JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

DEPARTMENT OF MATHEMATICS AND ACTUARIAL SCIENCE

STA 2408 REGRESSION MODELLING II: CAT I & CAT II

DECEMBER, 2024

CAT I

QUESTION ONE

(i) **Copy and complete the ANOVA Table** **(TOTAL MARKS: 6 mark)**

a. **Calculate the Total Sum of Squares Regression** (SS_{Total}): **(1 mark)**

Since $R^2 = \frac{SS_{\text{Reg}}}{SS_{\text{Total}}}$, we have;

$$SS_{\text{Reg}} = R^2 \times SS_{\text{Total}}$$

Rearranging to solve for SS_{Total} we have;

$$SS_{\text{Total}} = \frac{SS_{\text{Res}}}{1 - R^2} = \frac{410.4965}{1 - 0.9315} = \mathbf{5992.6496}$$

b. **Calculate SS_{Reg}:** **(1 mark)**

$$SS_{\text{Reg}} = R^2 \times SS_{\text{Total}} = 0.9315 \times 5992.6496 = \mathbf{5582.1532}$$

c. **Degrees of Freedom (df)**: **(1 mark)**
For the regression, $df = 3$ (number of predictors), and for the residuals, $df = n - 1 - 3$. To estimate this, we have;

$$df_{residual} = \frac{410.4965}{14.1216} = \mathbf{29}$$

The total df is $= n - 1$=33-1=**32** or 3+29=**32**

d. **Calculate the Mean Square for Regression** (MS_{Reg}): **(1 mark)**

$$MS_{\text{Reg}} = \frac{SS_{\text{Reg}}}{df_{\text{Reg}}} = \frac{5582.1531}{3} = \mathbf{1860.7177}$$

e. **Calculate the F-statistic**: **(1 mark)**

$$F = \frac{1860.7177}{14.1216} = 131.7639$$

f. **The completed values for the table are;** **(1 mark)**

Model	Sum of Squares	df	Mean Square	F	P-value
Regression	5582.1531	3	1860.7177	131.7639	0.0000
Residual	410.4965	29	14.1216	-	-
Total	5992.6496	32	-	-	-

(ii) **What was the sample size?** (2 marks)

The total degrees of freedom $\text{df}_{\text{Total}} = n - 1$. From the table, $\text{df}_{\text{Total}} = 32$, so,

$$n = 32 + 1 = 33$$

Therefore, the sample size is 33.

(iii) **How many independent variables were in the fitted model?** (2 marks)

The df for regression is p = 3, where p is the number of independent variables. Hence, there are 3 independent variables in the fitted model.

(iv) **Interpret the ANOVA Table Using the P-value** (5 marks)

Hypotheses:

(a) **Null Hypothesis** (H_0)
The model with the independent variables does not explain the variation in Y, implying that the regression coefficients are zero. That is, ($\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$)

(b) **Alternative Hypothesis** (H_1)
The model explains a significant portion of the variation in Y. That is, at least one of the regression coefficients is not zero.

Interpretation of the p-value
The p-value is 0.0000, and this is significantly less than 0.050. This indicates that we reject the null hypothesis and conclude that the independent variables in the model explain a significant portion of the variation in Y. Also, the high R^2 value (0.9315) and the significant p-value show that the model fits the data well, and the independent variables collectively have a statistically significant relationship with the dependent variable.

QUESTION TWO

(i) **Linearisation technique** (TOTAL MARKS: 15 marks)
Table 1 shows data, fitted values,and derivatives evaluated at each observation.

