Name: Aditya Sagave

Student no: 47541164

GLM STAT8111 - Assignment 1

Q1

```
data(cystfibr)
head(cystfibr)
```

```
##   age sex height weight bmp fev1  rv frc tlc pemax
## 1   7   0    109   13.1  68   32 258 183 137    95
## 2   7   1    112   12.9  65   19 449 245 134    85
## 3   8   0    124   14.1  64   22 441 268 147   100
## 4   8   1    125   16.2  67   41 234 146 124    85
## 5   8   0    127   21.5  93   52 202 131 104    95
## 6   9   0    130   17.5  68   44 308 155 118    80
```

```
# Select the variables of interest
variables_of_interest <- cystfibr[c("weight", "sex", "bmp", "fev1", "rv", "age", "frc")]

# Calculate correlation matrix
correlation_matrix <- cor(variables_of_interest)
correlation_matrix
```

```
##            weight        sex        bmp       fev1         rv        age
## weight  1.0000000 -0.1904400  0.6725463  0.4488393 -0.6215056  0.9058675
## sex    -0.1904400  1.0000000 -0.1375611 -0.5282571  0.2713516 -0.1671220
## bmp     0.6725463 -0.1375611  1.0000000  0.5455204 -0.5823729  0.3777643
## fev1    0.4488393 -0.5282571  0.5455204  1.0000000 -0.6658557  0.2944880
## rv     -0.6215056  0.2713516 -0.5823729 -0.6658557  1.0000000 -0.5519445
## age     0.9058675 -0.1671220  0.3777643  0.2944880 -0.5519445  1.0000000
## frc    -0.6172561  0.1836055 -0.4343888 -0.6651149  0.9106029 -0.6393569
##               frc
## weight -0.6172561
## sex     0.1836055
## bmp    -0.4343888
## fev1   -0.6651149
## rv      0.9106029
## age    -0.6393569
## frc     1.0000000
```

```
# Create scatter plot matrix
#ggpairs(variables_of_interest)
```

Answer 1.a:

Graphical Analysis:

****Positive correlations:**

- Weight and fev1, Weight and age, rv and frc

Negative correlation:

- Weight and rv, weight and frc, fev1 and rv, fev1 and frc, rv and age, age and frc

Answer 1.b: #### **Numerical Analysis:**

- Weight and FEV1:positive correlation (0.45); Age and RV: Moderate negative correlation (-0.55); FEV1 and RV: Strong negative correlation (-0.67); Weight and Age: Very strong positive correlation (0.91); RV and FRC: Strong positive correlation (0.91); residual volume (RV) and functional residual capacity (FRC) are closely related. FEV1 and FRC: Strong negative correlation (-0.67)

Linear Model for Relationship between Weight and Pemax:

The linear model to study the relationship between the response variable Pemax (maximum expiratory pressure) and the explanatory variable Weight can be expressed as:

$$y(pemax) = \beta_0 + \beta_1 x(weight) + \epsilon$$

$Where$,

Pemax is the response variable (maximum expiratory pressure).

x(Weight) is the co-variate.

β_0 is the intercept

β_1 is the Slopw OR coefficient for Weight, representing the change in Pemax for a unit change in weight.

ϵ is the error term, representing the variability not explained by the model.

Assumptions of the Linear Model:

- **Linearity**: The relationship between Weight and Pemax is assumed to be linear.
- **Independence**: The errors (ϵ) are assumed to be independent for each observation.
- **Multicollinearity**: There is no perfect multicollinearity among the predictor variables.
- **Normality of Errors**: The errors are assumed to be normally distributed.

