STAT-448

Assignment 1

Rujal Shrestha - 29954403

29 July, 2025

Question 1

Question 1(a and b)

Question 1 (a)

$Y = \{4, 10, 16\}$ ← Response variable

$X = \{4, 6, 8\}$ ← Explanatory variable

Steps to compute $\hat{\beta} = (X'X)^{-1} X'Y$

i) Calculate $X'X$

ii) Calculate $X'Y$

iii) Calculate $(X'X)^{-1}$

iv) Substitute into formula $\hat{\beta}$

Given (adding a row of 1's to match formula $\left[\hat{Y} = 1\times\hat{\beta}_0 + x\times\hat{\beta}_1\right]$)

$$X = \begin{bmatrix} 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \qquad Y = \begin{bmatrix} 4 \\ 10 \\ 16 \end{bmatrix}$$

$$X' = \begin{bmatrix} 1 & 1 & 1 \\ 4 & 6 & 8 \end{bmatrix}$$

Step 1 $X'X = \begin{bmatrix} 1 & 1 & 1 \\ 4 & 6 & 8 \end{bmatrix}_{2\times3} \begin{bmatrix} 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix}_{3\times2}$

$$X'X = \begin{bmatrix} 3 & 18 \\ 18 & 116 \end{bmatrix}_{2\times2}$$

STEP 2 $X'Y = \begin{bmatrix} 1 & 1 & 1 \\ 4 & 6 & 8 \end{bmatrix}_{2\times3} \begin{bmatrix} 4 \\ 10 \\ 16 \end{bmatrix}_{3\times1}$

$$X'Y = \begin{bmatrix} 30 \\ 204 \end{bmatrix}$$

Step 3

$(X'X)^{-1} = \frac{1}{|X'X|} \text{Adj}(X'X)$ ← formula for inverse of a matrix

$|X'X| = 3\times116 - 18\times18$

$= 24$ ← non-zero determinant, so determinant of $X'X$ exists

$\text{cofactor}(X'X) = \begin{bmatrix} 116 & -18 \\ -18 & 3 \end{bmatrix}$

$$\text{Adj}\,(X'X) = [\text{cofactor}\,(X'X)]^{T} \quad \text{(transpose)} = \begin{pmatrix} 116 & -18 \\ -18 & 3 \end{pmatrix}$$

$$\therefore (X'X)^{-1} = \frac{1}{24}\begin{bmatrix} 116 & -18 \\ -18 & 3 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} \frac{116}{24} & -0.75 \\ -0.75 & 0.125 \end{bmatrix}$$

now, to compute $\hat{\beta}$, we use the formula

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} \frac{116}{24} & -0.75 \\ -0.75 & 0.125 \end{bmatrix}\begin{bmatrix} 30 \\ 204 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} -8 \\ 3 \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

β_0: intercept, β_1: slope

$$\hat{\beta}_0 = -8, \quad \hat{\beta}_1 = 3$$

$\therefore$ the estimated model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\hat{y} = -8 + 3x \quad \leftarrow \text{ANSWER}$$

Question 1(b)

x	y	$\hat{y}$	$\epsilon = y_i - \hat{y}_i$
4	4	4	$4-4=0$
6	10	10	$10-10=0$
8	16	16	$16-16=0$

Question 1(c)

```
# define matrices

# explanatory variable
x <- matrix(c(1, 1, 1, 4, 6, 8), byrow = FALSE, nrow = 3)

# response variable
y <- matrix(c(4, 10, 16), nrow = 3)

print(x)
```

```
##      [,1] [,2]
## [1,]    1    4
## [2,]    1    6
## [3,]    1    8
```

```
print(y)
```

```
##      [,1]
## [1,]    4
## [2,]   10
## [3,]   16
```

```
# transpose of x

x_t <- t(x)

print(x_t)
```

```
##      [,1] [,2] [,3]
## [1,]    1    1    1
## [2,]    4    6    8
```

