Problem 1 - Least Squares

Define $\mathbf{b} = \begin{bmatrix} b_0 & b_1 & b_2 \dots b_m \end{bmatrix}$ where $\mathbf{b} \in \mathbb{R}^n$. Similarly, define $\mathbf{x}_i = \begin{bmatrix} 1 & x_{i1} & x_{i2} \dots x_{in} \end{bmatrix}$ where $\mathbf{x}_i \in \mathbb{R}^n$. Here we have defined $x_{i0} = 1$.

The vertical distance between a point $(y_i, \mathbf{x}_i)$ and the hyperplane $\mathbf{x}_i^T \mathbf{b}$ is $y_i - (\mathbf{x}_i^T \mathbf{b})$, $i = \{1, 2, \dots m\}$.

By summing the squared terms for every point y_i, we get the RSS function

$$RSS = \sum_{i=1}^{m} \left[y_i - \left(\mathbf{x}_i^T \mathbf{b} \right) \right]^2$$

If we define $X = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_m \end{bmatrix}$ and $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$, then we can minimize RSS with respect to $\mathbf{b}$ to get

$$\mathbf{b} = \left(X^T X \right)^{-1} XY$$

Problem 2 - Optimization

Please see `Problem2\problem.1.2.R` for work.

Problem 3 - Interpretation

See `Q3_considerations.R` for all calculations and marketing plan.

Problem 4 - Weighted Regression

1. We consider the model $y = X\beta + \epsilon$, where

$$\epsilon \sim \mathcal{N} \left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_m^2 \end{bmatrix} \right)$$

where m is the total number of observations.

W is a matrix withw_i on the diagonal and zeroes everywhere else, which corresponds to the reciprocal $w_i = \frac{1}{\sigma_i^2}$. Then, we have

$$W = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ & & \ddots & \vdots \\ \dots & \dots & & \\ 0 & 0 & \dots & \frac{1}{\sigma_m^2}. \end{bmatrix}$$

We can now minimize

$$WRSS = \frac{1}{n}(y - Xb)^T W(y - Xb).$$

Take the gradient of $WRSS$ and set it to zeros.

$$\begin{aligned} \frac{\partial WRSS}{\partial b} &= \frac{1}{n} X^T W(y - Xb) = 0 \\ &\implies \frac{1}{n} X^T (Wy - WXb) = 0 \\ &\implies \frac{1}{n} X^T Wy - X^T WXb = 0 \\ &\implies \frac{1}{n} X^T Wy = X^T WXb \\ &\implies b = \frac{1}{n} \left(X^T WX\right)^{-1} X^T Wy \end{aligned}$$

2. It prioritizes fitting to points with smaller variance than those with higher.

3. We want to weight the terms with a smaller variance σ_i^2 higher than terms with a smaller variance. In this case, $\frac{1}{\sigma_i^2} > \frac{1}{\sigma_j^2} \iff \sigma_i^2 < \sigma_j^2$.