STAT448 - ASS1

Noah KIng

2024-09-03

Question 1

a Y_hat = X$Beta_hat$ = $X(X\hat{}t$ X)^-1 X^tY

Firstly; define X = [6,7,8] Define Y= [9,16,32]

```
X <- c(6,7,8)
Y <- c(9,16,23)
```

\begin{document}

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

where $\hat{\beta}$ is the vector of estimates $[\hat{\beta}_0, \hat{\beta}_1]^T$, X is the design matrix, and Y is the vector of observed values.

1. Define the Matrices

The design matrix X and the vector Y are:

$$X = \begin{bmatrix} 1 & 6 \\ 1 & 7 \\ 1 & 8 \end{bmatrix}$$

$$Y = \begin{bmatrix} 9 \\ 16 \\ 23 \end{bmatrix}$$

2. Compute $X^T X$

First, we compute the ^t of X, and then multiply it by X (upack the equation above)

$$X^T = \begin{bmatrix} 1 & 1 & 1 \\ 6 & 7 & 8 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 6 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 6 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 3 & 21 \\ 21 & 149 \end{bmatrix}$$

3. Compute $(X^T X)^{-1}$

Calculate the inverse of $X^T X$ using the formula for a 2×2 matrix:

$$\det(X^T X) = 3 \cdot 149 - 21 \cdot 21 = 447 - 441 = 6$$

$$(X^T X)^{-1} = \frac{1}{6} \begin{bmatrix} 149 & -21 \\ -21 & 3 \end{bmatrix} = \begin{bmatrix} 24.8333 & -3.5 \\ -3.5 & 0.5 \end{bmatrix}$$

4. Compute $X^T Y$

Compute $X^T Y$:

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 \\ 6 & 7 & 8 \end{bmatrix} \begin{bmatrix} 9 \\ 16 \\ 23 \end{bmatrix} = \begin{bmatrix} 48 \\ 350 \end{bmatrix}$$

5. Compute $\hat{\beta}$

Finally, calculate the OLS estimates:

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \begin{bmatrix} 24.8333 & -3.5 \\ -3.5 & 0.5 \end{bmatrix} \begin{bmatrix} 48 \\ 350 \end{bmatrix} = \begin{bmatrix} -33 \\ 7 \end{bmatrix}$$

The OLS value estimates are $\hat{\beta}_0 = -33$ and $\hat{\beta}_1 = 7$.

i.e.

Goal:

b We can get an idea of the performance of our model by calculating our residuals (.e. the diff between prediction and value)

To calculate the residuals $\hat{\epsilon}$, we follow these steps:

1. Compute the Predicted Values

Given the design matrix X and the OLS estimates $\hat{\beta}$, the predicted values $\hat{Y}$ are calculated as:

$$X = \begin{bmatrix} 1 & 6 \\ 1 & 7 \\ 1 & 8 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} -33 \\ 7 \end{bmatrix}$$

The regression model is:

$$\hat{Y} = X\hat{\beta}$$

Compute $\hat{Y}$ as follows:

$$\hat{Y} = \begin{bmatrix} 1 & 6 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} -33 \\ 7 \end{bmatrix}$$

Calculate the predicted values:

$$\hat{Y}_1 = (1 \times -33) + (6 \times 7) = -33 + 42 = 9$$

$$\hat{Y}_2 = (1 \times -33) + (7 \times 7) = -33 + 49 = 16$$

$$\hat{Y}_3 = (1 \times -33) + (8 \times 7) = -33 + 56 = 23$$

Thus:

$$\hat{Y} = \begin{bmatrix} 9 \\ 16 \\ 23 \end{bmatrix}$$

2. Compute the Residuals

Given the observed values Y:

$$Y = \begin{bmatrix} 9 \\ 16 \\ 23 \end{bmatrix}$$

The residuals $\hat{\epsilon}$ are computed as:

$$\hat{\epsilon} = Y - \hat{Y}$$

Subtracting the predicted values from the observed values:

$$\hat{\epsilon}_1 = 9 - 9 = 0$$

$$\hat{\epsilon}_2 = 16 - 16 = 0$$

$$\hat{\epsilon}_3 = 23 - 23 = 0$$

Thus:

$$\hat{\epsilon} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The residuals are all zeros for each data point. This is expected because the datapoints are intuitively just a linear function (i.e. 678 then 6912 are obivously linear)

c We can validate our ideas by performing these steps programmatically (am abridged version of the linear algebra above in q1, q2) (note: with R code in this doc, we will redefine use static values in each chunk)