The matrix dimensions have changed from 3x2 to 2x3

```
x_t_x <- x_t %*% x

print(x_t_x)
```

```
##      [,1] [,2]
## [1,]    3   18
## [2,]   18  116
```

```
x_t_y <- x_t %*% y

print(x_t_y)
```

```
##      [,1]
## [1,]   30
## [2,]  204
```

```r
# now calculating the inverse of x_t_x

inv_x_t_x <- solve(x_t_x)
print(inv_x_t_x)
```

```
##           [,1]   [,2]
## [1,]  4.833333 -0.750
## [2,] -0.750000  0.125
```

```r
# as we have all the necessary elements, we now substitute the values
# in the formula and calculate the coefficient matrix

b <- inv_x_t_x %*% x_t_y
print(b)
```

```
##      [,1]
## [1,]   -8
## [2,]    3
```

Which matches the figures from manual calcuations perfectly.

Now, Calculating the estimates of residuals

```r
y_hat <- x %*% b
e <- y - y_hat

print(e) # output can be considered as zero
```

```
##              [,1]
## [1,] 1.598721e-14
## [2,] 1.421085e-14
## [3,] 1.065814e-14
```

Question 1(d)

```r
x <- c(4, 6, 8)
y <- c(4, 10, 16)
```

```r
df <- data.frame(x, y)

print(df)
```

```
##   x  y
## 1 4  4
## 2 6 10
## 3 8 16
```

```r
model <- lm(y ~ x, data = df)
summary(model)
```

```
## Warning in summary.lm(model): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##          1          2          3
##  2.421e-16 -4.842e-16  2.421e-16
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -8.000e+00  1.304e-15 -6.136e+15   <2e-16 ***
## x            3.000e+00  2.097e-16  1.431e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.93e-16 on 1 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.047e+32 on 1 and 1 DF,  p-value: < 2.2e-16
```

The coefficients match the earlier values exactly. But this method is extensively more convinient as all those efforts were easily abstracted within a few lines of code. Moreover, we get additional insights like R-squared, significance values too. According to the summary, the exploratory value seems to be very significant with a very small p-value.

Question 2

```r
new_x <- c(7, 7, 7)

df2 <- data.frame(new_x, y)

print(df2)
```

```
##   new_x  y
## 1     7  4
## 2     7 10
## 3     7 16
```

```r
model2 <- lm(y ~ new_x, data = df2)

summary(model2)
```

```
##
```

```
## Call:
## lm(formula = y ~ new_x, data = df2)
##
## Residuals:
##          1          2          3
## -6.000e+00 -2.034e-15  6.000e+00
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.000      3.464   2.887    0.102
## new_x             NA         NA      NA       NA
##
## Residual standard error: 6 on 2 degrees of freedom
```

Question 2(a)

The slope coefficient β_1 is `NA` or `undefined`.

The intercept coefficient β_0 is 10, which is just the average of Y

Question 2(b)

Statistical intuition is that the explanatory variable X is collinear with the intercept. The variance of X is zero, which when put into the formula of β_1, (ie $\sum(X_i - \bar{X})^2 = 0$), which ends up making the of β_1 undefined.

Question 2(c)

Geometrically speaking, there is no horizontal spread accross X axis, due to which the slope(β_1) is undefined. To clarify, slope(β_1) is the variation of Y with respect to variation in X. As X is constant for all the points, the regression algorithm is unable to separate the effect of X from the intercept

Question 3

```r
df <- read.csv("dataset/Student_Scores_Dataset.csv", header = TRUE)
head(df)
```

```
##      Hours   Scores
## 1 4.370861 49.59712
## 2 9.556429 93.88757
## 3 7.587945 82.78044
## 4 6.387926 71.93219
## 5 2.404168 31.84063
## 6 2.403951 34.44341
```

```r
model <- lm(Scores ~ Hours, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = Scores ~ Hours, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8154  -3.1299  -0.1679   3.2371  15.5653
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.52998    0.34948   18.68   <2e-16 ***
## Hours        9.76317    0.05809  168.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.827 on 998 degrees of freedom
## Multiple R-squared:  0.9659, Adjusted R-squared:  0.9658
## F-statistic: 2.825e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

Question 3(a)

The regression equation for the student score is as follows:

$$\hat{y} = 6.530 + 9.763x$$

Question 3(b)

The coefficient $\beta_1 = 9.763$ can be interpreted as - A student's score increases by 9.763 for every additional hour they study.

Question 3(c)

Yes, the hours of study as a very signification effect on the scores, this is also supported by the very small p-value of $2e - 16$, which denotes that changes in scores are heavily influenced by variations in study hours.

Question 3(d)

Yes, the model provides a very good fit for the observed data. This is justified by the following figures:

1. The $MedianResiduals = -0.1679$ indicates that the observed values (Y) and predicted values ($\hat{Y}$) is small and the model was able to fit very well to the data
2. The $R^2 = 0.9658$ denotes that the model explains most (96.6) of the variability in the data
3. Also, The small p-value ($2.2e - 16$) also further supports this by indicating that the model is highly significant

Question 3(e)

Residual Analysis