Table 1: Data, fitted values, and derivatives at $\hat{\theta}_0 = [1, 1]'$

i	x_i	y_i	f_i^0	$y_i - f_i^0$	Z_{i1}^0	Z_{i2}^0
1	1	4	1	3	1	0
2	3	9	3	6	3	3.2958
3	2	5	2	3	2	1.3863
4	3	8	3	5	3	3.2958
5	4	15	4	11	4	5.5452
6	5	19	5	14	5	8.0472
7	3	9	3	6	3	3.2958
8	2	8	2	6	2	1.3863
9	4	12	4	8	4	5.5452
10	3	7	3	4	3	3.2958

Fill Table 3 as follows:

(a) For f_i^0, use;

$$f_i^0 = f(x_i, \theta_0) = \alpha x_i^{\beta}$$

So, for $i = 1$, we have;

$$f_1^0 = (1)(1)^1 = 1$$

For $i = 2$, we have;

$$f_2^0 = (1)(3)^1 = 3$$

Finally, for $i = 10$, we have;

$$f_{10}^0 = (1)(3)^1 = 3$$

(2 marks. i.e., 0.2 mark for each correct answer)

(b) Furthermore, for $y_i - f_i^0$, we have;
For $i = 1$;

$$y_1 - f_1^0 = 4 - 1 = 3$$

For $i = 2$, we have;

$$y_2 - f_2^0 = 9 - 3 = 6$$

Finally, for $i = 10$, we have;

$$y_{10} - f_{10}^0 = 7 - 3 = 4$$

(2 marks. i.e., 0.2 mark for each correct answer)

(c) For the Z_{ij}^0, the partial derivatives wrt α and β are as follows;
For α, treat x_i^β as a constant since it does not depend on α, and we have;

$$\frac{\partial y}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left(\alpha x_i^\beta\right).$$

Since α is multiplied by x_i^β, the derivative wrt α is;

$$\frac{\partial y}{\partial \alpha} = x_i^\beta.$$

For β, treat α as a constant and differentiate x_i^β wrt β, given as;

$$\frac{\partial y}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\alpha x_i^\beta\right).$$

Using the chain rule and the formula for differentiating exponential terms , we have;

$$\frac{d}{d\beta} x_i^\beta = x_i^\beta \ln(x_i),$$

and this gives;

$$\frac{\partial y}{\partial \beta} = \alpha x_i^\beta \ln(x_i).$$

So, the partial derivatives wrt the parameters are;
α:

$$\frac{\partial y}{\partial \alpha} = x_i^\beta.$$

and
β:

$$\frac{\partial y}{\partial \beta} = \alpha x_i^\beta \ln(x_i).$$

For Z_{ij}^0, we have;

$$Z_{11}^0 = x_i^\beta$$

and

$$Z_{12}^0 = \alpha x_i^\beta \ln(x_i)$$

(d) **Estimating $Z_{i1}^0 = x_i^\beta$, with $\alpha = 1$, and $\beta = 1$**
From Table 3, the first observation on x is $x_1 = 1$ and we have;

$$Z_{11}^0 = (1)^1 = 1$$

For $i = 2$, $x_2 = 3$ and we have;

$$Z_{21}^0 = (3)^1 = 3$$

Finally, for $i = 10$, $x_{10} = 3$ and we have;

$$Z_{101}^0 = (3)^1 = 3$$

(2 marks. i.e., 0.2 mark for each correct answer)

(e) **Estimating $Z_{i2}^0 = \alpha x_i^\beta \ln(x_i)$, with $\alpha = 1$, and $\beta = 1$**
From Table 3, the first observation on x is $x_1 = 1$, we have;

$$Z_{12}^0 = (1)(1)^1 ln(1) = 0$$

For $i = 2$, $x_2 = 3$ and we have;

$$Z_{22}^0 = (1)(3)^1 ln(3) = 3 * 1.0986 = 3.2958$$

Finally, for $i = 10$, $x_{10} = 3$ and we have;

$$Z_{102}^0 = (1)(3)^1 ln(3) = 3 * 1.0986 = 3.2958$$

The process is continued and the Z_{ij}^0 are estimated at the observed values of x_i.
(2 marks, i.e., 0.2 mark for each correct answer)