Model Fitting and Interpretation: The linear model was fitted to the data to quantify the relationship between Pemax and Weight. The coefficients were estimated, with the results indicating that the intercept (β_0) represents the estimated Pemax when Weight is zero, which might not be meaningful in this context. The coefficient for Weight (β_1) represents the change in Pemax for a unit change in weight.

```
model1 <- lm(pemax ~ weight, data = cystfibr)
summary(model1)
```

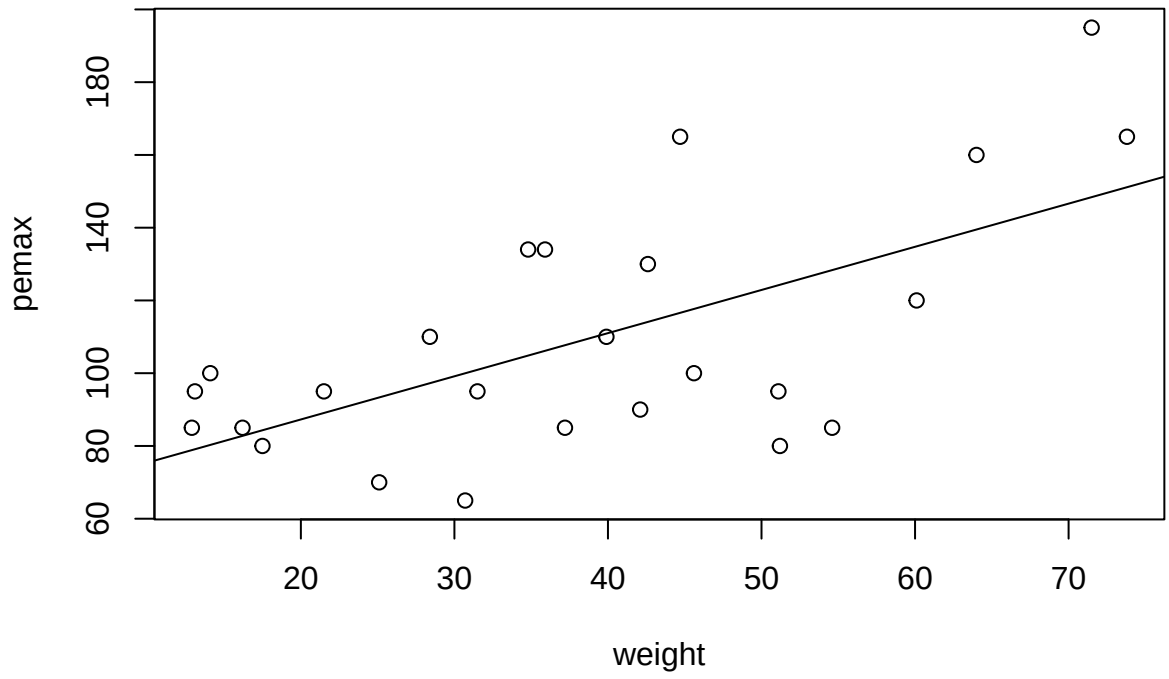
```
##
## Call:
## lm(formula = pemax ~ weight, data = cystfibr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.30 -22.69   2.23  15.91  48.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.5456    12.7016   5.003 4.63e-05 ***
## weight        1.1867     0.3009   3.944 0.000646 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.38 on 23 degrees of freedom
## Multiple R-squared:  0.4035, Adjusted R-squared:  0.3776
```

```
## F-statistic: 15.56 on 1 and 23 DF,  p-value: 0.0006457
```

