P} \boldsymbol{f}_i$ must be non-negative and on expansion is $h_i - \frac{1}{n}$

Leverages: $\sum h_i = p$

- This states that $\mathrm{tr}(\boldsymbol{P}) = p$. Can be seen by applying cyclic permutation identity for trace to formula for $\boldsymbol{P}$ in terms of $\boldsymbol{X}$
- A more geometrical approach is to note that $\boldsymbol{P}$ is symmetric, so is orthogonally diagonalisable. This gives $\mathrm{tr}(\boldsymbol{P}) = \sum \lambda_i$, where λ_i are the eigenvalues of $\boldsymbol{P}$
- As $\boldsymbol{P}$ is idempotent, the λ_is are all 0 or 1. Eigenvectors with eigenvalue 1 will form a basis for the range space of $\boldsymbol{X}$, so there are $\mathrm{rank}(\boldsymbol{X}) = p$ eigenvalues equal to 1
- This result gives a handle on the sizes of the h_i as they have mean p/n
- For models with intercepts, a handle on the size of all of the elements of $\boldsymbol{P}$ comes from $\mathbf{1}^T \boldsymbol{P} \mathbf{1} = \mathbf{1}^T \mathbf{1} = n$, so the mean of the off-diagonal elements is $(n-p)/[n(n-1)]$, i.e. of order $1/n$

Standardized residuals

- When assessing constancy of variance of residuals, sensible to remove the variation in the variance of $\hat{\epsilon}_i$ due to factor $1 - h_i$
- The *standardized residual*, r'_i, is

$$r'_i = \frac{\hat{\epsilon}_i}{s\sqrt{1-h_i}} = \frac{y_i - \hat{y}_i}{s\sqrt{1-h_i}}$$

- Most authors take opportunity to make the residual dimensionless by also dividing by s, the estimate of σ
- Reasonable to think this might be approximately standard Normal - 95% of values within (-2,2)
- But not exact. Not even *t*-distributed as denominator is independent of $\hat{y}_i$ but not of y_i

Deletion residuals

- Model-checking can be global (errors in specification) or local (outliers)
- If there is an outlying value, principal concern is for how it affects estimated parameters
- An example is a suspiciously large y_i. This might 'drag' the fitted line towards itself, reducing the size of the residual
- This led to notion of *deletion residuals*, r_i^*
- Analogous to standardized residual, but with fitted value and estimated RMS, s based on dataset with ith point omitted

Deletion residuals - detail

- In following, subscript (i) indicates an estimate/quantity found from data with ith point removed
- So $\hat{\boldsymbol{\beta}}_{(i)}$ is deletion estimate of $\boldsymbol{\beta}$, which has variance $\sigma^2(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}$, with $\boldsymbol{X}_{(i)}$ being design matrix with ith row omitted
- Definition is

$$r_i^* = \frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}}{s_{(i)}\sqrt{1+\boldsymbol{x}_i^T(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i}}$$

- Could find simply by refitting regression n times but there is a more elegant and insightful approach

Deletion residuals - technicalities

- Worth noting the following three items
  1. $\boldsymbol{X}^T\boldsymbol{X} = \sum_j \boldsymbol{x}_j\boldsymbol{x}_j^T$
  2. $\boldsymbol{X}^T\boldsymbol{y} = \sum_j \boldsymbol{x}_j y_j$
  3. $\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i = h_i$
- and also recall the Woodbury identity

$$(\boldsymbol{A} - \boldsymbol{U}\boldsymbol{V}^T)^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{I} - \boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1}$$

where $\boldsymbol{A}$ is a $p \times p$ matrix and $\boldsymbol{U}, \boldsymbol{V}$ are $p \times m$ matrices

Deletion residuals - more technicalities

- Putting $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{U} = \boldsymbol{V} = \boldsymbol{x}_i$ gives

$$1 + \boldsymbol{x}_i^T(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i = 1/(1-h_i)$$

- and

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{\epsilon}_i}{1-h_i}$$

- and

$$(n-p-1)s_{(i)}^2 = s^2(n-p-r_i'^2)$$

- And finally

$$r_i^* = r_i'\sqrt{\frac{n-p-1}{n-p-r_i'^2}}$$

- Shrunk towards zero if $\mid r_i' \mid < 1$, opposite if $\mid r_i' \mid > 1$. N.b. $\mid r_i' \mid < \sqrt{n-p}$, so can be neither Normal nor t

