# Heteroskedastic Linear Regression

***Junteng Jia;*** *jj585@cornell.edu* *August 12, 2019*

**Abstract.** Ordinary least square regression can be understood as the maximal likelihood fit of a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ to a linear function, assuming the observation noise variance is the same for all data points. In this report, we assume the noise variance is also a function of the independent variable $\mathbf{x}_i$. We derive the log-likelihood expression for observing $\mathcal{D}$, as well as the log-likelihood derivatives for gradient based maximum likelihood estimation (MLE) of model parameters.

**Model.** Consider the scenario where the independent variable $x_i$ denotes "amount of workload" of a job, and $y_i$ denotes its "time cost". Assume the workload can be divided into infinitesimal units, and each unit is processed independently with constant expected time. Then the central limit theorem tell us that the relationship between $x_i$ and $y_i$ can be written as,

$$y_i = kx_i + b + (\alpha + \beta x_i)^{1/2} \cdot \mathcal{N}(0, 1) \tag{1}$$

where $k$ is per unit time cost, $b$ is the overhead for launching the job, and $(\alpha + \beta x_i)$ is the time variance. Then the log-likelihood of observing $\mathcal{D}$ is,

$$\begin{aligned}
\log \mathcal{L}(\mathcal{D}|k, b, \alpha, \beta) &= \log \left[ \prod_i P(y_i|x_i) \right] \\
&= \log \left\{ \prod_i \frac{\exp\left[-\frac{|y_i - kx_i - b|^2}{2(\alpha+\beta x_i)}\right]}{[2\pi(\alpha + \beta x_i)]^{1/2}} \right\} \\
&= -\frac{1}{2} \sum_i \left[ \log(2\pi) + \log(\alpha + \beta x_i) + \frac{|y_i - kx_i - b|^2}{(\alpha + \beta x_i)} \right].
\end{aligned} \tag{2}$$

For maximum log-likelihood estimation, we have

$$\begin{aligned}
\max_{k,b,\alpha,\beta} \log \mathcal{L}(k, b, \alpha, \beta) &= \max_{\alpha,\beta} \max_{k,b} \log \mathcal{L}(k, b, \alpha, \beta) \\
&= -\min_{\alpha,\beta} \min_{k,b} - \log \mathcal{L}(k, b, \alpha, \beta) \\
&= -\min_{\alpha,\beta} \Omega(\alpha, \beta).
\end{aligned} \tag{3}$$

It turns out, the inner minimization problem has a closed-form solution,

$$\begin{aligned}
\arg\min_{k,b} - \log \mathcal{L}(k, b, \alpha, \beta) &= \arg\min_{k,b} \sum_i \frac{|y_i - kx_i - b|^2}{(\alpha + \beta x_i)} \\
&= \arg \min_{\mathbf{w} \equiv [k,b]} (\mathbf{y} - \mathbf{X}\mathbf{w})^\intercal \mathbf{U} (\mathbf{y} - \mathbf{X}\mathbf{w}) \\
&= (\mathbf{X}^\intercal \mathbf{U} \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{U} \mathbf{y}
\end{aligned} \tag{4}$$

where $\mathbf{X} \in \mathbb{R}^{n \times 2}$ is the feature matrix with each row $\mathbf{X}_{i,:} = [x_i, 1]$; $\mathbf{y} \in \mathbb{R}^n$ is the vector of labels; $\mathbf{U} \in \mathbb{R}^{n \times n}$ with $U_{ij} = (\alpha + \beta x_i)^{-1}$ if $i = j$ and $U_{ij} = 0$ otherwise.

**Computation.** Now let's consider eq. (3), where we need to minimize $\Omega(\alpha, \beta)$. Substituting eq. (2) into eq. (4) gives the **objective function**,

$$\Omega(\alpha, \beta) = \frac{n}{2} \log(2\pi) + \frac{1}{2} \mathtt{Tr}[\log(\mathbf{U}^{-1})] + \frac{1}{2} \left[ \mathbf{y}^\intercal \mathbf{U} \mathbf{y} - \mathbf{y}^\intercal \mathbf{U} \mathbf{X} (\mathbf{X}^\intercal \mathbf{U} \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{U} \mathbf{y} \right] \tag{5}$$

Furthermore, we can also compute the analytical **objective function gradients**,

$$\begin{aligned}
\frac{\partial \Omega}{\partial \alpha} =& \frac{1}{2}\texttt{Tr}(\mathbf{U}) - \frac{1}{2}\mathbf{y}^\intercal\mathbf{U}\mathbf{U}\mathbf{y} + \mathbf{y}^\intercal\mathbf{U}\mathbf{U}\mathbf{X}(\mathbf{X}^\intercal\mathbf{U}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{U}\mathbf{y} \\
&- \frac{1}{2}\mathbf{y}^\intercal\mathbf{U}\mathbf{X}(\mathbf{X}^\intercal\mathbf{U}\mathbf{X})^{-1}(\mathbf{X}^\intercal\mathbf{U}\mathbf{U}\mathbf{X})(\mathbf{X}^\intercal\mathbf{U}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{U}\mathbf{y} \\
\frac{\partial \Omega}{\partial \beta} =& \frac{1}{2}\texttt{Tr}(\mathbf{U}\mathbf{P}) - \frac{1}{2}\mathbf{y}^\intercal\mathbf{U}\mathbf{P}\mathbf{U}\mathbf{y} + \mathbf{y}^\intercal\mathbf{U}\mathbf{P}\mathbf{U}\mathbf{X}(\mathbf{X}^\intercal\mathbf{U}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{U}\mathbf{y} \\
&- \frac{1}{2}\mathbf{y}^\intercal\mathbf{U}\mathbf{X}(\mathbf{X}^\intercal\mathbf{U}\mathbf{X})^{-1}(\mathbf{X}^\intercal\mathbf{U}\mathbf{P}\mathbf{U}\mathbf{X})(\mathbf{X}^\intercal\mathbf{U}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{U}\mathbf{y} \qquad (6)
\end{aligned}$$

where $\mathbf{P} \in \mathbb{R}^{n\times n}$ with $P_{ij} = x_i$ if $i = j$ and $P_{ij} = 0$ otherwise. Since the objective function is not convex, restarting from multiple initial $[\alpha_0, \beta_0]$ is needed if a convex optimization solver is used in minimizing $\Omega(\alpha, \beta)$.