(f) Having filled in Table 4, $Z_{ij}^0 s$ are collected into the matrix Z_0, given as;

$$Z_0 = \begin{pmatrix} 1 & 0 \\ 3 & 3.2958 \\ 2 & 1.3863 \\ 3 & 3.2958 \\ 4 & 5.5452 \\ 5 & 8.0472 \\ 3 & 3.2958 \\ 2 & 1.3863 \\ 4 & 5.5452 \\ 3 & 3.2958 \end{pmatrix}$$

and the product, $Z_0^T Z_0$, is;

$$Z_0^T Z_0 = \begin{pmatrix} 102.000 & 129.692 \\ 129.692 & 173.549 \end{pmatrix}$$

(2 marks)

(g) The matrix, Y_0 is given as;

$$Y_0 = \begin{pmatrix} 3 \\ 6 \\ 3 \\ 5 \\ 11 \\ 14 \\ 6 \\ 6 \\ 8 \\ 4 \end{pmatrix}$$

and the product, $Z_0^T Y_0$, is;

$$Z_0^T Y_0 = \begin{pmatrix} 230.000 \\ 299.708 \end{pmatrix}$$

(2 marks)

(h) So,

$$\hat{\beta} = \begin{pmatrix} 102.000 & 129.692 \\ 129.692 & 173.549 \end{pmatrix}^{-1} \begin{pmatrix} 230.000 \\ 299.708 \end{pmatrix} = \begin{bmatrix} 1.187 \\ 0.840 \end{bmatrix}$$

(2 marks)

(i) Therefore, the revised estimates, $\hat{\theta}_1$ are;

$$\hat{\theta}_1 = \hat{\beta}_0 + \theta_0 = \begin{bmatrix} 1.187 \\ 0.840 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.187 \\ 1.840 \end{bmatrix}$$

(1 mark)

(ii) **R programme implementation** (5 marks)

```
1. Load the data from the Table 2
    x<- c(1, 3, 2, 3, 4, 5, 3, 2, 4, 3)
    y<-c(4, 9, 5, 8, 15, 19, 9, 8, 12, 7)

2. Initial parameter guesses
    alpha <- 1
    beta <- 1
    theta <- c(alpha, beta)

3. Fit the Non-linear model
    nls_model<-nls(y~alpha*x*exp(beta), start=list(theta1= 1, theta2=1))

4. Summary of the fitted model
    summary(nls_model)

5. Get estimated coefficients or print estimates
    coef(nls_model)
```

QUESTION THREE

(a) **Model Definition**

(i) **Model Equation** (3 marks)

The model equation for the mixed-effects is;

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \epsilon_{ij}$$

where;

- Y_{ij} is the test score for student i in school j
- β_0 is the fixed intercept
- β_1 is the fixed effect of the predictor X
- $u_j \sim N(0, \sigma_u^2)$ represents the random effect for school j
- $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is the residual error term.

(ii) **Explanations** (2 marks)

(a) The fixed effect β_1 measures the overall association between the predictor X (e.g., hours studied) and the test score, which is assumed to be constant across all schools.

(b) The random effect u_j represents school-specific deviations from the overall intercept, capturing the variation in test scores attributable to differences between schools.

(b) **Final Model Based on Estimates**

(α) **Final mixed model based on estimates obtained** (2 marks)

The model equation for the mixed-effects model based on the estimates is;

$$Y_{ij} = 70 + 0.5X_{ij} + u_j + \epsilon_{ij}$$

(β) **Interpretation of results**:

(a) **Fixed Intercept Interpretation** (1 mark)

The fixed intercept ($\beta_0 = 70$) suggests that, on average, the baseline test score for a student with $X = 0$ (e.g., zero study hours) is 70.

(b) **Fixed Slope Interpretation** (1 mark)

The fixed slope ($\beta_1 = 0.5$) indicates that for each additional unit increase in X (each additional hour studied), a student test score is expected to increase by 0.5 points on average.