Distributional properties of residuals: standard residuals

[hereinafter, we assume an intercept in the model]

- Consider the usual, standard, residual $\hat{\epsilon}_i$.
- These have expectation 0 and variance $\sigma^2(1-h_i)$
- They are dependent - $\sum \hat{\epsilon}_i = 0$ (plus another $p-1$ similar constraints)
- Correlation for cases i, j is $-h_{ij}/(1-h_i)(1-h_j)$
- Normalising as $\hat{\epsilon}_i/s$ gives values still summing to 0, etc., still independent of $\hat{\boldsymbol{y}}$ but not t-distributed

Properties: deletion residuals

$$r_i^* = \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}}{s_{(i)} \sqrt{1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} \boldsymbol{x}_i}}$$

- These are t-distributed ($n - p - 1$ df): numerator and denominator independent ($s_{(i)}$, independent of fitted value and y_i)
- So they have mean 0 and variance $(n - p - 1)/(n - p - 3)$
- Deletion residuals do not sum to 0, nor are they independent of the fitted values $\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}$

Properties: standardized residuals I

$$r_i' = \frac{\hat{\epsilon}_i}{s\sqrt{1-h_i}} = \frac{y_i - \hat{y}_i}{s\sqrt{1-h_i}}$$

- These will not sum to 0 but are functions of residuals alone, being, apart from factors depending on n, p

$$r_i' \propto \frac{\hat{\epsilon}_i}{\sqrt{\sum \hat{\epsilon}_k^2}}$$

  so are independent of the $\hat{y}_i$
- They do have expectation zero but a proof is quite delicate, as the numerator and denominator are dependent.

Properties: standardized residuals II

- If $\boldsymbol{\epsilon}$ is Normal, then $\hat{\boldsymbol{\epsilon}}$ is multivariate Normal, $N_n(\mathbf{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{P}))$, then $\hat{\boldsymbol{\epsilon}}$ and $-\hat{\boldsymbol{\epsilon}}$ have the same distribution, so

$$\mathsf{E}\left(\frac{\hat{\epsilon}_i}{\sqrt{\sum_j \hat{\epsilon}_j^2}}\right) = \mathsf{E}\left(\frac{-\hat{\epsilon}_i}{\sqrt{\sum_j (-\hat{\epsilon}_j)^2}}\right) = -\mathsf{E}\left(\frac{\hat{\epsilon}_i}{\sqrt{\sum_j \hat{\epsilon}_j^2}}\right),$$

  and so $\mathsf{E}(r_i') = 0$.
- Similar argument applies if one only assumes $\hat{\boldsymbol{\epsilon}}$ is symmetric

Exact distribution of r_i'

We start with

$$\frac{r_i'^2}{n-p} = \frac{r_i^{*2}}{n-p-1+r_i^{*2}}$$

and recall r_i^* is t on $n-p-1$ df.

- Remember $Z/\sqrt{\chi^2_\nu/\nu}$ is t, with numerator and denominator independent
- Z^2 is χ^2 on 1 df
- A χ^2 variable is a form of Gamma variable
- If U, V are independent Gammas with same scale parameter, then $U/(U+V)$ and $U+V$ are independent, with the former a Beta variable and the latter a Gamma.

