IME692A Assignment 2

Rajarshi Dutta
200762

Saubhagya Bawari
200889

1 Question 1

This question discusses about a Multiple Linear Regression Model used for the **cadata.txt** dataset which consits of features like **Median House Value**, **Median Income**, **Housing Median Age**, **Total Rooms**, **Population**, **Total Bedrooms**, **Lattitude**, **Longitude**, **Households**. The equation describing the MLR model is represented below:

$$\begin{aligned}\ln(MHV) =&\beta_0 + \beta_1(MI) + \beta_2(MI^2) + \beta_3(MI^3) + \beta_4\ln(MA)\\ &+ \beta_5\ln\left(\frac{TR}{P}\right) + \beta_6\ln\left(\frac{B}{P}\right) + \beta_7\ln\left(\frac{P}{H}\right) + \beta_8\ln(H) + e\end{aligned}$$

The beta values are mentioned in the questions and the errors are calculated for the given MLR model to the data. These errors are further tested for conditions like **zero expectation** since the independent variables should have no information regarding the expected value of the errors. The errors follow a **Gaussian distribution** with a zero mean and the errors are also independent of each other.

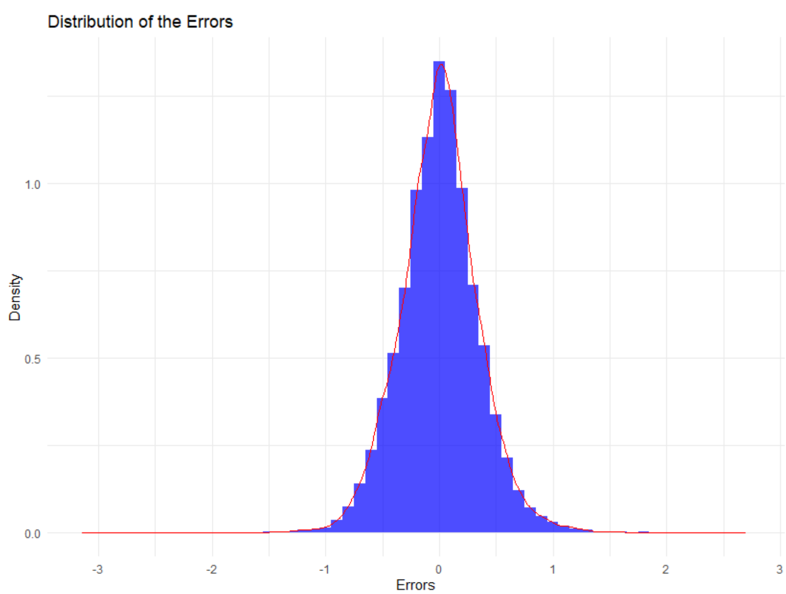

Figure 1: Distribution of Error values

Then we calculate the **Ordinary Least Squares Estimator** $\hat{\beta} = (X^TX)^{-1}X^Ty$. As for the second portion of the question, we are required to choose **100** random samples with a **sample size** of **200**. We are then required to calculate the expected value of $\hat{\beta}$ which is $E[\hat{\beta}_i]\forall\beta_i, i = 0...8$. We then

β **Index**	**Actual Value** β_i	**Expected value** $E[\hat{\beta}_i]$	**Density for** $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2(X^TX)^{-1})$
0	11.4939	11.5111	0.5626
1	0.4790	0.5270	0.7425
2	-0.0166	-0.0173	0.8167
3	-0.0002	-0.0350	0.7794
4	0.1570	0.1306	0.7753
5	-0.8582	-0.8616	0.7756
6	0.8043	0.8203	0.7716
7	-0.4077	-0.4231	0.7746
8	0.0477	0.0440	0.8074

Table 1: Expected Values and pdfs for $\hat{\beta}_i$

calculate the probability density values for each of the $\hat{\beta}_i$ corresponding to the distribution $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2(X^TX)^{-1})$. The given table shows the expected values and the probability densities for β_i's.