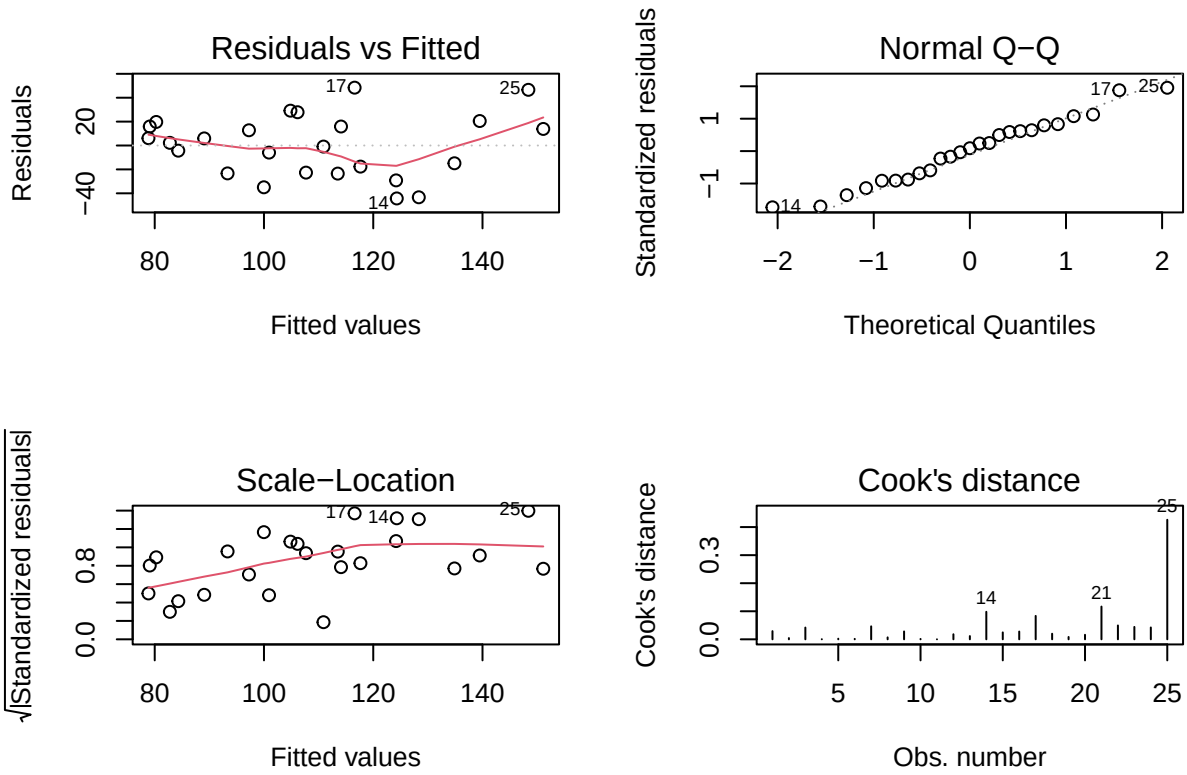
$$pemax = 63.5456 + 1.1867 * weight$$

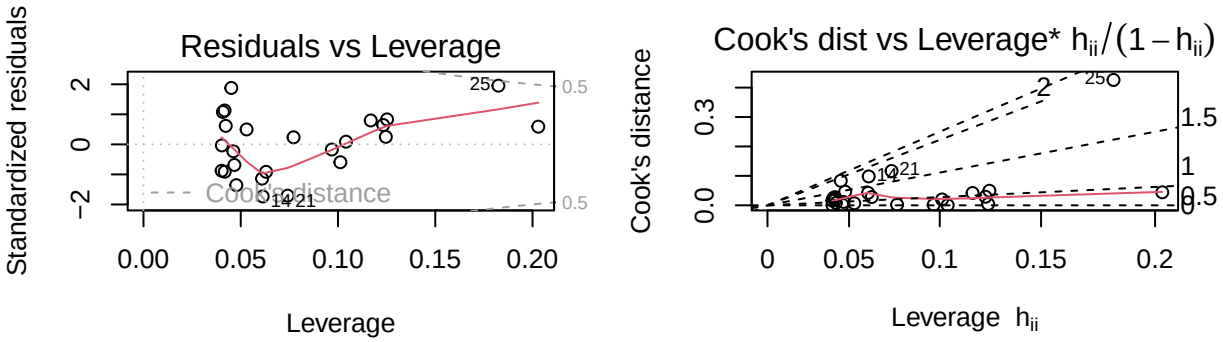
- $\beta_0 = 63.5456$
- $\beta_1 = 1.1867$
- Both the intercept and weight coefficient are statistically significant ($p < 0.001$), indicating that they have a significant impact on the model.
- The multiple R-squared value (0.4035) indicates that approximately 40.35% of the variability in Pemax can be explained by the linear relationship with weight.
- The adjusted R-squared (0.3776) considers the number of predictors and provides a more realistic estimate of the model's explanatory power.

```
plot(pemax ~ weight, data = cystfibr)
abline(model1)
```




```
par(mfrow = c(2, 2))
plot(model1, which = 1:6)
```