```
residuals <- model$residuals
fitted <- fitted(model)

df_residuals <- data.frame(residuals = residuals, fitted = fitted)

df_residuals |>
  ggplot(aes(x = fitted, y = residuals)) +
  geom_point(size = 0.5) +
  geom_hline(yintercept = 0, color = "red", size = 2) +
  geom_hline(yintercept = 2 * sd(model$residuals), color = "blue", size = 1) +
  geom_hline(yintercept = -2 * sd(model$residuals), color = "blue", size = 1) +
  labs(
    caption = "Figure: Fitted vs Residuals plot for residual analysis"
  ) +
  theme(
    plot.caption = element_text(hjust = 0.5)
  )
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

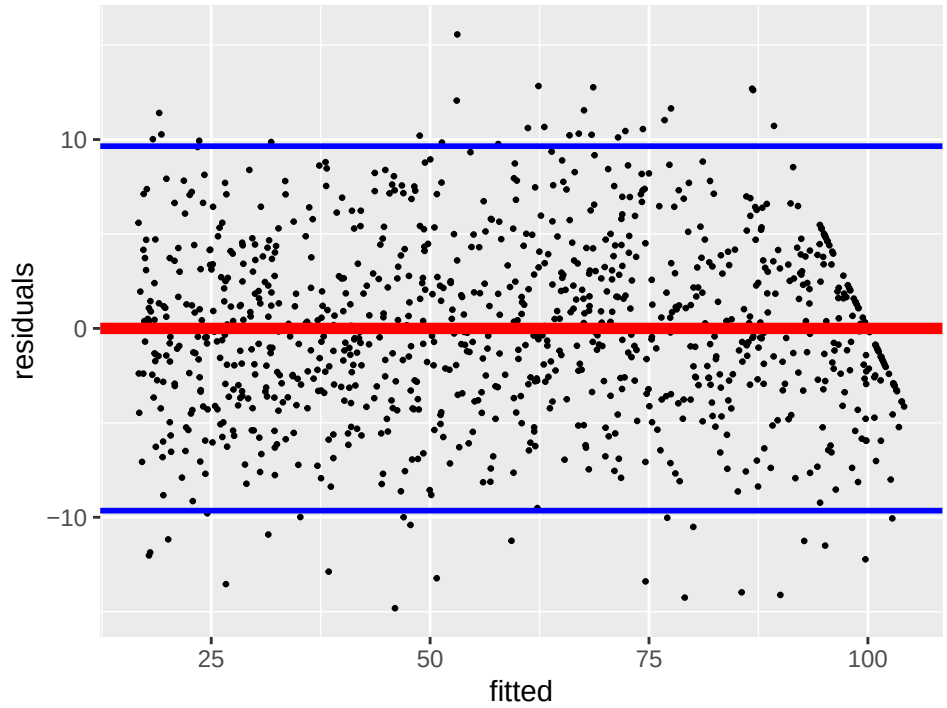

Figure: Fitted vs Residuals plot for residual analysis

As show in the plot above, the residual points are scattered evenly around zero (the red line), which suggests that the assumption of the model's linearity is reasonable. Moreover, as the points are evenly scattered throughout all the values of fitted axis (without any funneling patterns), there seems to be no presence of homoscadesticity.

A point to be noted is the funneling line at the right end of the plot. This is due to the scores maxing out at 100, also called "boundary effect".