The probability distributions are also represented in the following bar plot:

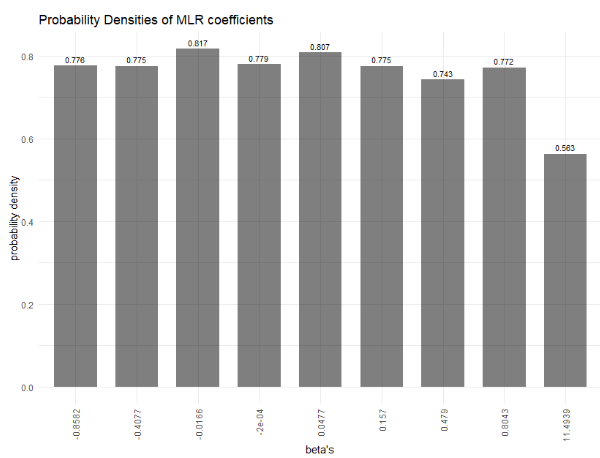

Figure 2: Caption

The second part involves the calculations of **90%**, **95%** and **99%** confidence intervals for cases corresponding to **known variance (Z distribution)** and **unknown variance (t distribution)**. The plots for the confidence intervals for each of the $\hat{\beta}_i$'s are represented below:

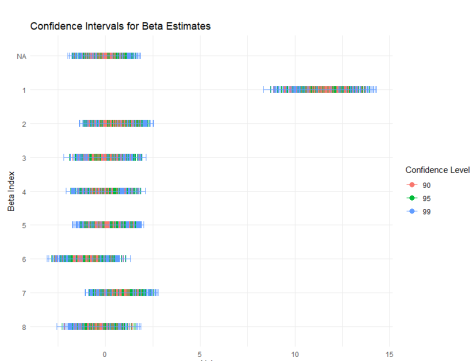

Figure 3: T distribution Confidence Intervals

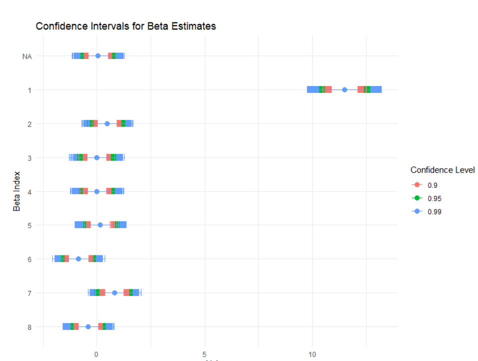

Figure 4: Z distribution Confidence Intervals

The data samples from **20600** to **20640** are calculated using the estimated values of $\hat{\beta}$ and compared to the original one. The total deviation is around **55.0203** . The table for true values vs estimated values is given below:

Index	Value	Index	Value
20600	-2.43861331	20621	1.33648232
20601	-0.55377642	20622	-1.18527660
20602	-1.80225954	20623	-1.02931905
20603	-2.51177635	20624	-0.84893814
20604	-2.89936993	20625	-0.48928029
20605	-1.88715946	20626	1.71713742
20606	-2.35839122	20627	-1.45428905
20607	-2.02958382	20628	0.43717245
20608	-1.98800848	20629	-1.20333717
20609	-1.94451654	20630	-2.57668380
20610	-1.61251297	20631	-0.08315018
20611	-2.63626494	20632	-0.19283215
20612	-2.68061651	20633	-0.54832525
20613	-2.27428666	20634	-1.34027409
20614	-2.50547289	20635	-0.06188863
20615	-1.54043292	20636	-2.28538561
20616	-1.23961149	20637	-0.68049152
20617	-1.74645184	20638	-2.15750645
20618	0.08541121	20639	-1.89336037
20619	-1.41760042	20640	-1.52173984
20620	-0.97845290		