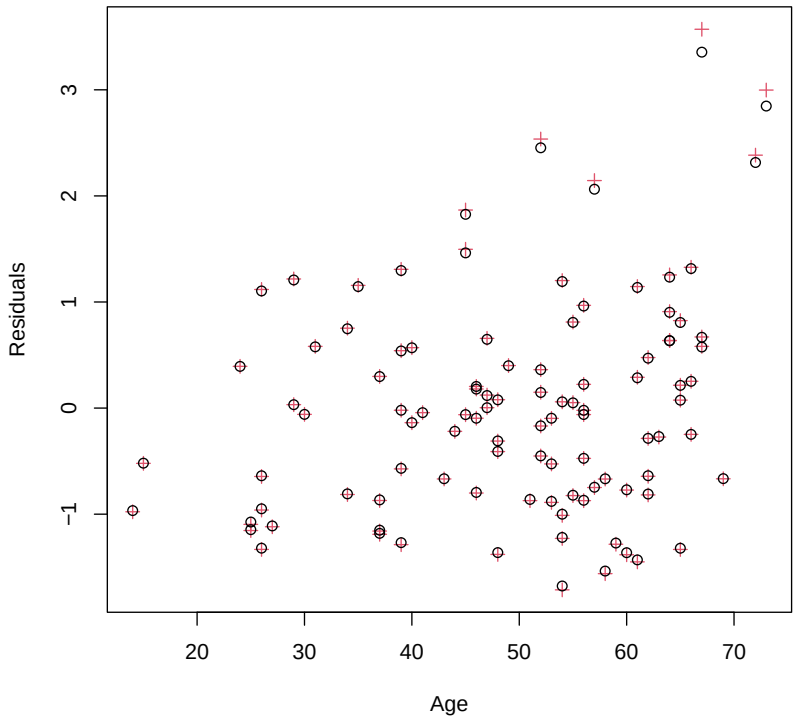
Exact distribution of r'_i, continued

From the above we find

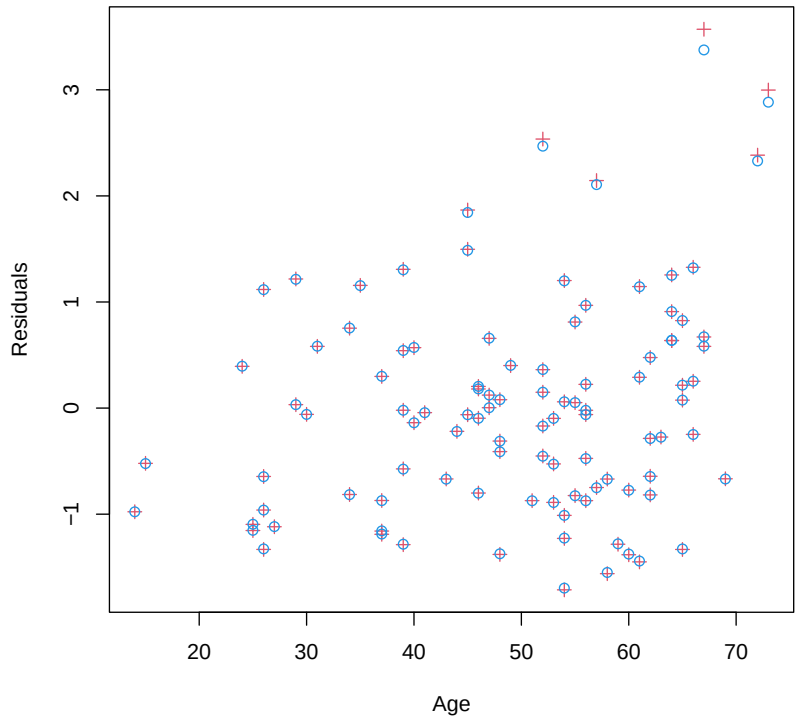
- r'^2_i is $n-p$ times a Beta variable with parameters $\frac{1}{2}$ and $\frac{1}{2}(n-p-1)$.
- From this we have $\mathrm{E}[r'^2_i] = 1$
- Can be shown correlation between r'_i, r'_j is $-h_{ij}/(1-h_i)(1-h_j)$

Does it matter?

- To compare all residuals directly, need scaled residual, $\hat{\epsilon}_i/s$ so all dimensionless
- Consider OI data