Question 3(f)

```
coefficients <- coef(model)

beta_0 <- coefficients[1]
beta_1 <- coefficients[2]

df |>
  ggplot(aes(x = Hours, y = Scores)) +
  geom_point(size = 0.5) +
  geom_abline(intercept = beta_0, slope = beta_1, color = "red", size = 1) +
  labs(
    title = "Regression plot for Scores vs hours studied",
    subtitle = expression(hat(y) == 6.530 + 9.763 * x)
  ) +
  theme_light() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5)
  )
```

Question 3(g)

```
new_x <- data.frame(Hours = c(3.63, 5.68, 7.48))

predicted_scores <- round(predict(model, newdata = new_x), 2)

new_x$predicted_scores <- predicted_scores

kable(new_x)
```

Hours	predicted_scores
3.63	41.97
5.68	61.98
7.48	79.56

It is not valid to make predictions outside of the hours of study range (or generally known as extrapolation) as the linear regression model was estimated using the provided data which might not be in similar trend outside of the range, due to which the model's predictions can end up varying.

Question 4

```
monkey <- read.csv("./dataset/macaque.csv")

df <- monkey |>
  select(age, mean_fertility)
```

Question 4(a)

```
plot_1 <- df |>
  ggplot(aes(x = age, y = mean_fertility)) +
  geom_point() +
  labs(
    title = "Macque average fertility for different age",
    x = "Age",
    y = "Mean age specific fertility",
    caption = "Fig: Fertility trend of female macaques accross age"
  ) +
  theme_light() +
  theme(
    plot.caption = element_text(hjust = 0.5)
  )

plot_1
```

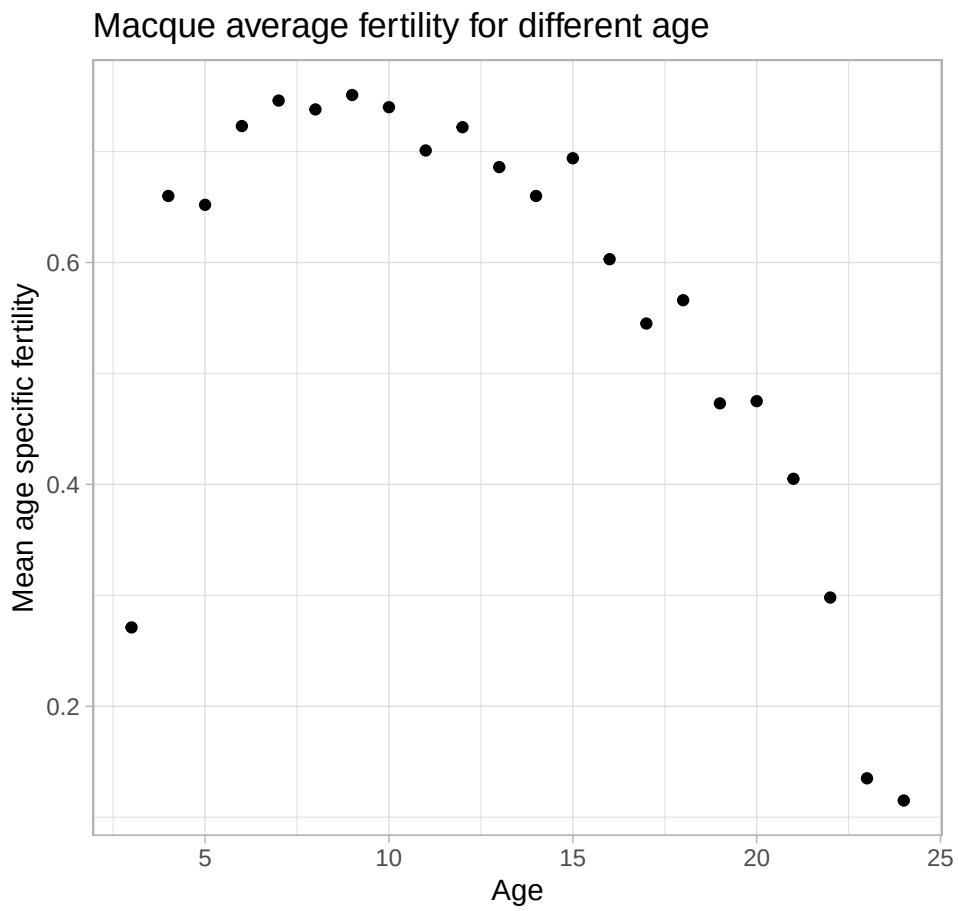

Fig: Fertility trend of female macaques accross age

The association is clearly non-linear as illustrated by the rising, peaking and falling curve of the plot. The fertility grows as the macaque ages and peaks at around 9 years of age. Following this is a sharp declining as the macaques age and attaining minimal values at the end of 24 years.