(c) **Random Intercept Variance Interpretation** (1 mark)

The random intercept variance ($\sigma_u^2 = 15$) suggests that there is variation in test scores across schools, with a standard deviation of $\sqrt{15} \approx 3.87$ points attributable to differences between schools. This implies that the baseline test score differs between schools, capturing variability due to school-specific factors.

(d) **Residual Error Term Interpretation** (1 mark)

The residual error term or variance of $(\sigma_\epsilon^2) = 30$ suggests that there is variation in test scores within schools, with a standard deviation of $\sqrt{30} \approx 5.477$ points attributable to differences within schools. This within-school variance captures unexplained differences in test scores among students within the same school.

(c) **Model Modifications**

(a) **Modified Model Equation** (2 marks)

The modified model is defined as;

$$Y_{ij} = \beta_0 + (\beta_1 + u_{1j})X_{ij} + u_{0j} + \epsilon_{ij}$$

where,

i. u_{0j} is the random intercept for school j

ii. u_{1j} is the random slope for X in school j, both assumed to follow a normal distribution.

(b) **Interpretation of Adding a Random Slope** (2 marks)

By including a random slope u_{1j}, we allow the effect of X (hours studied) on test scores to vary across schools. This accounts for differences in how much study hours impact test scores from one school to another, capturing additional between-school variability in the relationship between X and Y.

CAT II

QUESTION ONE

(a) **Explain the concept of non-parametric regression analysis** **(2 Marks)**

Non-parametric regression analysis is a type of regression analysis that does not assume a specific functional form for the relationship between the dependent and independent variables. Instead, it estimates the relationship directly from the data, allowing for more flexibility to capture complex patterns. It is particularly useful when the underlying relationship is unknown or cannot be adequately described using a parametric model.

(b) **How does non-parametric regression analysis differs from parametric regression analysis?**

(1) For non-parametric regression analysis, no fixed functional form is assumed WHILE a specific functional form such as linear, quadratic, etc is assumed for parametric regression analysis.

(2) Non-parametric regression analysis is Flexible and adapts to the data structure WHILE parametric regression analysis is less flexible and constrained by a specified model.

(3) Non-parametric regression analysis is more computationally intensive WHILE parametric regression analysis is relatively simpler to compute

(4) Non-parametric regression has higher risk of overfitting with small data WHILE parametric regression analysis has lower risk if the model is correctly specified.

(5) Non-parametric regression analysis is used for complex and unknown relationships WHILE parametric regression analysis is used for relationships that are aligned with assumed functional form.

(Any 1 difference for 1 Mark)

(c) **Mention the type of models commonly used in non-parametric regression**

(a) Kernel Regression

(b) Local Polynomial Regression

(c) Spline Regression

(d) Nearest Neighbor Regression

(e) Generalized Additive Models (GAMs)

(Any 2 for 2 Marks)

(d) **Briefly explain the concept of Kernel smoothing in the context of non-parametric regression analysis** **(2 Marks)**

Kernel smoothing is a technique used in non-parametric regression to estimate the regression function at a specific point by averaging nearby observations. It assigns weights to the observations based on their distance from the target point, with closer observations receiving higher weights. The weights are determined using a kernel function, and the degree of smoothing is controlled by a parameter called *bandwidth.*

(e) **Mention three Kernel functions used in estimation in non-parametric regression analysis and state their formulas**

The kernel function, $K(u)$ is usually a symmetric, non-negative function that integrates to 1. Commonly used kernel functions include;

(1) Gaussian Kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

(1 Mark)

(2) Epanechnikov Kernel:

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2), & \text{for } |u| \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

(1 Mark)

(3) Uniform Kernel:

$$K(u) = \begin{cases} \frac{1}{2} & \text{for } |u| \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

(1 Mark)

(f) **Define the Nadaraya-Watson estimator used in Kernel regression analysis and clearly define each component in the formula**

