HW1

Juan Arroyo Miranda

Problem 1

In order to answer this problem, we are going to make the following assumptions regarding the residuals (errors) within the framework of Linear Regression:

- Errors have mean 0 - that is $E(u) = 0$
- Errors are uncorrelated with the data. This assumption also implies that errors are uncorrelated with any linear function of the data.

  Plot A: From the plot, it seems that the residuals are symmetric around 0. That is to say, $E(u) = 0$. However, the residuals could be represented as a function of the data. Perhaps, a function such as $u = a \cdot sin(x)$ for some constant a. This violates our second assumption and therefore, under our assumptions, we cannot conclude that the residuals are the result of fitting least squares.

  Plot B: The plot shows a positive linear relationship between the residuals and the data. This is a violation of our second assumption (no correlation between data and residuals), and therefore it is not a plausible plot of the residuals of a least squares regression.

  Plot C: The plot shows that all the values for the residuals lie above 0. This implies that the expected value of the residuales is greater than zero. Once again, we violate one of the assumptions of the least squares and therefore, this is not a plausible plot of the residuals of a least squares regression.

  Plot D: The plot shows that the spread of the residuals is symmetric around zero. This satisfies the first condition above. With regards to the correlation of residuals with the data, it seems that, errors have no correlation with the data. A good way to think about (y_i, x_i) is the following:

  When we fit the model, the (y_i, x_i) pairs are either above or below the line fitted by least squares. This would result in the pattern of residuals we observe in plot d).

Problem 2

In this case, we need to minimize the following loss function with respect to **w**:

$$L(\mathbf{w}, \mathbf{X}, \mathbf{y'}) = L(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(y'_i - \mathbf{w} \cdot x_i)^2$$

Taking the derivative with respect to **w**, we arrive to the following solution in a similar way as we did in class $\mathbf{w'} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y'}$

$$\begin{aligned}
\mathbf{w'} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y'}, \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(a \cdot \mathbf{y} + \vec{b}), \\
&= a \cdot (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{b}, \\
&= a \cdot \mathbf{w^*} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{b}
\end{aligned}$$

where $\vec{b} = [b, b, \ldots, b]^T$. We can think about the expression $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{b}$ as the result of the regression of $\vec{b}$ on **X**. We can use the following result from least squares regression:

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY = \begin{pmatrix} \bar{Y} - \bar{X}\,Cov(X,Y)/Var(X) \\ Cov(X,Y)/Var(X) \end{pmatrix} = \begin{pmatrix} \bar{Y} - \bar{X}\hat{\beta} \\ \hat{\beta} \end{pmatrix} \text{where } \hat{\beta} \text{ is a vector of weights}$$

When we are regressing on a constant all the weights are equal to zero. Therefore, $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{b}$ reduces to the following vector $[b, 0, 0, \ldots, 0]^T$.

So, we can express $\mathbf{w'} = a \cdot \mathbf{w^*} + [b, 0, 0, \ldots, 0]^T$.

Problem 3

In this case, we minimize the following problem with respect to **w**

$$L(\mathbf{w}, \mathbf{X}, \mathbf{y'}) = L(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(y'_i - \mathbf{w} \cdot \tilde{x_i})^2$$

where $\tilde{x_i} = c_j x_i$

Taking the derivative with respect to **w**, we arrive to the following solution: $\tilde{\mathbf{w}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{y}$

where

$$\tilde{\mathbf{X}} = \begin{pmatrix} c_1x_{11} & c_2x_{12} & \cdots & c_dx_{1d} \\ c_1x_{21} & c_2x_{22} & \cdots & c_dx_{2d} \\ c_1x_{31} & c_2x_{32} & \cdots & c_dx_{3d} \\ \vdots & \ddots & \ddots & \vdots \\ c_1x_{d1} & c_2x_{d2} & \cdots & c_dx_{dd} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ x_{31} & x_{32} & \cdots & x_{3d} \\ \vdots & \ddots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix} \begin{pmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & c_d \end{pmatrix} = \mathbf{XC}$$

Here, we assume that **X** is a square matrix with dimensions $d \times d$ and that **C** is a square diagonal matrix with c_j in the elements of the diagonal.

Therefore, the inverse of this matrix exists and is given by

$$\mathbf{C^{-1}} = \begin{pmatrix} 1/c_1 & 0 & \cdots & 0 \\ 0 & 1/c_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/c_d \end{pmatrix}$$

We can rewrite the expression $(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T$ as

$$\begin{aligned}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T &= ((\mathbf{XC})^T(\mathbf{XC}))^{-1}(\mathbf{XC})^T \\ &= (\mathbf{C}^T\mathbf{X}^T\mathbf{XC})^{-1}\mathbf{C}^T\mathbf{X}^T\end{aligned}$$