Question 4(b)

```
model <- lm(mean_fertility ~ age, data = df)

summary(model)
```

```
##
## Call:
## lm(formula = mean_fertility ~ age, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49299 -0.05819  0.05484  0.09969  0.16112
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.821764   0.080313  10.232 2.15e-09 ***
## age         -0.019259   0.005384  -3.577  0.00189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1602 on 20 degrees of freedom
## Multiple R-squared:  0.3901, Adjusted R-squared:  0.3596
## F-statistic: 12.79 on 1 and 20 DF,  p-value: 0.001887
```

The simple linear regression equation is:

$$\hat{y} = 0.822 - 0.019x$$

Although it is apparent that a simple linear equation is not the best fit model for this dataset, the above equation can be interpreted as:

$\beta_{\mathbf{0}} = \mathbf{0.822}$: mean fertility at age 0 is 0.822, which is soundly impractical for this scenario

$\beta_{\mathbf{1}} = -\mathbf{0.019}$: there is drop of 0.019 mean fertility for every year a female macaque ages

Question 4(c)

```
plot_1 + geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```

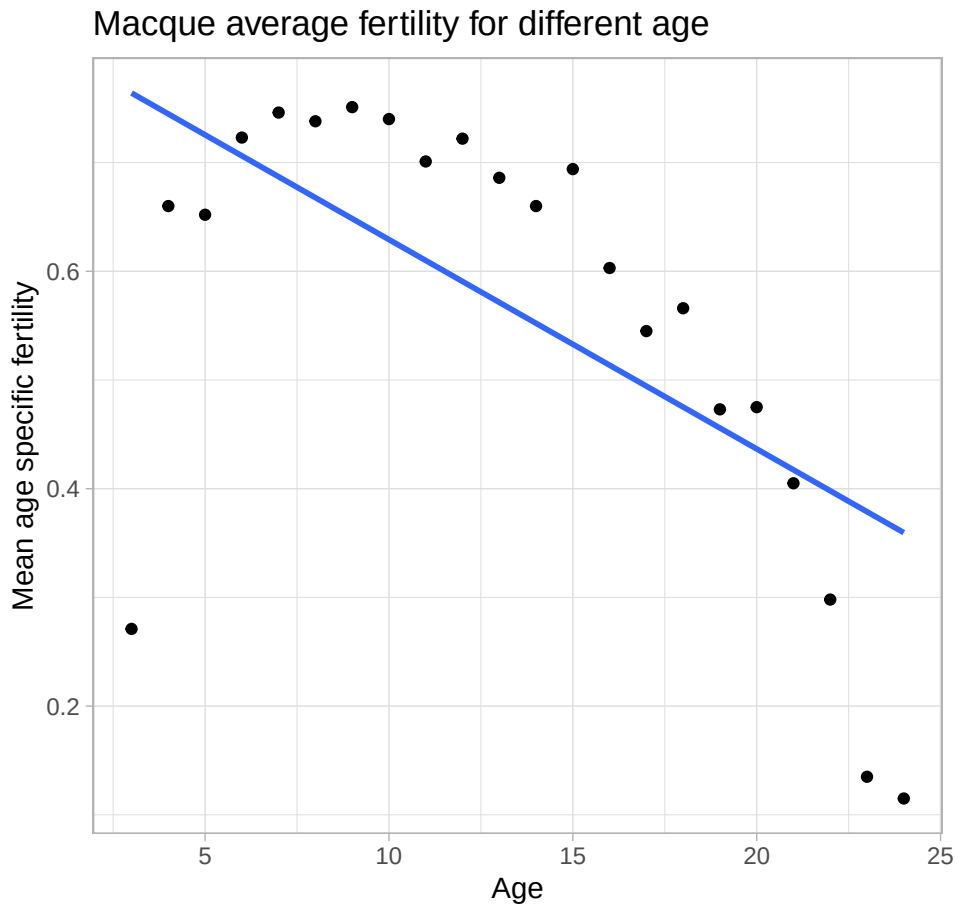

Fig: Fertility trend of female macaques accross age

The plot suggests an inadequate fit. There is a curving trend on the points, whereas the linear regression line passes straight under it.