By definition, the N-W kernel estimator is given as;

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K_h\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K_h\left(\frac{x-x_i}{h}\right)}$$

(2 Marks)

where;

(1) $K_h(\cdot)$ is the Gaussian kernel function

(2) x is the point at which we estimate the regression function

(3) x_i and y_i are the observed data points

(4) h is the bandwidth

(0.5 mark for each correct description given, making 2 Marks)

(g) **Given the following data points:**

$$x = (2, 4, 6, 8, 10) \textbf{ and } y = (2, 3, 5, 7, 9),$$

use the Nadaraya-Watson kernel estimator you defined in Question (f) above and the Gaussian kernel function to estimate the value of the regression function $f(4)$ by assuming $h = 1$

Using the Nadaraya-Watson kernel estimator and the Gaussian kernel function to estimate the value of the regression function $f(4)$ by assuming $h = 1$, we have;

$$\hat{f}(4) = \frac{\sum_{i=1}^{n} K_h(4 - x_i) y_i}{\sum_{i=1}^{n} K_h(4 - x_i)}$$

where,

(1) $K(\cdot)$ is defined as;

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

(2) $x = 4$ is the point at which we estimate the regression function

(3) x_i and y_i are the observed data points

(4) $h = 1$ is the bandwidth.

(2 Marks)

Now, given,

$$x = (2, 4, 6, 8, 10) \text{ and } y = (2, 3, 5, 7, 9),$$

we are to estimate $f(4)$ at $h = 1$.

(1) **Compute weights, w_i for each data point, x_i**
For each x_i, we calculate the weight, w_i as follows;

i. For $x_1 = 2$, we have;

$$w_1 = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4-2)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp(-2) \approx 0.0540$$

(1 Mark)

ii. For $x_2 = 4$,

$$w_2 = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4-4)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp(0) = \frac{1}{\sqrt{2\pi}} \approx 0.3989$$

(1 Mark)

iii. For $x_3 = 6$,

$$w_3 = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4-6)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp(-2) \approx 0.0540$$

(1 Mark)

iv. For $x_4 = 8$,

$$w_4 = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4-8)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp(-8) \approx 0.0001$$

(1 Mark)

v. For $x_5 = 10$,

$$w_5 = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4-10)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp(-18) \approx 0.000000006$$

(1 Mark)

(2) **Compute Denominator and Numerator**

(i) Compute Denominator
$Denom = w_1 + w_2 + w_3 + w_4 + w_5 + w_6$
$Denom = 0.0540 + 0.3989 + 0.0540 + 0.0001 + 0.00000001$
$Denom = 0.5071$ **(1 Mark)**

(ii) Compute Numerator
$Num = w_1y_1 + w_2y_2 + w_3y_3 + w_4y_4 + w_5y_5$
$Num = (0.0540\times2)+(0.3989\times3)+(0.0540\times5)+(0.0001\times7)+(0.00000001\times9)$
$Num = 0.1080 + 1.1968 + 0.2700 + 0.0009 + 0.0000001$
$Num = 1.5757$

(2 Marks)

(3) **Compute $f(4)$:**
Divide numerator by denominator to get $f(4)$ using the formula;

$$\hat{f}(4) = \frac{1.5757}{0.5071} \approx 3.1075$$

Therefore, the estimate of $f(4)$ is 3.1075. **(2 Marks)**

(h) **Now, assuming $h = 2.5$, use the same information in Question (g) above to estimate the value of the regression function**.
Given $h = 2.5$, and using the same information in (g), we estimate $f(4)$ as follows;

(1) **Compute weights, w_i for each data point, x_i**
For $h = 2.5$, the procedure is the same as in (g), but the kernel weights are adjusted as;

$$w_i = K\left(\frac{x - x_i}{h}\right) = \exp\left(-\frac{(4 - x_i)^2}{2(2.5)^2}\right)$$