Table 2: Data Values

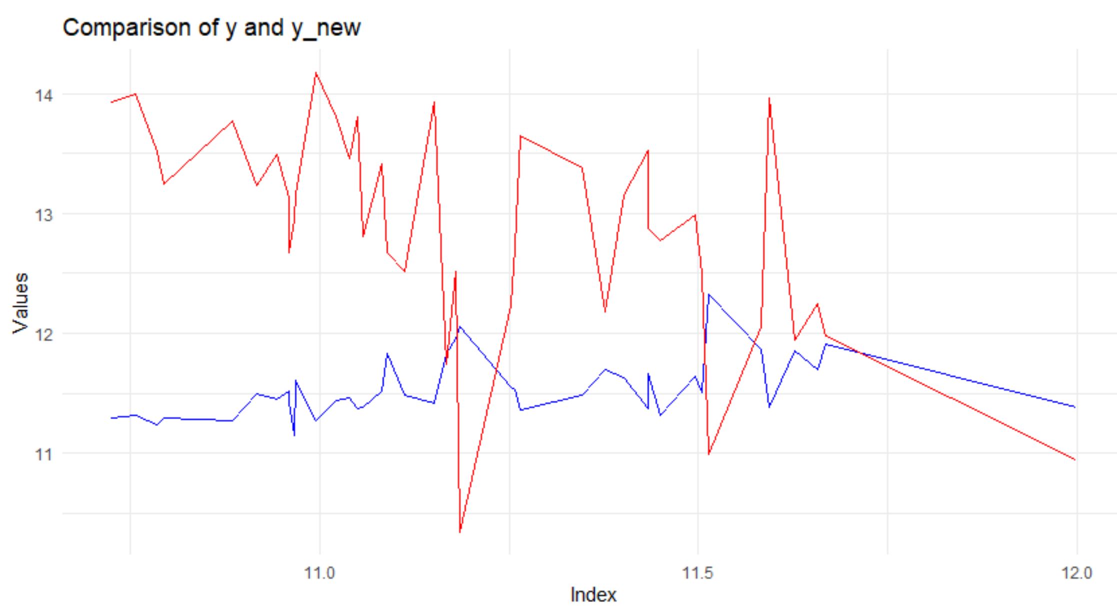

Figure 5: Variation of y and y_estimated

As for the **b** part, the tables for the loss values across different models:

- for **Model 1**, the value of $\hat{\beta}$ is $[11.39946, 0.28464, -0.36161, 0.04353, 0.4154, -0.9716, 0.78819, -0.42018, 0.4959]^T$ and the value of $\hat{\theta}$ is **11.6723**.

- for **Model 2**, the value of $\hat{\beta}$ is $[11.39946, 0.28464, -0.36161, 0.04353, 0.4154, -0.9716, 0.78819, -0.42018, 0.4959]^T$ and the value of $\hat{\theta}$ is **11.6723**.
- for **Model 3**, the value of $\hat{\beta}$ is $[11.39946, 0.28464, -0.36161, 0.04353, 0.4154, -0.9716, 0.78819, -0.42018, 0.4959]^T$ and the value of $\hat{\theta}$ is **11.67422**.
- for **Model 4**, the value of $\hat{\beta}$ is $[11.39946, 0.28464, -0.36161, 0.04353, 0.4154, -0.9716, 0.78819, -0.42018, 0.4959]^T$ and the value of $\hat{\theta}$ is **11.67422**.
- for **Model 5**, the value of $\hat{\beta}$ is $[11.39946, 0.28464, -0.36161, 0.04353, 0.4154, -0.9716, 0.78819, -0.42018, 0.4959]^T$ and the value of $\hat{\theta}$ is **11.67423**.
- for **Model 6**, the value of $\hat{\beta}$ is $[11.39946, 0.28464, -0.36161, 0.04353, 0.4154, -0.9716, 0.78819, -0.42018, 0.4959]^T$ and the value of $\hat{\theta}$ is **11.67423**.
- for **Model 7**, the value of $\hat{\beta}$ is $[11.39946, 0.28464, -0.36161, 0.04353, 0.4154, -0.9716, 0.78819, -0.42018, 0.4959]^T$ and the value of $\hat{\theta}$ is **11.67422**.
- for **Model 8**, the value of $\hat{\beta}$ is $[11.39946, 0.28464, -0.36161, 0.04353, 0.4154, -0.9716, 0.78819, -0.42018, 0.4959]^T$ and the value of $\hat{\theta}$ is **11.67423**.