```
draw_residual_plot <- function(df) {
  # nolint start
  ggplot(data = df, aes(x = fitted, y = residuals)) +
    geom_point() +
    geom_hline(yintercept = 0, color = "red", size = 2) +
    geom_hline(yintercept = 2 * sd(model$residuals), color = "blue", size = 1) +
    geom_hline(yintercept = -2 * sd(model$residuals), color = "blue", size = 1) +
    labs(
      caption = "Fig: Residual plot for linear regression model",
      x = "Fitted",
      y = "Residuals"
    ) +
    theme(
      plot.caption = element_text(hjust = 0.5)
    )
  # nolint end
}

df_residuals <- data.frame(
  residuals = model$residuals,
```

```
  fitted = fitted(model)
)

draw_residual_plot(df_residuals)
```

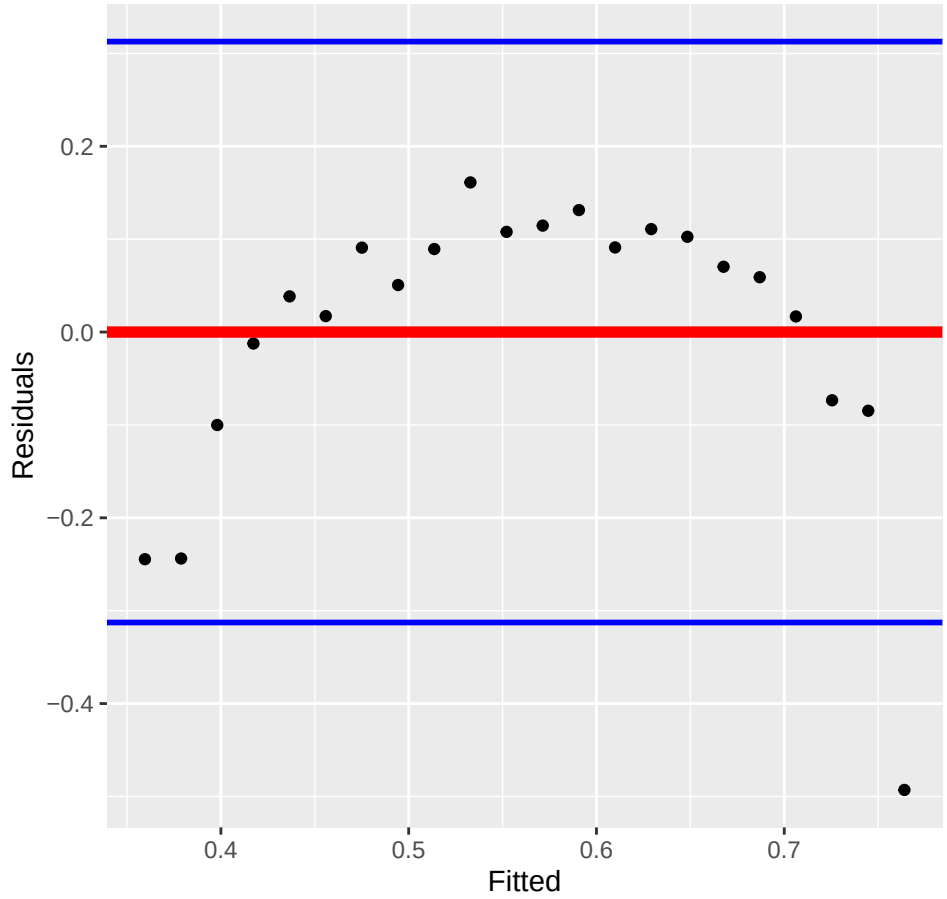

Fig: Residual plot for linear regression model

The curved pattern in the residual plot above clearly indicates that this model is **not adequate**, ie the linear model fails to fully capture the underlying relationship between the variables. A good fitting model usually is expected to have no patterns and all the points evenly spread out near the center line.

Question 4(d)