Residuals vs Fitted
Residuals
Fitted values
17
25
14
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Cook's distance
Obs. number
21



Diagnostic plots interpretations:

- **Residuals vs Fitted Plot**: The Residuals are randomly scattered along the zero ($y = 0$) line without any clear funnel-like shape or patterns.
- **Normal Q-Q Plot**: Residuals follow a straight line, indicating normality, except 3 outliers (14, 17, 25).
- **Residuals vs Leverage Plot**: Except a residual of 25 is close to the 0.5 dashed line, no points exceeding it indicate that there are no influential observations significantly impacting the model.
- **Cook's Distance Plot**: None of the points surpass the Cook's threshold of 1, implying that there are no influential observations significantly affecting the model.

Answer 1.c:

Model 1 - Including Sex as a Categorical Variable:

$$pemax = \beta_0 + \beta_1 * weight + \beta_2 * sex + \epsilon$$

- The coefficient β_2 represents the change in pemax when sex changes from male (0) to female (1), while keeping weight constant.

Model 2 - Interaction Model with Sex:

$$pemax = \beta_0 + \beta_1 * weight + \beta_2 * sex + \beta_3 * (weight * sex) + \epsilon$$

- The interaction term β_3 represents how the relationship between Weight and pemax changes depending on the sex. It indicates whether the effect of weight on pemax differs between males and females.

```
model2 <- lm(pemax ~ weight + sex, data = cystfibr)
model3 <- lm(pemax ~ weight * sex, data = cystfibr)
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = pemax ~ weight + sex, data = cystfibr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.388 -16.850   0.073  13.168  43.748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.9719    14.4644   4.907 6.61e-05 ***
## weight        1.1248     0.3056   3.681  0.00131 **
## sex         -11.4776    10.7963  -1.063  0.29926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.31 on 22 degrees of freedom
## Multiple R-squared:  0.4327, Adjusted R-squared:  0.3811
## F-statistic: 8.388 on 2 and 22 DF,  p-value: 0.00196
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = pemax ~ weight * sex, data = cystfibr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.464 -14.565  -2.096  14.247  42.973
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  61.3603    15.9335   3.851 0.000927 ***
## weight        1.3572     0.3471   3.910 0.000805 ***
## sex          22.0905    27.2923   0.809 0.427358
## weight:sex   -0.9240     0.6922  -1.335 0.196187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.85 on 21 degrees of freedom
## Multiple R-squared:  0.477,  Adjusted R-squared:  0.4023
## F-statistic: 6.385 on 3 and 21 DF,  p-value: 0.003025
```

Analysis of models:

The adjusted R-squared for **model1** is 0.4035, suggesting that around **40.35%** of the variability in Pemax is explained by the combinations of co-variates in the model and the p-value is significant (p = 0.0006457), indicating that the model is statistically significant.The adjusted R-squared for **model2** is 0.4327, suggesting that around **43.27%** of the variability in Pemax is explained by the combinations of co-variates in the model and the p-value is significant (p = 0.00196), indicating that the model is statistically significant. The adjusted R-squared for **model3** is 0.477, suggesting that around **47.7%** of the variability in Pemax is explained by the combinations of co-variates in the model and the p-value is significant (p = 0.003025), indicating that the model is statistically significant.

