Nachtrag gewichtete Regression, Aufgabe 3

Man geht vom Modell: $Y = X\boldsymbol{\beta} + \epsilon$ aus mit $\epsilon \sim \mathcal{N}(0, \sigma^2\Sigma)$. Dabei ist Σ bekannt, aber σ^2 unbekannt.

Rückführung auf unser bekanntes Modell:

Wir definieren die Matrix A, so dass $AA^T = \Sigma$.

$$\tilde{Y} = A^{-1}Y = A^{-1}(X\boldsymbol{\beta} + \epsilon) = \underbrace{A^{-1}X}_{\tilde{X}}\boldsymbol{\beta} + \underbrace{A^{-1}\epsilon}_{\tilde{\epsilon}} = \tilde{X}\boldsymbol{\beta} + \tilde{\epsilon}$$

Dabei gilt nun, wie in der Übungsstunde angetönt:

$$\begin{aligned} Cov[\tilde{\epsilon}] &= Cov[A^{-1}\epsilon] = A^{-1}Cov[\epsilon](A^{-1})^T \\ &= A^{-1}\sigma^2(AA^T)(A^{-1})^T = \sigma^2\mathbb{I} \end{aligned}$$

Damit sind wir wieder in der Standardsituation.

Spezialfall: Gewichtete Regression

Ist der Spezialfall, wo Σ diagonal ist, z.B.:

$$\Sigma = \begin{pmatrix} d_1^2 & 0 & 0 \\ 0 & d_2^2 & 0 \\ 0 & 0 & d_3^2 \end{pmatrix} \Longrightarrow A = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix}$$

Führt man nun Kleinste Quadrate durch, so wird $\boldsymbol{\beta}$ folgendermassen bestimmt:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{X}_i\boldsymbol{\beta})^2 &= \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} ((A^{-1}Y)_i - (A^{-1}X)_i\boldsymbol{\beta})^2 \\ &= \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{d_i^2}(Y_i - X_i\boldsymbol{\beta})^2 \qquad (1) \end{aligned}$$

Dies ist der Grund, wieso es gewichtete Regression heisst. Die Gleichung sieht aus, wie eine normale Kleinste Quadrate Regression, einfach werden die einzelnen Terme nun gewichtet mit den Gewichten $\frac{1}{d_i^2}$.

Bei Aufgabe 3 entspechen die v_i unseren d_i, daher sind die Gewichte $\frac{1}{v_i^2}$ zu wählen, analog zu (1).

Ich hoffe es ist nun ein wenig klarer.

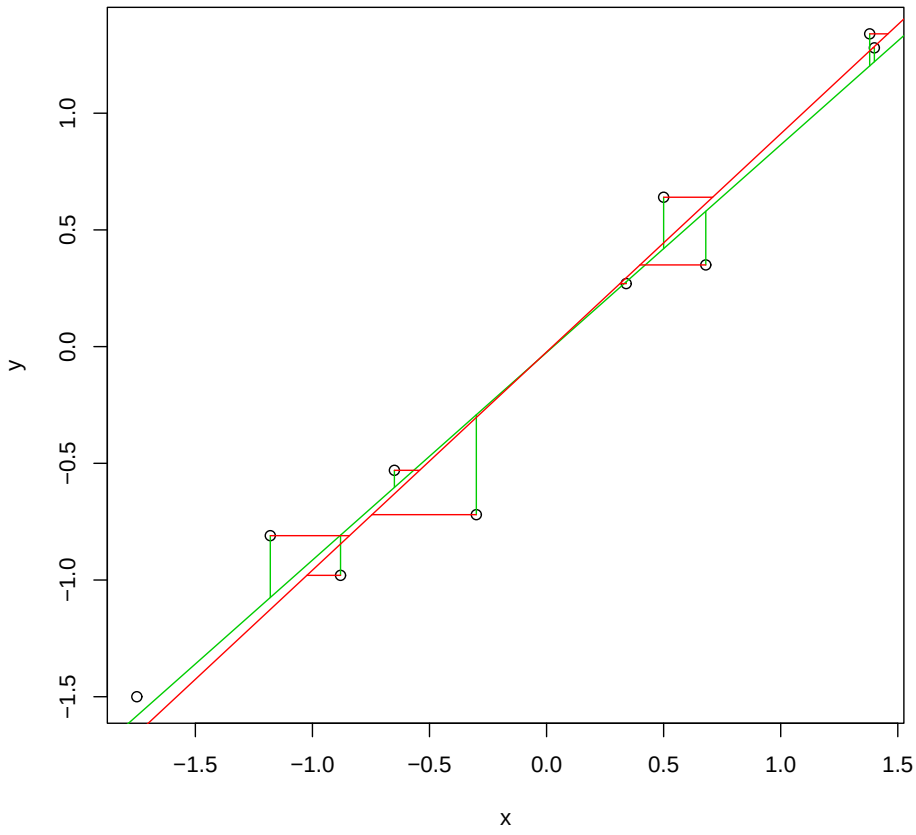
Nachbesprechung Serie 2:

Aufgabe 1

Ein häufiger Fehler war, dass man $\log N = c$ korrekt berechnet hat, und dann daraus geschlossen hat, dass $N = 10^c$. Dies ist nicht richtig, da R standarmässig den natürlichen Logarithmus nimmt. Das heisst die korrekte Lösung wäre: $N = e^c$.

Aufgabe 2

Die meisten haben herausgefunden, dass die beiden geraden von $lm(y \sim x)$ und $lm(x \sim y)$ nicht übereinstimmen. Den Grund dazu sieht man in der unteren Abbildung. In der Kleinsten Quadrate Regression, wie wir das uns gewohnt sind, wird die Gerade so angepasst, dass die grünen Abschnitte minimiert werden. Vertauscht man nun x un y, so werden die roten Abschnitte minimiert.Man sieht in der Zeichnun, dass dies eine andere Lösung ergibt und man versteht intuitiv, dass die rote Gerade steiler sein muss, als die grüne.



Aufgabe 3

Das Hauptproblem in Aufgabe 3 habe ich in der Übungsstunde bereits angesprochen. Das war die predict() - Funktion, welche Probleme bereitet hat.