```
plot_1 +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2), se = FALSE)
```

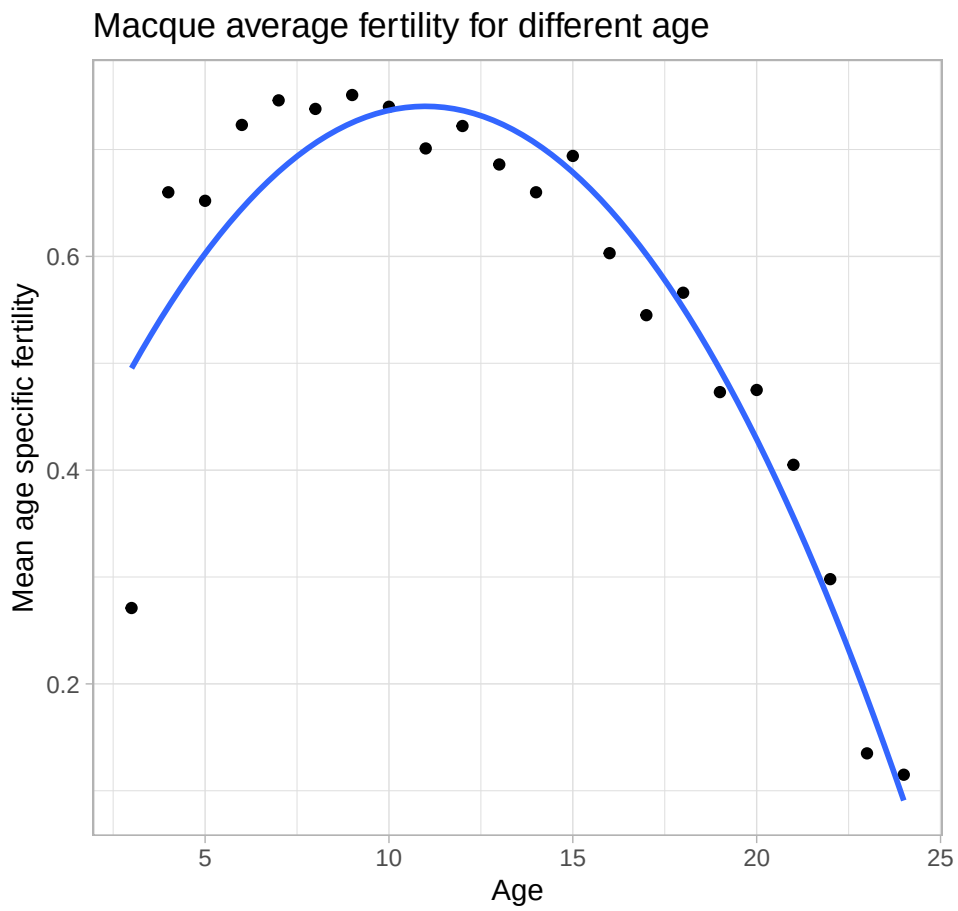

Fig: Fertility trend of female macaques accross age

It makes sense to add a quadratic term of degree 2 to the equation to better fit the curve. After changing the equation to quadratic of degree 2, the model was able to fit the datapoints better.

```
quadratic_model <- lm(mean_fertility ~ age + I(age^2), data = df)

summary(quadratic_model)
```

```
##
## Call:
## lm(formula = mean_fertility ~ age + I(age^2), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22433 -0.03929  0.01470  0.04254  0.10717
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2767658  0.0697498   3.968 0.000824 ***
## age          0.0843678  0.0116488   7.243 7.11e-07 ***
## I(age^2)    -0.0038380  0.0004223  -9.089 2.40e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.07108 on 19 degrees of freedom
## Multiple R-squared:  0.886,  Adjusted R-squared:  0.874
## F-statistic: 73.81 on 2 and 19 DF,  p-value: 1.101e-09
```