Model no.	λ **1**	λ **2**	λ **3**	λ **4**
1	23.3076	47.5032	71.6987	95.8943
2	-0.7091	-0.5303	-0.3516	-0.1729
3	23.2900	47.5801	71.8702	96.1603
4	-0.7997	-0.5994	-0.3992	-0.1989
5	-0.3949	-0.2958	-0.1968	-0.0978
6	-0.3997	-0.2994	-0.1992	-0.0989
7	0.0048	0.0036	0.0024	0.0012
8	0.00026	0.00005	0.0007	0.0010

Table 3: Risk Values for $\hat{\lambda}_i$

Model no.	λ **1**	λ **2**	λ **3**	λ **4**
1	0.82236	1.64473	2.46709	3.28940
2	0.0896	0.1793	0.2690	0.3587
3	0.8223	1.6447	2.4670	3.2894
4	0.0896	0.1793	0.2690	0.3586
5	0.00081	0.00076	0.00071	0.00067
6	0.00013	0.00025	0.00037	0.00050
7	0.00081	0.00076	0.00071	0.00067
8	0.00013	0.00025	0.00034	0.00050

Table 4: Loss values for $\hat{\lambda}_i$

2 Question 2

2.1 Chi Square Distribution

For the first part of the problem, we are asked to find the **probability distribution function** (pdf) of X_{min} and X_{max} of a set of random variables X_i for all i up to n. Here $X_{\text{min}} = \min(X_1, X_2, X_3, \dots, X_n)$ and $X_{\text{max}} = \max(X_1, X_2, X_3, \dots, X_n)$ where the random variables follow a **Chi-square distribution** with pdf represented as

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$$

The calculations for the $f_{X_{\text{min}}}(x)$ and $f_{X_{\text{max}}}(x)$ are presented as follows:

$$F_{X_{\text{min}}}(x) = 1 - P(X_{\text{min}} > x) \tag{1}$$
$$F_{X_{\text{min}}}(x) = 1 - P(\min(X_1, X_2, X_3, \dots, X_n) > x) \tag{2}$$
$$F_{X_{\text{min}}}(x) = 1 - P(X_1 > x)P(X_2 > x)P(X_3 > x)\dots P(X_n > x) \tag{3}$$
$$F_{X_{\text{min}}}(x) = 1 - \prod_{i=1}^{n} P(X_i > x) \tag{4}$$
$$F_{X_{\text{min}}}(x) = 1 - \prod_{i=1}^{n}(1 - F(x)) \quad \text{where } F(x) = P(X \leq x) \tag{5}$$

Now, for the calculation of the **Cumulative probability distribution** $F(x)$, we have:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \tag{6}$$
$$F(x) = \int_{-\infty}^{0} f(t)\, dt + \int_{0}^{x} f(t)\, dt \tag{7}$$
$$F(x) = \int_{0}^{x} \frac{1}{2^{n/2}\Gamma(n/2)} t^{n/2-1} e^{-t/2}\, dt \tag{8}$$

Performing a variable substitution of $p = t/2$, we get the following integral which can be expressed in terms of the **lower incomplete gamma function** where $\gamma(n, x) = \int_0^x x^{n-1} e^{-x}\, dt$.

$$F(x) = \int_{0}^{x/2} \frac{1}{2^{n/2}\Gamma(n/2)} (2p)^{n/2-1} e^{-p}\, 2dp \tag{9}$$
$$= \int_{0}^{x/2} \frac{1}{2^{n/2}\Gamma(n/2)} 2^{n/2} p^{n/2-1} e^{-p}\, 2dp \tag{10}$$
$$= \frac{\int_0^{x/2} e^{-p} p^{n/2-1}\, dp}{\Gamma(n/2)} \tag{11}$$
$$= \frac{\gamma(n/2, x/2)}{\Gamma(n/2)} \tag{12}$$
$$F_{X_{\text{min}}} = 1 - \left(1 - \frac{\gamma(n/2, x/2)}{\Gamma(n/2)}\right)^{n-1} \tag{13}$$
$$f_{X_{\text{min}}}(x) = \frac{dF_{X_{\text{min}}}}{dx} \tag{14}$$
$$f_{X_{\text{min}}}(x) = \frac{n x^{n/2-1} e^{-x/2}}{2^{n/2}\Gamma(n/2)} \left(1 - \left(1 - \frac{\gamma(n/2, x/2)}{\Gamma(n/2)}\right)^{n-1}\right) \tag{15}$$

