Weighted Least Squares

- when the noise is different for each instance i
- recall: OLS assumes that noise variance σ^2 is equal for all i
- maximum likelihood approach: $p(\{TS\}) = \prod_{i=1}^{N} p_i(Y_i \mid X_i) = \prod_{i=1}^{N} \mathcal{N}(Y_i \mid X_i\beta, \sigma_i^2)$

  $\prod_{i=1}^{N}$: instances are independent; p_i: prob. depends on instance; σ_i^2: μ_i

$$-\log p(\{TS\}) = \sum_{i=1}^{N} \frac{1}{2} \frac{(Y_i - X_i\beta)^2}{\sigma_i^2} + \sum_{i=1}^{N} \frac{1}{2} \log \sigma_i^2 + \underbrace{\sum_{i=1}^{N} \log \sqrt{2\pi}}_{\text{independent of model} \Rightarrow \text{drop}}$$

- 2 cases: σ_i^2 is known or not known
- <u>case 1</u>: - σ_i^2 is known, e.g. from a pilot experiment that characterises the measurement device

  $\Rightarrow \sum_{i=1}^{N} \frac{1}{2} \log \sigma_i^2$ is constant, i.e. independent of the model β $\Rightarrow$ drop

  define the covariance matrix as diagonal matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_N^2 \end{pmatrix}$$

weighted least squares objective:

$$\boxed{\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \frac{(Y_i - X_i\beta)^2}{\sigma_i^2} \overset{\arg\min}{=} (Y - X\beta)^T \Sigma^{-1} (Y - X\beta)}$$

$$\frac{\partial \text{Loss}}{\partial \beta} = -2 X^T \Sigma^{-1} (Y - X\beta) \overset{!}{=} 0 \Rightarrow \text{weighted normal equations } \boxed{(X^T \Sigma^{-1} X)\beta = X^T \Sigma^{-1} Y}$$

formal solution : weighted pseudo-inverse $\boxed{\hat{\beta} = \underbrace{(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}}_{\text{weighted pseudo-inverse}} Y}$

reduce to OLS by scaling : $\tilde{X}_i = \frac{X_i}{\sigma_i}$ $\quad \tilde{Y}_i = \frac{Y_i}{\sigma_i}$

$\boxed{\hat{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}}$ OLS of new variables

$\Rightarrow$ data preparation: scaling & centering $\tilde{X}_i = \frac{X_i - \bar{X}}{\sigma_i}$ $\quad \tilde{Y}_i = \frac{Y_i - \bar{Y}}{\sigma_i}$

weighted pseudo-inverse also works if Σ is not diagonal, i.e. the noise between different instance can be correlated, but we must know correlation

- case 2 σ_i^2 are unknown $\Rightarrow$ we must learn them along with β
  more model parameters $\theta = \{\beta, \{\sigma_i^2\}_{i=1}^N\}$
  mixed supervised ~~learning~~ and unsupervised learning: we have supervision for $Y_i \approx X_i \beta$, but no training information for σ_i^2
- $\sum_{i=1}^N \frac{1}{2} \log \sigma_i^2$ is now dependent on the model and cannot be dropped

$\Rightarrow$ heteroscedastic loss, Dawid-Sebastiani score

$$\boxed{\hat{\beta}, \{\hat{\sigma}_i^2\}_{i=1}^N = \arg\min_{\beta, \{\sigma_i^2\}} \sum_i \left[\frac{(Y_i - X_i\beta)^2}{\sigma_i^2} + \cancel{\tfrac{1}{2}} \log \sigma_i^2 \right]}$$

Solve by <u>alternating optimization</u>: fundamental approach when the optimization cannot be solved analytically, but the unknown parameters can be split into two groups (here: $(\beta, \{\sigma_i^2\}) = (\theta_1, \theta_2)$)

generic strategy: ⓪ initialization: make an initial guess $\theta_2^{(0)}$

① for $t = 1, \dots, T$

I. optimize for $\theta_1^{(t)}$, treating $\theta_2^{(t-1)}$ as constants

II. optimize for $\theta_2^{(t)}$, treating $\theta_1^{(t)}$ as constants

advantage: I and II are often much simpler than the full optimization, may be analytically tractable

disadvantage: the final solution is generally not the global optimum of full problem

application to heteroscedastic loss: <u>Iteratively Reweighted Least Squares</u>

abbreviate: $\sigma_i^2 = \tau_i$ ⓪ initialization: $\tau_i^{(0)} = 1$

① for $t = 1, \dots, T$:

I: solve for β, keeping $\tau_i^{(t-1)}$ fixed

$\Rightarrow$ weighted LSQ (case 1) $\beta^{(t)} = \arg\min_\beta \sum_{i=1}^N \frac{(Y_i - X_i\beta)^2}{\tau_i^{(t-1)}}$

II solve for τ_i, keeping $\beta^{(t)}$ fixed

define residual: $r_i = Y_i - X_i \beta^{(t)}$

$\{\tau_i^{(t)}\} = \arg\min_{\{\tau_i\}} \sum_{i=1}^N \left[\frac{r_i^2}{\tau_i} + \log \tau_i \right]$

$$\frac{\partial\,\text{Loss}}{\partial \tau_i} = -\frac{r_i^2}{\tau_i^2} + \frac{1}{\tau_i} \overset{!}{=} 0 \quad \Rightarrow \hat{\tau}_i^{(t)} = r_i^2$$

- how to choose T (# of iterations)? – IRLS in theory requires only $T=2$, but due to possible numerical inaccuracies, bit bigger T is safer $T=5$, $T=10$
  - in general for alternating optimization: numerical analysis may provide a theory about convergence speed $\Rightarrow$ choose T accordingly, otherwise trial & error
- heteroscedastic loss can also be used for non-linear optimization

$$Y_i = f(X_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma_i^2 = g(X_i))$$

train two non-linear predictors e.g. neural networks