```
quadtratic_df_residuals <- data.frame(
  residuals = quadratic_model$residuals,
  fitted = fitted(quadratic_model)
)

draw_residual_plot(quadtratic_df_residuals)
```

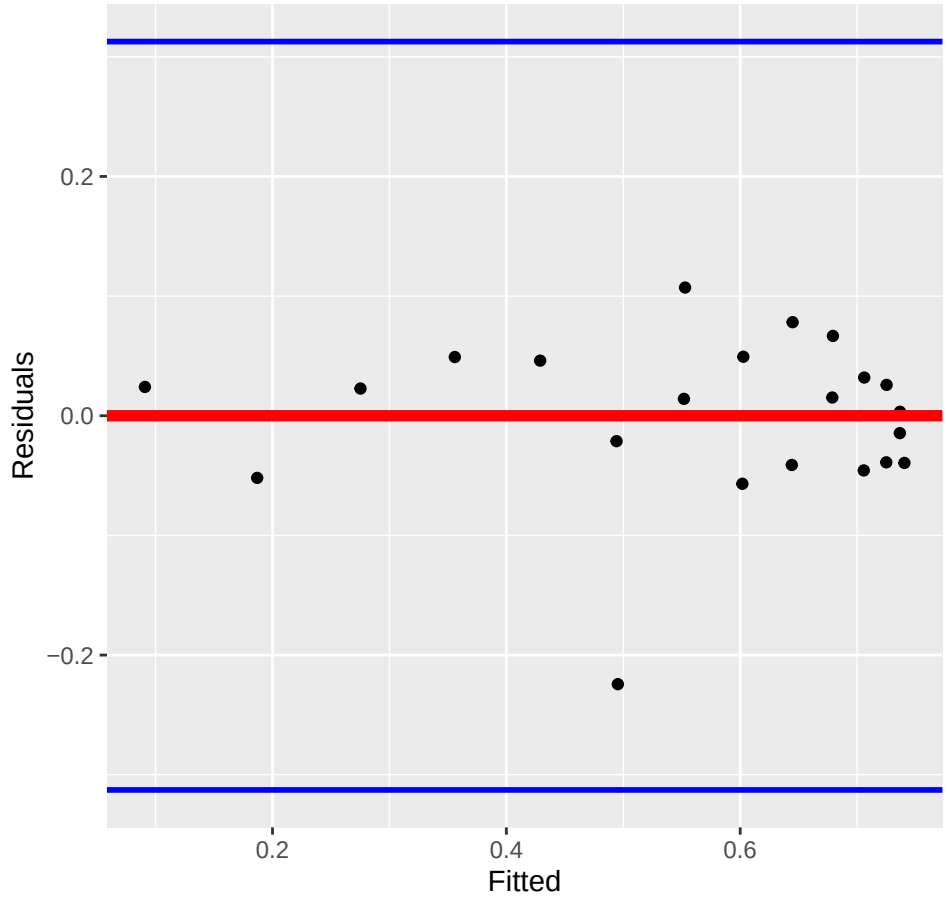

Fig: Residual plot for linear regression model

The residual plot of the quadratic linear model also is better. There is no more presence of a pattern, the points are scattered randomly near the center line which suggests that there are no more hidden patterns unaccounted by the model.

Additionally, there seems to be a uniform distribution of vertical spaces accross the fitted axes, which indicates lack of homoscedasticity.

Lastly, the curvature evident on the previous plot is no longer visible, which proves linearity, ultimately supporting this model to be more adequate as compared to a simple linear model of degree 1.

Question 4(e)

```r
df_new <- data.frame(age = c(6.95, 12.35, 15.87))

predictions <- round(predict(model, newdata = df_new), 2)

df_new$predicted_mean_fertility <- predictions

knitr::kable(df_new, caption = "Table: Predicted Fertility for Given Ages")
```

Table 2: Table: Predicted Fertility for Given Ages

age	predicted_mean_fertility
6.95	0.69
12.35	0.58
15.87	0.52

The predicted value make sense even from eye-level examination, it was already evident from the summarization performed earlier that the fertility in female macaques peaked between 7 - 10 and started degrading after that. This trend can be seen in the above predictions, the higher the age, the lower the fertility fell.