Choice and Justification:

Amongst Model 1, Model 2, and Model 3, Model 3 stands out with the highest Adjusted R-squared value, which is 0.4023. This suggests that Model 3 explains approximately 40.23% of the variability in Pemax using the combination of weight, sex, and their interaction. While Model 2 and Model 3 both include the sex predictor, Model 3's higher Adjusted R-squared indicates a potentially better fit.

Answer 1.d:

In constructing the statistical model for Pemax based on the normal response distribution and the variables weight, bmp, fev1, rv, and frc, a stepwise approach can be used to select the most relevant predictors and build an appropriate model. Starting with a full model including all variables, a stepwise process involves iteratively adding and removing predictors based on statistical significance and model fit improvement.

Diagnostic: For Diagnostic we assume the same factore that we assumed for the linear model earlier. These are Linearity, Independence, Multicollinearity & Normality of Errors

Final model equation:

$$pemax = \beta_0 + \beta_1 * weight + \beta_2 * bmp + \beta_3 * fev1 + \beta_4 * rv + \beta_5 * frc + \epsilon$$

```
model4 <- lm(pemax ~ weight + bmp + fev1 + rv + frc, data = cystfibr)
summary(model4)
```

```
##
## Call:
## lm(formula = pemax ~ weight + bmp + fev1 + rv + frc, data = cystfibr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39.72 -12.17   4.83  15.29  34.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.18640   54.73550   1.173 0.255423
## weight       1.73556    0.42529   4.081 0.000637 ***
## bmp         -1.35105    0.66763  -2.024 0.057303 .
## fev1         1.53087    0.62948   2.432 0.025078 *
## rv           0.13612    0.15668   0.869 0.395787
## frc         -0.02477    0.31278  -0.079 0.937703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.34 on 19 degrees of freedom
## Multiple R-squared:  0.6142, Adjusted R-squared:  0.5127
## F-statistic:  6.05 on 5 and 19 DF,  p-value: 0.001637
```

$$pemax = 64.18 + 1.73 * weight - 1.35 * bmp + 1.53 * fev1 + 0.13 * rv - 0.02 * frc + \epsilon$$

Interpretation of Model Parameters:

- β_0: The baseline Pemax value when all predictors are zero.
- β_1: A unit increase in weight corresponds to a change of 1.7356 in Pemax, holding other factors constant.
- β_2: A unit increase in bmp relates to a decrease of 1.3511 in Pemax, though this effect is marginally significant.
- β_3: A unit increase in fev1 results in a change of 1.5309 in Pemax, with other predictors unchanged.

- β_4: The effect of rv (0.1361) on Pemax is not statistically significant (p > 0.05).
- β_5: The effect of frc (-0.0248) on Pemax is not statistically significant (p > 0.05).
- ϵ: The vector of residual errors associated with each data point that the model doesn't perfectly explain.

Q2

Answer 2.a:

The General exponential distribution is represented as :

$$f(y|\theta,\phi) = exp\bigg(\frac{y\theta = b(\theta)}{a(\theta)} + c(y,\theta)\bigg)$$

Given Inverse Gaussian distribution is:

$$f(y|\mu,\gamma) = \sqrt{\frac{\gamma}{2(\pi)x^3}}exp\bigg(-\frac{\gamma(x-\mu)^2}{2\mu^2 x}\bigg)$$

$$f(y|\mu,\gamma) = exp\bigg(-\frac{1}{2}log(\frac{2(\pi)x^3}{\gamma}) - \frac{\gamma(x^2-2x\mu+\mu^2)}{2\mu^2 x}\bigg)$$

$$f(y|\mu,\gamma) = exp\bigg(-\frac{\gamma(x^2)}{2\mu^2 x} - \frac{\gamma\mu^2}{2\mu^2 x} + \frac{\gamma\mu(x)}{2\mu^2 x} - \frac{1}{2}log(\frac{2(\pi)x^3}{\gamma})\bigg)$$

$$f(y|\mu,\gamma) = exp\bigg(x\gamma\frac{\mu-\frac{1}{2}\mu^2}{2\mu^2 x} - \frac{1}{2}\bigg[\frac{\gamma(x^2)}{\mu^2 x}log(\frac{2(\pi)x^3}{\gamma})\bigg]\bigg)$$

