# Ibrahim Gabr - 12144496

## Problem 1

The two central assumptions regarding residuals (errors) within the framework of Linear Regression are as follows:

- Errors have mean 0 - that is $E(e) = 0$
- Errors are uncorrelated with the data. Specifically, the errors are uncorrelated with any linear function of the data.

In the plots, we are supposedly plotting $e_i$ vs $x_i$. Let us now inspect these plots and see if the assumptions above hold.

**Plot A:** Visual inspection shows no linear correlation between the errors and the data. Furthermore, it seems that the errors are symmetric around 0. That is to say, $E(e) = 0$. As such, this plot does indeed show $e_i$ vs $x_i$. The dependency of would be either $\sin$ or $\cos$.

**Plot B:** Visual inspections shows a linear correlation between the errors and the data. This is a violation of the above assumptions and as such, we can immediately say that this is **not** a plot of $e_i$ vs $x_i$. In addition, we can see a clear linear correlation in this plot.

**Plot C:** Visual inspection shows that all points like above a prediction error of 0. As such, it is impossible for $E(e) = 0$. As such, this is **not** a plot of $e_i$ vs $x_i$.

**Plot D:** Visual inspection shows that spread of the residuals are 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. That is to say, our variance is constant. The dependency could be described as having multiple y values for a given x. These would be separated a by an "empty band" of space with no points. The linear regression line would be drawn in this "empty band".

## Problem 2

Let us start with writing out the explicit form of Penrose pseudoinverse:

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

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 $\mathbf{X}$. We can use the following result from least squares regression:

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

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

This problem is similar to the one above, however, we know manipulate the matrix $\boldsymbol{X}$. The questions stipulates that that every $\boldsymbol{x_{ij}} = \boldsymbol{cx_{ij}}$. In order for this to happen, we must multiply the matrix $\boldsymbol{X}$ by a matrix $\boldsymbol{C}$. This matrix $\boldsymbol{C}$ has a 1 at $\boldsymbol{c_{1,1}}$ followed by the constant $\boldsymbol{c}$ on the main diagonal from $\boldsymbol{c_{2,2}}$ to $\boldsymbol{c_{i,j}}$ where $\boldsymbol{i = j}$.

Since $\boldsymbol{C}$ is a diagonal matrix and we can assume that $\boldsymbol{c_{ij} \neq 0}$ we know that the inverse $\boldsymbol{C^{-1}}$ exists. $\boldsymbol{C^{-1}}$ Simply contains the reciprocal of all values on the diagonal on $\boldsymbol{C}$. That is, $\frac{1}{c_{ij}}$.

Let us now recall the Moore-Penrose pseudoinverse: $\boldsymbol{w^* = (X^TX)^{-1}X^T\mathbf{y}}$

This can be re-written 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 us now recall the matrix rule: $\boldsymbol{(AB)^{-1} = B^{-1}A^{-1}}$. Let us say that $\boldsymbol{A = C^TX^T}$ and $\boldsymbol{B = XC}$.

Let us also call our new parameter vector $\tilde{\boldsymbol{w}}$ which is constructed from the Moore-Penrose pseudoinverse using $\tilde{\boldsymbol{X}}$. As such:

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

Substituting this into $\tilde{\boldsymbol{w}}$ we get:

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

From the above, we can see that in order to calculate $\tilde{\mathbf{w}}$ we simply need $\boldsymbol{C^{-1}}$ and $\mathbf{w^*}$. This shows that we do **not** need to look at the data, that is $\boldsymbol{X}$, in order to find $\tilde{\mathbf{w}}$.

## Problem 4

Using the slides from class we see that the Maximum Likelihood Estimator is as follows:

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

Taking the partial derivative with respect to $\mathbf{w}$ of this expression would be unwieldily. As such, we can take the log of this expression in order to transform it into a sum. Recall $Log(AB) = Log(A) + Log(B)$.

This transforms our expression to:

$$\widehat{w}_{ML} = -\frac{1}{N} \sum_{i=1}^{N} \log \sigma_x \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i; \mathbf{w}))^2}{2\sigma_x^2}$$

Note that this is the *Normalized Log Likelihood*.

The red terms in the above equation are **independent** of $w$. As such, they play no role in solving for the argmax of w. As such, we are left with:

$$Eq(1) : \underset{\mathbf{w}}{\operatorname{argmax}} - \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i; \mathbf{w}))^2}{2\sigma_x^2}$$

Furthermore, the variance of x, while unknown, is constant and independent of $w$. Thus it can be pulled out of them sum. This leaves us with:

$$\underset{\mathbf{w}}{\operatorname{argmax}} - \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i; \mathbf{w}))^2$$

Mathematically, we know this is equivalent to minimizing the log-loss:

$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i; \mathbf{w}))^2$$

Notice that the negative $(-)$ is no longer present. From this, we know it is possible to take the partial derivative of $L$ with respect to $w_j$ and thus solve for $w^*$.

**Assumption:** I am following the assumption that $\sigma_x^2$ does **not** depend on any $x_i$. If this assumption does not hold true, then we will be **unable** to solve for $w^*$

## Problem 5

Problem 5 is a similar variation to that of problem 4. However, this time, the variance of each $x_i$ is known and as such, is not constant. We cannot simply pull out the denominator as we did in $EQ(1)$. However, we are still able to solve for $w^*$ as we have each $\sigma_{x_i}^2$. Mathematically this would be:

$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i; \mathbf{w}))^2}{\sigma_{x_i}^2}$$

We could then take the derivative with respect to $w_j$ and solve for $w^*$.

## Problem 6 - 8

Please see Jupyter Notebook!