The steps for the calculation of $F_{X_{\text{max}}}(x)$ are mentioned as follows:

$$F_{X_{\text{max}}}(x) = P(X_{\text{max}} \leq x) \tag{16}$$
$$F_{X_{\text{max}}}(x) = P(\max(X_1, X_2, X_3, \dots, X_n) \leq x) \tag{17}$$
$$F_{X_{\text{max}}}(x) = P(X_1 \leq x)P(X_2 \leq x)P(X_3 \leq x)\dots P(X_n \leq x) \tag{18}$$
$$F_{X_{\text{max}}}(x) = \prod_{i=1}^{n} P(X_i \leq x) \tag{19}$$
$$F_{X_{\text{max}}}(x) = \prod_{i=1}^{n} F(x) \quad \text{where } F(x) = P(X \leq x) \tag{20}$$

Performing a similar variable substitution of $p = t/2$ for the calculation of $F_{X_{\max}}(x)$ we get the following integral which can be expressed in terms of the **lower incomplete gamma function** where $\gamma(n,x) = \int_0^x x^{n-1}e^{-x}\,dt$.

$$F(x) = \int_0^{x/2} \frac{1}{2^{n/2}\Gamma(n/2)}(2p)^{n/2-1}e^{-p}\,2dp \tag{21}$$

$$= \int_0^{x/2} \frac{1}{2^{n/2}\Gamma(n/2)}2^{n/2}p^{n/2-1}e^{-p}\,2dp \tag{22}$$

$$= \frac{\int_0^{x/2} e^{-p}p^{n/2-1}\,dp}{\Gamma(n/2)} \tag{23}$$

$$= \frac{\gamma(n/2, x/2)}{\Gamma(n/2)} \tag{24}$$

$$F_{X_{\max}} = \left(\frac{\gamma(n/2, x/2)}{\Gamma(n/2)}\right)^{n-1} \tag{25}$$

$$f_{X_{\max}}(x) = \frac{dF_{X_{\max}}}{dx} \tag{26}$$

$$f_{X_{\max}}(x) = \frac{nx^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}\left(\frac{\gamma(n/2, x/2)}{\Gamma(n/2)}\right)^{n-1} \tag{27}$$

2.2 t distribution

For the first part of the problem, we are asked to find the **probability distribution function** (pdf) of X_{min} and X_{max} of a set of random variables X_i for all i up to n. Here $X_{\text{min}} = \min(X_1, X_2, X_3, \ldots, X_n)$ and $X_{\text{max}} = \max(X_1, X_2, X_3, \ldots, X_n)$ where the random variables follow a **t distribution** with pdf represented as

$$f(x) = \frac{\Gamma\frac{(n+1)}{2}}{\sqrt{\pi n}\Gamma(n/2)}\left(1 + \frac{x^2}{n}\right)^{-\frac{(n+1)}{2}}$$

Similar to the previous part, Now, for the calculation of the **Cumulative probability distribution** $F(x)$, we have:

$$F(x) = \int_{-\infty}^{x} f(u)\,du \tag{28}$$

$$F(x) = \int_{-\infty}^{0} f(u)\,du + \int_0^x f(u)\,dt \tag{29}$$

$$F(x) = \frac{1}{2} + \left(\frac{1}{2} - \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\int_x^{\infty}\left(1 + \frac{u^2}{n}\right)^{-\frac{n+1}{2}}\,dt\right) \tag{30}$$

$$F(x) = 1 - I_{x(u)}\left(\frac{n}{2}, \frac{1}{2}\right) \quad \text{where } x(u) = \frac{n}{u^2 + n} \tag{31}$$

$$\tag{32}$$

Here the term $I_{t(u)}(t)(n/2, 1/2)$ refers to the **incomplete beta function**. The final steps for calculations of $f_{X_{\min}}(x)$ and $f_{X_{\max}}(x)$ are:

$$F_{X_{\max}} = \left(1 - I_{x(u)}\left(\frac{n}{2}, \frac{1}{2}\right)\right)^{n} \tag{33}$$

$$f_{X_{\max}}(x) = n\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}\left(1 - I_{x(u)}\left(\frac{n}{2}, \frac{1}{2}\right)\right)^{n-1} \tag{34}$$

$$F_{X_{\min}} = \left(I_{x(u)}\left(\frac{n}{2}, \frac{1}{2}\right)\right)^{n} \tag{35}$$

$$f_{X_{\max}}(x) = n\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}\left(I_{x(u)}\left(\frac{n}{2}, \frac{1}{2}\right)\right)^{n-1} \tag{36}$$