Let $A = \mathbf{C}^T\mathbf{X}^T$ and $B = \mathbf{XC}$ and use the following definition $(AB)^{-1} = B^{-1}A^{-1}$ (Eq. (1)) to compute

$$\begin{aligned}(\mathbf{C}^T\mathbf{X}^T\mathbf{XC})^{-1} &= (\mathbf{XC})^{-1}(\mathbf{C}^T\mathbf{X}^T)^{-1} \\ &= \mathbf{C}^{-1}\mathbf{X}^{-1}\mathbf{X^T}^{-1}\mathbf{C^T}^{-1} \\ &= \mathbf{C}^{-1}(\mathbf{X^T X})^{-1}\mathbf{C^T}^{-1} \quad \text{by Eq. (1)}\end{aligned}$$

dNow, we can replace this in our original expression for $\tilde{\mathbf{w}}$

$$\begin{aligned}\tilde{\mathbf{w}} &= (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{y} \\ &= \mathbf{C}^{-1}(\mathbf{X^T X})^{-1}\mathbf{C^T}^{-1}\mathbf{C^T X^T y} \\ &= \mathbf{C}^{-1}(\mathbf{X^T X})^{-1}\mathbf{I X^T y} \\ &= \mathbf{C}^{-1}(\mathbf{X^T X})^{-1}\mathbf{X^T y} \\ &= \mathbf{C}^{-1}\mathbf{w^*}\end{aligned}$$

Therefore, we can get $\tilde{\mathbf{w}}$ without looking at the data. We just need the matrix $\mathbf{C^{-1}}$ and $\mathbf{w^*}$.

Problem 4

In this case the Maximum Likelihood estimator would be given by

$$\widehat{\mathbf{w}}_{ML} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma_x)$$

Under the assumptions 1, this becomes

$$\widehat{\mathbf{w}}_{ML} = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{N} \frac{1}{\sigma_x\sqrt{2\pi}} \exp\left(-\frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{2\sigma_x^2}\right)$$

Taking the log, we can express the log- likelihood in the following way

$$
\begin{aligned}
\log p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma_x) &= \log \prod_{i=1}^{N} \frac{1}{\sigma_x\sqrt{2\pi}} \exp\Big(-\frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{2\sigma_x^2} \Big) \\
&= \sum_{i=1}^{N} \left[\log\left(\frac{1}{\sigma_x\sqrt{2\pi}} \right) - \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{2\sigma_x^2} \right] \\
&= \sum_{i=1}^{N} \left[-\log\sigma_x\sqrt{2\pi} - \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{2\sigma_x^2} \right] \\
&= -\sum_{i=1}^{N} \log\sigma_x\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{2\sigma_x^2}
\end{aligned}
$$

Looking at the expression above, we see that $\sum_{i=1}^{N} \log\sigma_x\sqrt{2\pi}$ is independent of **w**, and we only need to maximize **w** with the expression on the right-hand side.

$$
\begin{aligned}
\underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(y_i|x_i;\mathbf{w},\sigma_x) &= \underset{\mathbf{w}}{\operatorname{argmax}} -\sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{\sigma_x^2} \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{\sigma_x^2}
\end{aligned}
$$

Taking the derivative with respect to w_j, we get the following

$$
\begin{aligned}
\frac{\delta}{\delta w_j}\mathbf{L}(\mathbf{w}) &= 2\sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))}{\sigma_x^2} \frac{\delta}{\delta w_j} f(\mathbf{x}_i;\mathbf{w}) = 0 \\
&= \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))}{\sigma_x^2} \frac{\delta}{\delta w_j} f(\mathbf{x}_i;\mathbf{w}) = 0
\end{aligned}
$$

In the expression above, $\sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{\sigma_x^2} \frac{\delta}{\delta w_j} f(\mathbf{x}_i;\mathbf{w})\ \sigma_{\mathbf{x}}^2$ is the same for every **x** and does not depend on **w**, and it is unknown.

If $\sigma_{\mathbf{x}}^2$ does not depend on any individual $\mathbf{x}_i$ and we assume is constant, then the solution for the minimization problem becomes $\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i;\mathbf{w}))\frac{\delta}{\delta w_j} f(\mathbf{x}_i;\mathbf{w}) = 0$, which has the same solution as $\mathbf{w}^*$.

Problem 5

If we know the values of the noise variance $\sigma_{x_i}^2$ at every training input in $\mathbf{x}_i$ the equation in problem 4 becomes the following:

First, let $\sigma_i = \sigma_{x_i}^2$ for $i = 1, \ldots, N$, where σ_i is a known constant and

$$
\begin{aligned}
\log p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma_i) &= \log \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\Bigg(- \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{2\sigma_i^2} \Bigg) \\
&= \sum_{i=1}^{N} \left[\log\left(\frac{1}{\sigma_i\sqrt{2\pi}} \right) - \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{2\sigma_i} \right] \\
&= \sum_{i=1}^{N} \left[- \log \sigma_i\sqrt{2\pi} - \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{2\sigma_i} \right] \\
&= - \sum_{i=1}^{N} \log \sigma_i\sqrt{2\pi} - \frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{\sigma_i}
\end{aligned}
$$

$$
\begin{aligned}
\underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(y_i|x_i;\mathbf{w},\sigma_i) &= \underset{\mathbf{w}}{\operatorname{argmax}} - \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{\sigma_i} \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{\sigma_i}
\end{aligned}
$$

Taking the derivative with respect to w_j, we get the following

$$
\begin{aligned}
\frac{\delta}{\delta w_j}\mathbf{L}(\mathbf{w}) &= 2 \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))}{\sigma_i^2} \frac{\delta}{\delta w_j} f(\mathbf{x}_i;\mathbf{w}) = 0 \\
&= \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))}{\sigma_i^2} \frac{\delta}{\delta w_j} f(\mathbf{x}_i;\mathbf{w}) = 0
\end{aligned}
$$

Therefore, maximum likelihood estimator for this noise model **w'** can be estimated, but it will be different from $\mathbf{w^*} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.