scaled black, deletion red



standardized blue, deletion red

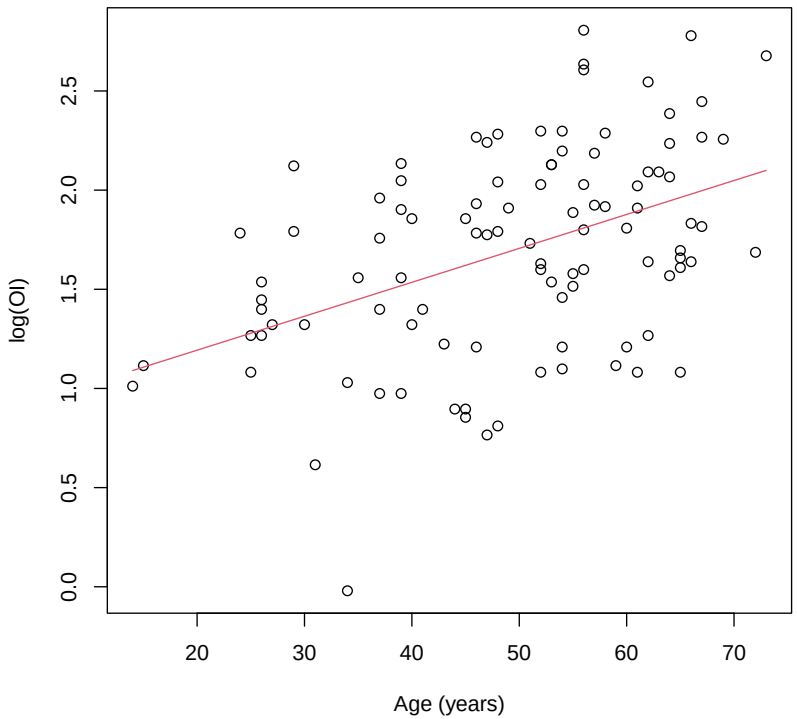
What to do next?

- Different types of residuals differ little
- Clear from all that for OI data, σ^2 increases with age.
- What should be done about it?

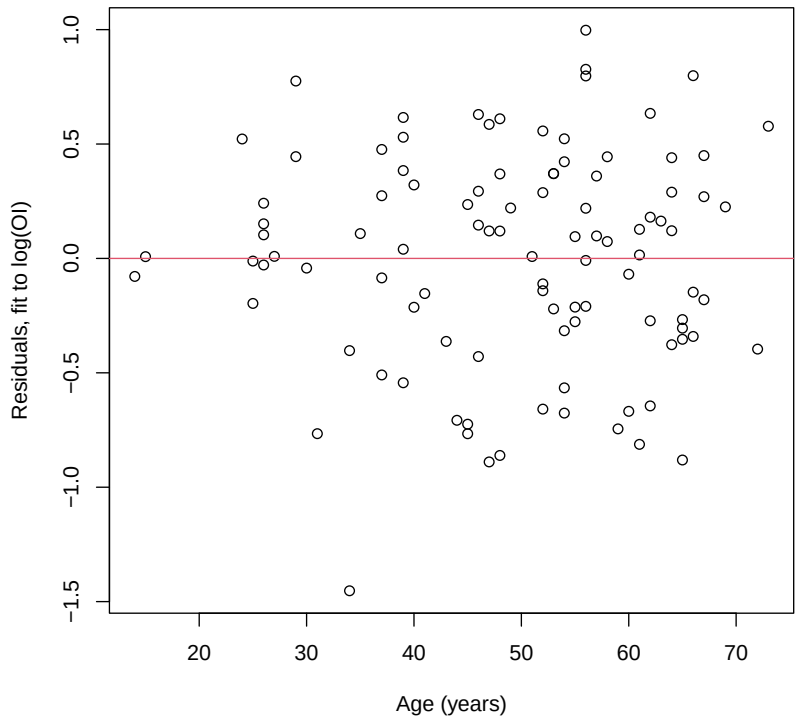
Model revision

- If some of the diagnostic plots suggest model inadequacies, what to do?
- Should not adopt an approach that is too algorithmic
- Diagnostic plots might guide thinking but changes should be rooted in the context
- Increasing OI with age *and* increasing variance - perhaps error variance is multiplicative?
- ?Try modelling log OI

Trying log OI



Regression



$\hat{\epsilon}_i$

Looks better - can probably stop model revision here.

What isn't in this lecture

- Cook's statistic - measures effect of individual point on parameter estimates

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{ps^2} = \frac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})^T (\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})}{ps^2}$$
$$= \frac{r_i'^2 h_i}{p(1-h_i)},$$

- Very relevant aim but I haven't used them much
  1. Considers all parameters - would often want to focus on particular items, e.g. treatment effects
  2. Puts items on a dimensionless scale but often more meaningful on original scale
  3. Seldom interested in looking at omission of all points, rather than a few suspicious ones

Normal probability plots

- Main omission is no assessment of whether the ϵ_i are Normal?
- Usually done with Normal probability plots
- Although there are tests of Normality
- This is for another exciting episode