We now compute w_i for $h = 2.5$ for each X_i as follows:
For each x_i, we calculate the weight, w_i as;

i. For $x_1 = 2$, we have;

$$w_1 = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(4-2)^2}{2(2.5)^2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{4}{12.5}\right) = \frac{1}{\sqrt{2\pi}}\exp(0.0033) \approx 0.2897$$

(1 Mark)

ii. For $x_2 = 4$,

$$w_2 = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(4-4)^2}{2(2.5)^2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{0}{12.5}\right) = \frac{1}{\sqrt{2\pi}} \approx 0.3989$$

(1 Mark)

iii. For $x_3 = 6$,

$$w_3 = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(4-6)^2}{2(2.5)^2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{4}{12.5}\right) = \frac{1}{\sqrt{2\pi}}\exp(0.0033) \approx 0.2897$$

(1 Mark)

iv. For $x_4 = 8$,

$$w_4 = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4-8)^2}{2(2.5)^2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{16}{12.5}\right) = \frac{1}{\sqrt{2\pi}} \exp(1.28) \approx 0.1109$$

(1 Mark)

v. For $x_5 = 10$,

$$w_5 == \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4-10)^2}{2(2.5)^2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{36}{12.5}\right) = \frac{1}{\sqrt{2\pi}} \exp(2.88) \approx 0.0224$$

(1 Mark)

(2) **Compute Denominator and Numerator**

(i) **Compute Denominator**
$Denom = w_1 + w_2 + w_3 + w_4 + w_5 + w_6$
$Denom = 0.2897 + 0.3989 + 0.2897 + 0.1109 + 0.0224$
$Denom =$1.1116 **(1 Mark)**

(ii) **Compute Numerator**
$Num = w_1 y_1 + w_2 y_2 + w_3 y_3 + w_4 y_4 + w_5 y_5$
$Num = (0.2897 \times 2) + (0.3989 \times 3) + (0.2897 \times 5) + (0.1109 \times 7) + (0.0224 \times 9)$
$Num =$ 0.5794 + 1.1968 + 1.4485+ 0.7764 + 0.2016
$Num =$ 4.2027

(2 Marks)

(3) **Compute $f(4)$:**
Divide numerator by denominator to get $f(4)$ using the formula;

$$\hat{f}(4) = \frac{Num}{Denom} \approx \frac{4.2027}{1.1116} = 3.7806$$

Therefore, the estimate of $f(4)$ is 3.7806. **(2 Marks)**

(i) **By comparing the results of Question (g) and Question (h), explain the effect of the different bandwidth parameters on the smoothness of the non-parametric regression curve**.

(1) When h is smaller ($h = 1$), the kernel weights $K(u_i)$ are sharper, giving more weight to points closer to $x_0 = 4$. This results in a less smooth regression curve, as it captures more local variations. So, a smaller bandwidth results in a less smooth curve, as it focuses more on the local structure and may overfit the data.

(2) When h is larger ($h = 2.5$), the kernel weights $K(u_i)$ are smoother, spreading the influence over a broader range of data points. This results in a smoother regression curve, as it averages out variations over a wider range. So, a larger bandwidth produces a smoother curve, as it incorporates more points in the averaging process, potentially underfitting the data.

(2 Marks)

(j) **Briefly explain how bandwidth parameters are selected in the context of non-parametric regression analysis?**
Bandwidth selection is crucial for balancing bias and variance. Common methods include:

(1) **Cross-Validation:** Selecting h by minimizing the prediction error on a validation dataset. The cross-validation is given as;

$$\hat{h}_{CV} = \arg\min_{h} \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - \hat{f}_{-i}(X_i) \Big)^2$$

Where, $\hat{f}_{-i}(X_i)$ is the leave-one-out kernel regression estimator, computed without using the i^{th} observation.

(2) **Rule-of-Thumb:** Using a formula based on the data's standard deviation and sample size:

$$h = 1.06\sigma n^{-1/5}$$

where σ is the standard deviation and n is the number of data points.