Here, $\theta = \mu, \phi = 2\mu^2 x$

Putting these values back in the inverse gaussian distribution we get,

$$exp\bigg(\frac{\gamma(x)\theta - \frac{1}{2}\theta^2}{\phi} - \frac{1}{2}\bigg[\frac{\gamma(x^2)}{\phi} + log\bigg(\frac{2(\pi)x^3}{\gamma}\bigg)\bigg]\bigg)$$

Here,

$a(\theta) = \phi, b(\theta) = \frac{1}{2}\theta^2$

$c(y,\phi) = -\frac{1}{2}\bigg[\frac{\gamma(x^2)}{\phi} + log\bigg(\frac{2(\pi)x^3)}{\gamma}\bigg)\bigg]$

The natural parameter is $\theta = \mu$

Answer 2.a: (By other method)

The Inverse Gaussian distribution can be shown to be a member of the exponential family by expressing it in the general form:

$$f(x;\theta) = h(x)\cdot\exp\left(\frac{T(x)\cdot\theta - A(\theta)}{\phi}\right)$$

Where:

$f(x;\theta)$ is the probability density function (PDF) of the distribution.

θ is the natural parameter.

$h(x)$ is the base measure.

$T(x)$ is the sufficient statistic.

$A(\theta)$ is the log partition function.

ϕ is the dispersion parameter.

Comparing this with the Inverse Gaussian PDF:

$$f(x;\mu,\gamma) = \sqrt{\frac{\gamma}{2\pi x^3}} \cdot \exp\left(-\frac{\gamma(x-\mu)^2}{2\mu^2 x}\right)$$

We can see that the Inverse Gaussian distribution can be written in the exponential family form with:

- Natural parameter $\theta = -\frac{\mu^2}{\gamma}$
- Sufficient statistic $T(x) = x$
- Log partition function $A(\theta) = -\frac{1}{2}log\,(\gamma\theta)$
- Dispersion parameter $\theta = 1$
- Base measure $h(x) = \sqrt{\frac{\gamma}{2(\pi)x^3}}$

Answer 2.b:

Natural and Scale Parameters: From the derived exponential family form, the natural parameter is $\theta = -\frac{\mu^2}{\gamma}$ The scale parameter is not directly present in the exponential family form for this distribution.

Answer 2.c:

Mean and Variance:

The mean and variance of the Inverse Gaussian distribution can be derived from the natural parameter θ and the scale parameter ϕ

Mean: The mean (μ) of the Inverse Gaussian distribution is given by $E(X) = -\frac{\partial A(\theta)}{\partial\theta} = -\frac{1}{\theta}$

Variance: The variance σ^2 of the Inverse Gaussian distribution is given by $Var(X) = \frac{\phi}{-\frac{\partial^2 A(\theta)}{\partial\theta^2}} = \frac{\phi}{\theta^3}$

Using the relation between the natural parameter θ and the distribution parameters μ and γ, we can express the mean and variance in terms of μ and γ:

- Mean: $E(X) = \mu$
- Variance: $Var(X) = \frac{\mu^3}{\gamma}$

Q3

Answer 3.a:

Given linear model: $Y_i = X_i^T\beta + \epsilon_i$, where ϵ_i arr i.i.d normally distributed noise with mean 0 and variance σ^2

The likelihood function for a single observation (X_i, X_n) is the probability density function (PDF) of the normal distribution:

$$f(Y_i|X_i,\beta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(Y_i - X_i^T\beta)^2}{2\sigma^2})$$