We break out each step of the L to R equation from the top for our betas:

Firstly; the Beta coefficients...

```
X <- c(6,7,8)
Y <- c(9,16,23)
X <- cbind(1, X)

# XTX
X_t <- t(X)
X_t_X <- X_t %*% X


# inverse of X^T X
X_t_X_inv <- solve(X_t_X)
#XT Y
X_t_Y <- X_t %*% Y


beta_hat <- X_t_X_inv %*% X_t_Y

print(beta_hat)
```

```
##   [,1]
##    -33
## X    7
```

```
Y_hat <- X %*% beta_hat
Y_hat <- as.vector(Y_hat)
residuals <- Y - Y_hat
Y
```

```
## [1]  9 16 23
```

```
Y_hat
```

```
## [1]  9 16 23
```

```
#The below is just to account for floating point errors in R.
threshold <- 1e-10
residuals[abs(residuals) < threshold] <- 0



print(residuals)
```

```
## [1] 0 0 0
```

and we confirm programmatically that our L.A. calculations and the residuals are correct.

d And finally we confirm the coeffcieints using R's inbuilt LM function the intercept is B_0 and slope is the 'X' value we see is B_1

```
X <- c(6,7,8)
Y <- c(9,16,23)
model <- lm(Y~X)
coefficients(model)
```

```
## (Intercept)           X
##         -33           7
```

Question 2

(10 marks) In the context of question 1, consider the case where the values observed for the explanatory variable X are {2, 2, 2}.

Effectively; replacing with X = [2,2,2]; that is all explanatory vars are the same

a a). What happens to the coefficient estimates? (3 marks).

Intuitively; our slope becomes zero. We have no data points to use. In terms of computing using (non LA) methods, we define B_1_hat in terms of covariance and variance; in which we divide by zero. There is no valid slope. We see this in the inbuilt function:

The intercept is simply the mean of Y.

```
X <- c(2,2,2)
Y <- c(9,16,23)
model <- lm(Y~X)
coefficients(model)
```

```
## (Intercept)           X
##          16          NA
```

b b). Using appropriate terminology give both a statistical and geometric explanation of this situation. (7 marks). All observed variables are the same; so there is zero variability in X; this means that the X variable provides noinformation about the response variable Y (in terms of statistics, as above, the var = 0; cov = 0 therefore slope = 0)

Geometrically, as all observed values of X are the same, the data points lie on a vertical line in the (X, Y) plane. In this case, there is no spread or variation along the axis-X, which means that the regression line cannot be oriented (i,e no slope) to describe the relationship between X and Y .

Question 3

(35 marks) Using the data in the provided CSV file (happy.csv), generate a simple linear regression model to describe the association between Income and Happiness. Then answer the questions below. Note: 1 unit of Income (feature) is $10,000 and Happiness (target) is a scale, 1-10.

```r
happiness_data <- read.csv("Happy data.csv", header=TRUE, sep=",")
model <- lm(happiness ~ income, data = happiness_data)
summary(model)
```