(3) **Plug-In Methods:** Estimating the bandwidth by approximating the optimal smoothing parameter.

(Any 1 method for 2 Marks)

QUESTION TWO

(a) **Define a simple neural network for regression and explain how it differs from traditional linear regression models (3 Marks)**
A simple neural network for regression consists of layers of interconnected neurons that learn to approximate a target function by adjusting weights and biases through optimization algorithms. Unlike traditional linear regression models, which assume a linear relationship between input and output, neural networks can model complex, non-linear relationships using activation functions and multiple layers.

(b) **Explain the architecture of a simple neural network used for regression tasks. Highlight the roles of the input layer, hidden layers, and output layer (4 Marks)**

(1) Input Layer: This takes in the features or independent variables (x_1, x_2, . . . , x_n).

(2) Hidden Layers: These are intermediate layers that process input using weights, biases, and activation functions to learn complex patterns.

(3) Output Layer: This produces the predicted output (y) based on hidden layers' output. For regression, this layer usually has no activation function.

(c) **Briefly define an activation function and give 3 examples**.
An activation function introduces non-linearity to the network, allowing it to model complex patterns **(1 Mark)**
Examples includes;

(1) Rectified Linear Units (ReLU) Activation Function:

$$\text{ReLU}(z) = \max(0, z)$$

(2) Sigmoid Activation Function:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

(3) Tanh Activation Function:

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

(4) Leaky ReLU Activation Function:

$$\text{Leaky ReLU}(z) = \begin{cases} z & \text{if } z > 0, \\ \alpha z & \text{if } z \leq 0 \end{cases}$$

where α is a small constant (e.g., $\alpha = 0.01$).

(5) Softmax Activation Function:

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

where, z_i is the input to the i^{th} neuron and n is the total number of neurons.

(6) Exponential Linear Unit (ELU) Activation Function:

$$ELU(z) = \begin{cases} z & \text{if } z > 0, \\ \alpha(e^z - 1) & \text{if } z \leq 0 \end{cases}$$

where α is a positive constant.

(7) Swish Activation Function:

$$Swish(z) = z \cdot \sigma(z) = z \cdot \frac{1}{1+e^{-z}}$$

(8) Maxout Activation Function:

$$Maxout(z) = \max(z_1, z_2, \ldots, z_k)$$

where $z_1, z_2, \ldots, z_k$ are the inputs.

(Any 3 for 3 Marks)

(d) **Explain the role of activation functions in neural network used for regression (1 Mark)**
Activation functions enable the network to capture non-linear relationships between inputs and outputs. Without them, the neural network behaves like a linear model, limiting its ability to solve complex problems.

(e) **You are given data with an input feature, x and an output feature, y. A simple feedforward neural network has the following:**
The network has the following details from the question:

(1) Input data: $x = (2, 3)$, where $x_1 = 2$, $x_2 = 3$

(2) Hidden layer weights and biases:

(i) Neuron 1: $w_{1,1} = 0.5$, $w_{1,2} = 0.1$, $b_1 = 0.1$

(ii) Neuron 2: $w_{2,1} = -0.3$, $w_{2,2} = 0.2$, $b_2 = -0.2$

(3) Output layer weights and bias:

(1) $w_{3,1} = 0.8$, $w_{3,2} = 0.2$, $b_3 = -0.3$

(α) Calculate the input to each neuron in the hidden layer.
For neuron i in the hidden layer, we have;

$$z_i = w_{i,1}x_1 + w_{i,2}x_2 + b_i$$

So, for neuron 1, we have;

$$z_1 = (0.5)(2) + (0.1)(3) + 0.1 = 1 + 0.3 + 0.1 = 1.4$$

(1 Mark)

Similarly, for neuron 2, we have;

$$z_2 = (-0.3)(2) + (0.2)(3) - 0.2 = -0.6 + 0.6 - 0.2 = -0.2$$

(1 Mark)