The likelihood function for all n observations $(X_1, Y_1), ..., (X_n, Y_n)$ is the product of the individual likelihoods:

$$L(\beta|X,Y,\sigma^2) = \prod_{i=1}^{n} f(Y_i|X_i,\beta,\sigma^2)$$

Taking the logarithm of the likelihood (log-likelihood) simplifies the product into a sum and is a common practice for mathematical convenience:

$$logL(\beta|X,Y,\sigma^2) = -\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(Y_i - X_i^T\beta)^2$$

The MLE of the parameter vector β is the value that maximizes the log-likelihood, which is equivalent to minimizing the negative log-likelihood or, equivalently, minimizing the squared error loss:

$$LS(\beta) = \frac{1}{n}\Sigma_{i=1}^{n}(Y_i - X_i^T\beta)^2$$

This is the same form as the squared error loss function, which is minimized when the parameter vector β is chosen such that the observed responses Y_i are as close as possible to the predicted responses $X_i^T\beta$.

In summary, the MLE for the parameter vector β in the linear model with normally distributed noise is also the solution to the squared error loss minimization problem, commonly known as the least squares estimator.

Answer 3.b:

Given the Laplace distribution for the noise, the likelihood of a single observation Y_i is:

$$f(Y_i|X_i,\beta,\sigma^2) = \frac{1}{2\sigma^2}exp(-\frac{|Y_i - X_i^T\beta|}{\sigma})$$

The likelihood of the entire sample is the product of individual likelihoods:

$$L(\beta|X,Y,\sigma^2) = \prod_{i=1}^{n}\frac{1}{2\sigma^2}exp(-\frac{|Y_i - X_i^T\beta|}{\sigma})$$

Taking the logarithm of the likelihood:

$$logL(\beta|X,Y,\sigma^2) = -nlog(2\sigma) - \frac{1}{\sigma}\Sigma_{i=1}^{n}|Y_i - X_i^T\beta|$$

This is the log-likelihood for the Laplace-distributed noise.

To show that maximizing the log-likelihood is equivalent to minimizing the absolute error loss (AL), notice that the term $\frac{1}{\sigma}\Sigma_{i=1}^{n}|Y_i - X_i^T\beta|$ is proportional to the absolute error loss. Maximizing the log-likelihood is equivalent to minimizing the negative log-likelihood, which in this case is proportional to the absolute error loss:

$-\frac{1}{\sigma}\Sigma|Y_i - X_|$

Thus, maximizing the log-likelihood is equivalent to minimizing the absolute error loss.