Pre:

```
##
## Call:
## lm(formula = happiness ~ income, data = happiness_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02479 -0.48526  0.04078  0.45898  2.37805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20427    0.08884   2.299   0.0219 *
## income       0.71383    0.01854  38.505   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7181 on 496 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7488
## F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

a

The regression equation is therefore:

Y = 0.204 + 0.714X

b

(i.e happiness = b0 + b1income)

I.e. happiness = 0.204 when income is zero per unit increase in income, happiness increases by 0.714

c

The association between income and happiness is reasonable; i.e it has a significant effect on happiness. We can infer this from the P-value; which is very small (2e-16) and typically <0.05 is used to determine statisical significance. (Also R puts three Asterixes next to highly significant variables)

The T-value is large. (38.505), which supports the p-value as above (high t-value usually means low p-value)

d

Per the residuals etc here:

R^2 is 0.7493; i.e. ~~75% of the variability is explained by income, which suggest a strong relationship (between the estimator and the observed, hence a good fit)

Adjusted R-squared suggests similar strengths in the relationship

RSE

F-stat is again small (similar idea to the p/t-value as above)

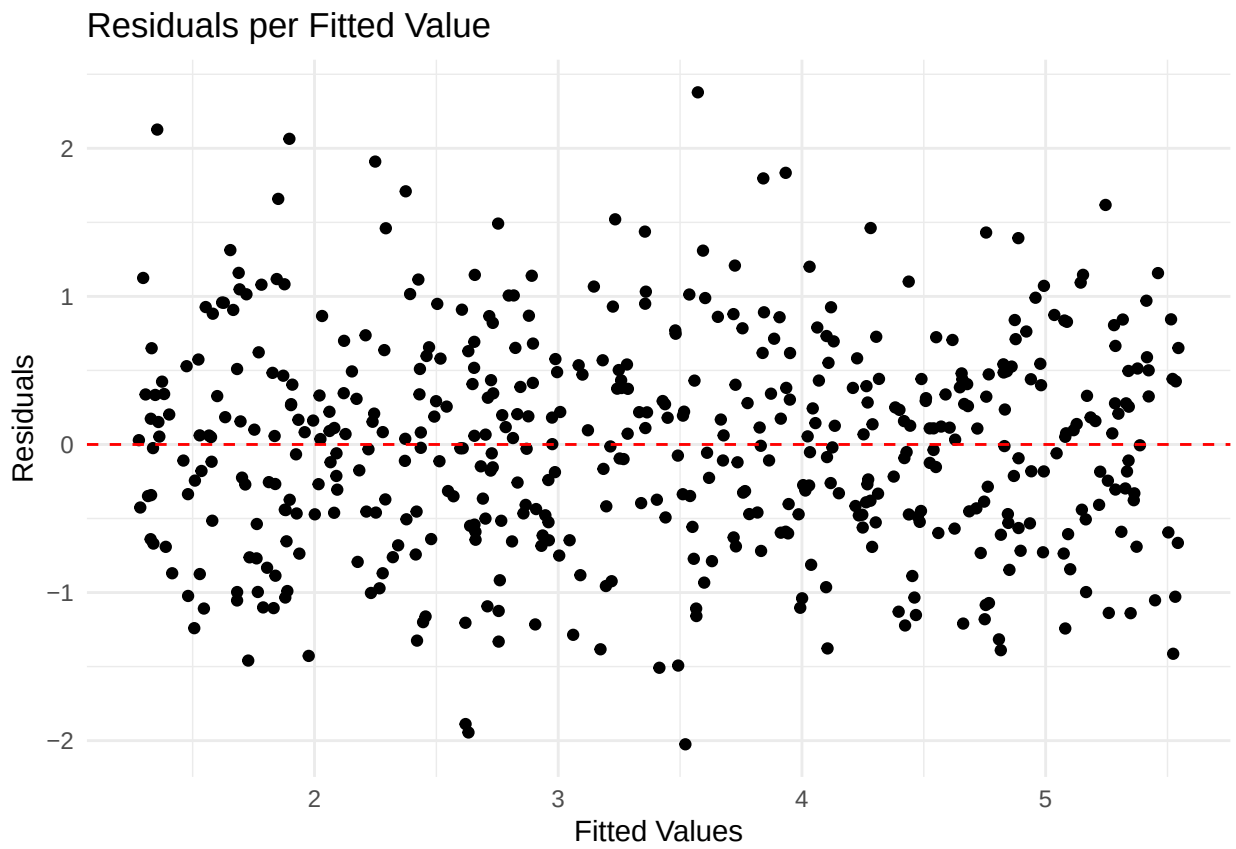
e

We need to assess whether our regression model is a good fit for the data. In particular, we want to ensure that the explanatory variable (income) has a linear relationship with the response variable (happiness) and that the residuals exhibit constant variance (homoscedasticity). To check these assumptions, we plot the residuals against the fitted values below:

```
library(ggplot2)

# pop these into the df
happiness_data$fitted_values <- fitted(model)
happiness_data$residuals <- residuals(model)


ggplot(happiness_data, aes(x = fitted_values, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals per Fitted Value",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()
```



Since the Residuals vs. Fitted Values plot shows no obvious pattern or curve, it suggests that the relationship between the predictors and the response is linear, and you're not violating the assumption of linearity. The absence of any funnel-like pattern or significant change in the spread of residuals indicates that the variance is likely constant, meaning homoscedasticity is not violated.

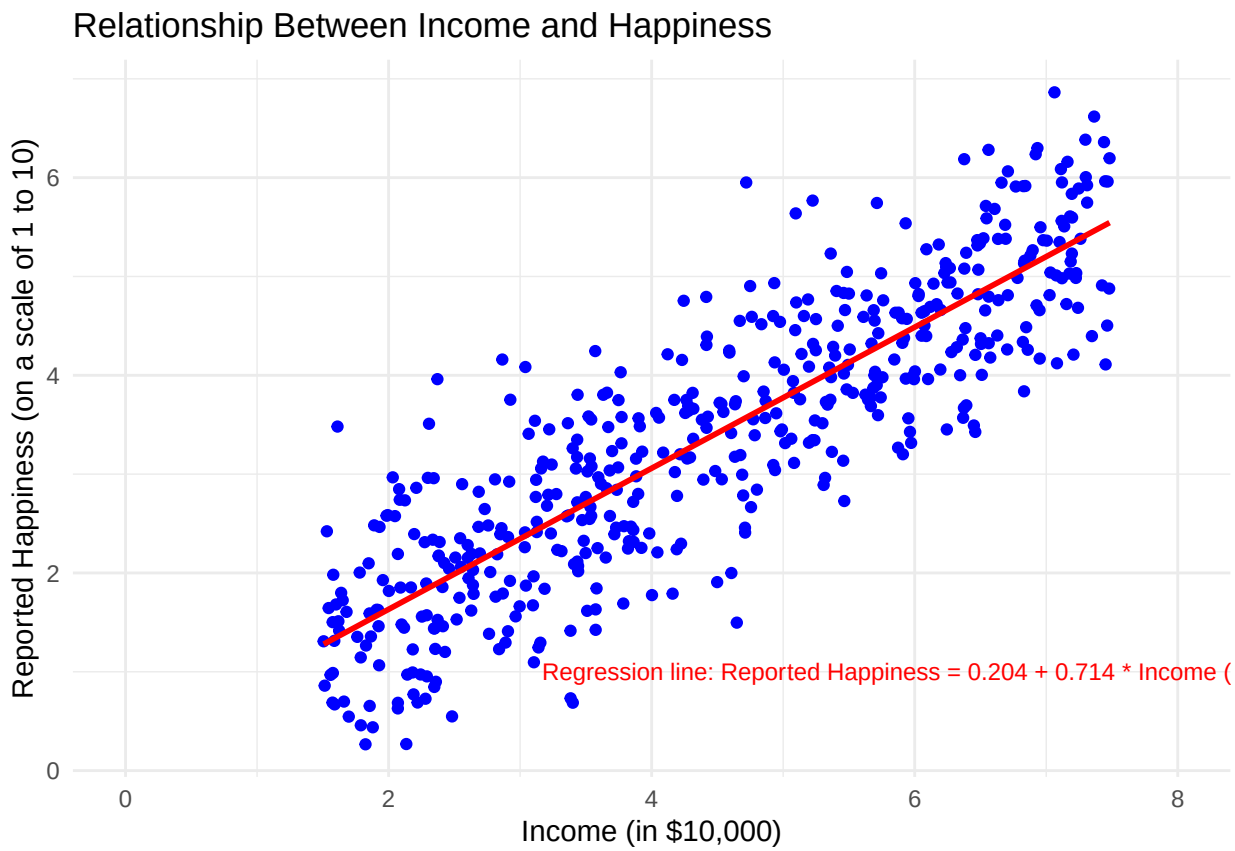
This implies that the model assumptions of linearity and constant variance are reasonably satisfied, and the model seems to be a good fit based on this (above_ diagnostic)

5 As we are satisfied that our model provides a good fit, we now consider the model itself (as a graph):