(β) Apply the ReLU to each hidden input of the neuron.
The ReLU activation function is defined as;

$$f(z) = \max(0, z)$$

So, for neuron 1, we have;

$$f(z_1) = \max(0, 1.4) = 1.4$$

(1 Mark)

Similarly, for neuron 2, we have;

$$f(z_2) = \max(0, -0.2) = 0$$

(1 Mark)

(γ) Calculate the final output of the network for the given input.
For the output layer, we have;

$$\hat{y} = w_{3,1}f(z_1) + w_{3,2}f(z_2) + b_3$$

So, this gives;

$$\hat{y} = (0.8)(1.4) + (0.2)(0) - 0.3 = 1.12 + 0 - 0.3 = 0.82$$

(1 Mark)

So, the final output gives;

$$\hat{y} = 0.82$$

(1 Mark)

(f) **What is a loss function? Give 2 examples of loss functions**

A loss function measures the difference between predicted values ($\hat{y}$) and actual values (y). It quantifies the error the model needs to minimize during training. **(1 Mark)**

Examples of loss functions are;

(1) Mean Squared Error (MSE) Loss Function

(2) Mean Absolute Error (MAE) Loss Function

(3) Binary Cross-Entropy Loss Function

(4) Categorical Cross-Entropy Loss Function

(5) Huber Loss Function

(6) Hinge Loss Function

(7) Kullback-Leibler (KL) Divergence Loss Function

(Any 2 for 2 Marks)

(g) **Define the Mean squared Error (MSE) loss function and explain how it is used to train a neural network for regression. Hence, estimate the MSE loss given the following data points:**

$$y = (5, 10, 15, 20, 25), \text{ and } \hat{y} = (4, 8, 18, 15, 20)$$

A loss function measures the difference between predicted values ($\hat{y}$) and actual values (y). **(2 Marks)**

The MSE loss function is defined as;

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

For $y = (5, 10, 15, 20, 25)$ and $\hat{y} = (4, 8, 18, 15, 20)$, we have;

$$MSE = \frac{1}{5}\left[(5-4)^2 + (10-8)^2 + (15-18)^2 + (20-15)^2 + (25-20)^2\right]$$

$$MSE = \frac{1}{5}\left[1 + 4 + 9 + 25 + 25\right] = \frac{64}{5} = 12.8$$

Therefore, the MSE is 12.8 **(2 Marks)**

(h) **Using the same data points in Question (g), estimate the Mean Absolute Error loss (MAE)**.

The MAE loss function is defined as;

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

For the same data in (g), we have;

$$MAE = \frac{1}{5}\left[|5-4| + |10-8| + |15-18| + |20-15| + |25-20|\right]$$

$$MAE = \frac{1}{5}\left[1 + 2 + 3 + 5 + 5\right] = \frac{16}{5} = 3.2$$

Therefore, the MAE is 3.2 **(2 Marks)**

(i) **Compare the results of Question (g) and Question (h) above** **(1 Mark)**

(1) MSE is more sensitive to large errors because it squares the differences. This makes it suitable for cases where large deviations are particularly undesirable.

(2) MAE provides a linear measure of error and is less sensitive to outliers.

(3) For this dataset, the higher MSE indicates that large deviations significantly impact the error.

(j) **Two state optimization algorithms that are used to adjust weights and biases to minimize losses**

(1) Stochastic Gradient Descent (SGD) Optimizer: Iteratively adjusts weights and biases by moving in the direction of the negative gradient.

(2) Adam Optimizer: Combines momentum and adaptive learning rates for efficient weight updates.

(3) Deterministic Finite Gradient Search (DFGS) Optimizer: Designed to iteratively find the minimum of a differentiable objective function. It is often applied in machine learning and numerical optimization tasks, where smooth, convex (or non-convex) loss functions are minimized.

(Any 2 for 2 Marks)