Problem 1

Assume the sample size is sufficiently large and the plots only show small part of the picture.

To verify whether the above plots are showing the distribution of residuals, we need to check whether residuals satisfies the following OLS assumptions related to error term:

1. The expected value of the mean of the error terms should be zero given the values of independent variables.

   i.e. $E(e|X) = 0$

2. There is homoscedasticity: the error terms in the regression should all have the same variance

   i.e. $V(e|X) = \sigma^2$

3. There is no autocorrelation: the error terms of different observations should not be correlated with each other.

   i.e. $Cov(e_i, e_j) = 0 \; for \; i \neq j$

(b) and (c) don't show the residuals of least square regression because:

(b) does not satisfies the assumption of no autocorrelation. There seems to be positive autocorrelation.

(c) does not satisfies the assumption of zero mean

For (a), Y is a sinusoidal function or a cubic polynomial function of X.

For (d), the relationship between Y and X is something like some up and down points on a spark line.

End of Problem 1

Problem 2

From least square solution, we know:

$$w^* = (X^T X)^{-1} X^T y = X^+ y$$

Substitute y with $y' = ay + b$, where $a \; and \; b$ are both scalar. We get w':

$$\begin{aligned} w' &= (X^T X)^{-1} X^T (ay + b) \\ &= a(X^T X)^{-1} X^T y + b(X^T X)^{-1} X^T \\ &= aw^* + bX^+ \end{aligned}$$

→$w' = aw^* + bX^+$. We can't compute it directly from w^*. To get w', we still need to get pseudoinverse of X

End of Problem 2

Problem 3

From least square solution, we know:

$$w^* = (X^T X)^{-1} X^T y = X^+ y$$

Assuming $X \in R^{n \times d}, y \in R^{n \times 1}, w \in R^{d \times 1}$

Substitute X with $\tilde{X} = XC$, where $C \in R^{d \times d}$ is a diagonal matrix with $c_1, c_2 .. c_d$ on tis diagonal. We get $\widetilde{w}$:

$$\begin{aligned} \widetilde{w} &= ((XC)^T (XC))^{-1} (XC)^T y = (C^T X^T X C)^{-1} C^T X^T y \\ &= C^{-1} X^{-1} (X^T)^{-1} (C^T)^{-1} C^T X^T y = C^{-1} X^{-1} (X^T)^{-1} X^T y \end{aligned}$$

$$= C^{-1}(X^T X)^{-1} X^T y = C^{-1} X^+ y$$
$$= C^{-1} w^*$$

→$\widetilde{w} = C^{-1} w^*$. We can compute it directly from w^* by left multiply it by C^{-1}

End of Problem 3

Problem 4

Maximum Likelihood Estimation is give as:

$$\widehat{w} = \underset{w}{\text{argmax}}\ p\,(y|X; w, \sigma_x)$$

Because under Gaussian noise model, data is i.i.d:

$$\widehat{w} = \underset{w}{\text{argmax}} \prod_{i=1}^{N} \frac{1}{\sigma_x \sqrt{2\pi}} \exp\,(-\frac{\big(yi - f(xi; w)\big)^2}{2{\sigma_x}^2})$$

Take log on both side. Max the above equation is equivalent to max log likelihood

$$\widehat{w} = \text{argmax} \sum_{i=1}^{N} -ln\sigma_x - \frac{\big(yi - f(xi; w)\big)^2}{2{\sigma_x}^2} - ln\big(\sqrt{2\pi}\big)$$

$-ln\sigma_x$ and $-ln\big(\sqrt{2\pi}\big)$ are independent of w, so maximizing the above equation is equivalent to:

$$\text{argmin} \sum_{i=1}^{N} \frac{\big(yi - f(xi; w)\big)^2}{{\sigma_x}^2}$$

Take the derivative w.r.t w:

$$\frac{\partial \sum_{i=1}^{N} \frac{\big(yi - f(xi; w)\big)^2}{{\sigma_x}^2}}{\partial w} = 0$$

→This is definitely different from minimizing squared loss. We can't compute the MLE because we can't solve the above equation without knowing what exactly is the value of σ_x for each xi. In practice, likelihood methods are most useful when the distribution of the data can be written using a relatively small number of parameters which are same for all the observations (when OLS works). A unique variance for each observation would lead to more parameters than observations.

End of Problem 4

Problem 5

Following problem 4, when we know the value of each σ_{xi}, we can describe how the variance varies, maybe as a function of x. For example, $\sigma = g(x)$. Such that:

$$\frac{\partial \sum_{i=1}^{N} \frac{\big(yi - f(xi; w)\big)^2}{{\sigma_{xi}}^2}}{\partial w} = 0$$

→

$$\frac{\partial \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \frac{\left(yi - f(xi; w^*)\right)^2}{g(x)^2}}{\partial w} = 0$$

We can get $w^*$by solving the above equation, but it is computationally inefficient and will be much more complicated than OLS estimation.

End of Problem 5

Problem 6

I select the model (with degree 3) with minimum symmetric validation loss because that model has less over fitting problem, thus it has less symmetric validation error and predict more accurately.

End of Problem 6

Problem 7

We use gradient descent to calculate the value of loss because our loss function is not closed. I select the model (with degree 3) with minimum asymmetric validation loss because that model has less over fitting problem, thus it has less asymmetric validation error and predict more accurately.

End of Problem 7

Problem 8

Best degree (sym):3

Asym loss: 10.9697

Sym loss: 22.1586

Log-likelihood (symmetric): -2.4930

Best degree (asym):3

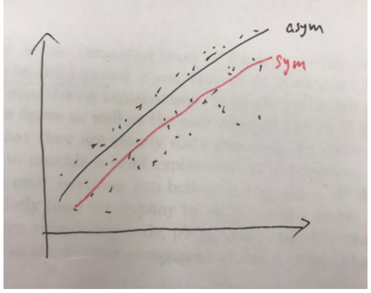
Asym loss: 3.7321

Sym loss: 48.7741

Log-likelihood (symmetric): -2.8875

Observation (compare asym loss and sym loss for both sym and asym model):

Both asym model and sym model has less asym loss compared to sym loss. This is intuitively explainable because we gave add less loss value for underestimation. So the asym fitted line will be above the sym fitted line (as it shown in the graph below)



To choose between models, there is two things to concern: the first is of course the accuracy, and the second is the computational efficiency. i.e. how long it will take to get the results.

Form the perspectives of accuracy in prediction: we can see that asym model is better because it has the lowest loss value (3.7) among all model and all loss evaluation method. In addition, the Log-likelihood for asy model is larger. Furthermore, we would naturally prefer asmy model because of the assumption that it is worse to overestimate.

From the perspectives of computational efficiency: we usually want to choose a model that requires less computation time. The asym model takes much more time than the sym model because of the gradient descent method.

Overall, we would have a trade off between accuracy and calculation speed. But in this case, we would go with asym model because it's accuracy is much higher. But computation time is okay. But if we have larger dataset, we might have different decision.

End of Problem 8