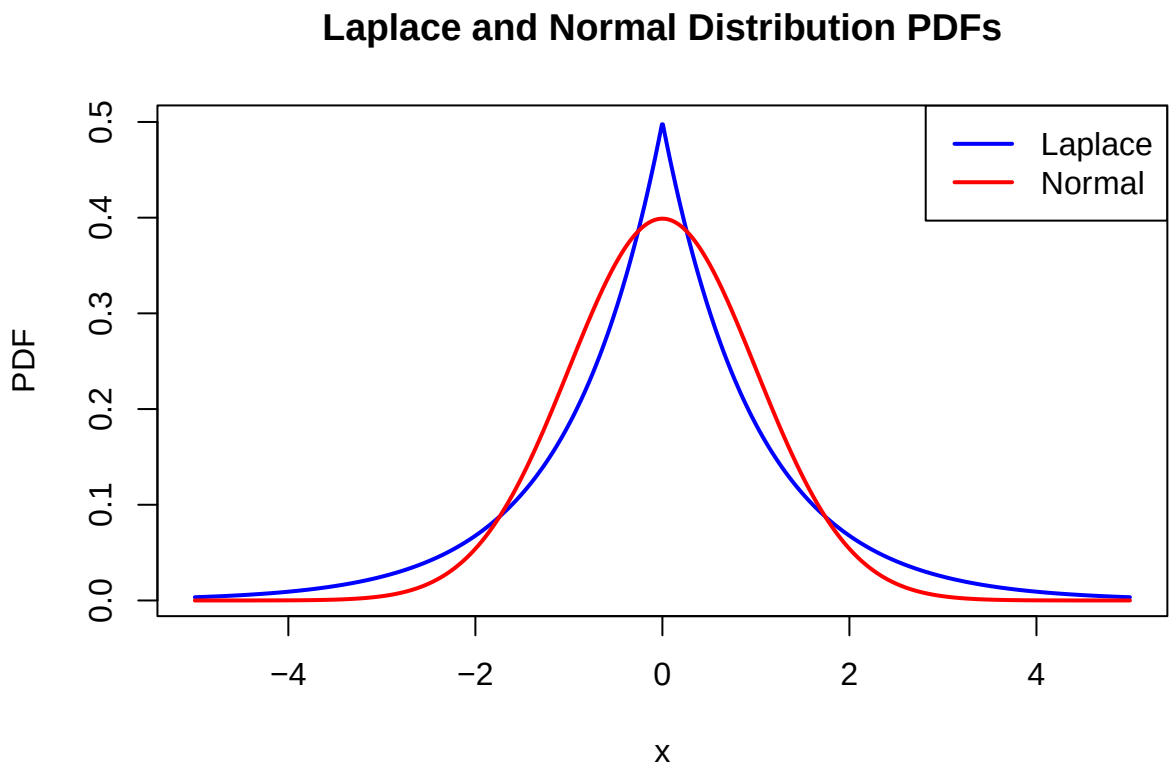
The absolute error loss (AL) is defined as:

$$AL(\beta) = \frac{1}{n}\Sigma_{i=1}^{n}|Y_i - X_i^T\beta|$$

In summary, for the Laplace distribution noise, maximizing the log-likelihood is equivalent to minimizing the absolute error loss (AL), which is the sum of the absolute differences between observed responses Y_i and predicted responses $X_i^T\beta$

Answer 3.c:

```r
# Parameters of the distributions
mu_laplace <- 0    # Laplace distribution: Location parameter
b_laplace <- 1     # Laplace distribution: Scale parameter

mu_normal <- 0     # Normal distribution: Mean
sd_normal <- 1     # Normal distribution: Standard deviation

# Generate x values
x <- seq(-5, 5, length.out = 1000)

# Calculate the Laplace PDF values for each x
laplace_pdf_values <- (1 / (2 * b_laplace)) * exp(-abs(x - mu_laplace) / b_laplace)

# Calculate the Normal PDF values for each x
normal_pdf_values <- dnorm(x, mean = mu_normal, sd = sd_normal)


# Create a line plot
plot(x, laplace_pdf_values, type = "l", lwd = 2, col = "blue",
     xlab = "x", ylab = "PDF", main = "Laplace and Normal Distribution PDFs")

# Add Normal distribution plot to the existing plot
lines(x, normal_pdf_values, type = "l", lwd = 2, col = "red")

# Add legend
legend("topright", legend = c("Laplace", "Normal"), col = c("blue", "red"), lwd = 2)
```



Answer 3.d:

The linear model using Laplace error noise is known to be more robust to outliers due to following reasons:

1. **Heavy Tails:** Laplace distribution has heavier tails, assigning higher probabilities to extreme values, reducing sensitivity to outliers.
2. **Resistant:** Laplace distribution's slower decay away from mean makes it less influenced by extreme observations.
3. **Absolute Error Minimization:** Laplace minimizes absolute differences, less affected by outliers than Normal's squared differences.
4. **Visual Comparison:** In the plot, Laplace's fat tails and sharp peak show higher tolerance for extreme values.
5. **Robust Regression:** Laplace's properties lead to stable parameter estimates in robust regression, unaffected by outliers.