```
library(ggplot2)

# Scatter plot of the observations with regression line and adjusted income scale
ggplot(happiness_data, aes(x = income, y = happiness)) +
  geom_point(color = "blue") +  # Plot the observations
  geom_smooth(method = "lm", se = FALSE, color = "red") +  # Add the regression line
  labs(
    title = "Relationship Between Income and Happiness",
    x = "Income (in $10,000)",
    y = "Reported Happiness (on a scale of 1 to 10)"
  ) +
  xlim(0, 8) +  # Adjust income scale (modify limits as needed)
  annotate("text", x = 6, y = 1, label = "Regression line: Reported Happiness = 0.204 + 0.714 * Income
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



6:

We see the above regression model (i.e the red line on the graph above) and we can therefore create predictions on unseen data: (we did not create a train set for this data, so let's use some predefined calues)

Income(units of 10k) -> Happiness (rating of 10)

4.76 -> 3.61

6.48 -> 4.83

8.98 -> 6.62

Why shouldn't we predict outside of the training set income range?

If we were to do this; the accuracy of the model relies on the relationship observed in the training data continuing further 'up' the scale; we simply don't know this to be true. Hence; our uncertainty increases. Interpolating or extrapolating the data is fine in domains where we should very reasonably expect a linear ongoing relationship

Predictions outside the observed income range may not accurately reflect the true relationship between income and happiness.

Question 4

(45 marks) Simple linear regression models serve as crucial tools for exploring relationships between variables. These models offer insights into how the response variable (typically plotted on the y-axis) changes as the explanatory variable (usually on the x-axis) varies. Given that many biological variables are continuous and tend to follow a normal distribution, simple linear regression models play a foundational role in elucidating these associations between variables of interest.

In this context, we will employ a simple linear regression model to investigate the connections between fertility rate and age in female rhesus macaques from Cayo Santiago. We will use reproductive data from Cayo Santiago rhesus macaque females, as documented in Luevano et al. (2022). Our goal is to determine whether female fertility is influenced by age through the application of simple linear regression analysis. This dataset represents authentic information collected via daily visual censuses conducted by the staff of the Caribbean Primate Research Center (CPRC) at the University of Puerto Rico-Medical Sciences Campus.

The original dataset contains a total of 14,401 rows, each providing information on the reproductive performance of females at various stages of their lives. We have calculated the mean age-specific fertility rate, which is defined as the number of offspring produced at age 'x' divided by the total number of females of age 'x'. To achieve this, we have grouped the rows by age and computed the mean fertility rate for each age category. You can access the resulting dataset, 'macaque.csv', to assist in completing the tasks and addressing the questions posed in this analysis.

Pre

No need to remove cols.

```
repro_data <- read.csv("Macaque data.csv", header=TRUE, sep=",")
repro_data
```

```
##    age  variable    n min max median q1 q3 iqr mad mean_fertility    sd    se
## 1    3 offspring 2108   0   1      0  0  1   1   0          0.271 0.445 0.010
## 2    4 offspring 1823   0   1      1  0  1   1   0          0.660 0.474 0.011
## 3    5 offspring 1608   0   1      1  0  1   1   0          0.652 0.477 0.012
## 4    6 offspring 1370   0   1      1  0  1   1   0          0.723 0.448 0.012
## 5    7 offspring 1170   0   1      1  0  1   1   0          0.746 0.435 0.013
## 6    8 offspring 1021   0   1      1  0  1   1   0          0.738 0.440 0.014
## 7    9 offspring  896   0   1      1  1  1   0   0          0.751 0.433 0.014
## 8   10 offspring  780   0   1      1  0  1   1   0          0.740 0.439 0.016
## 9   11 offspring  665   0   1      1  0  1   1   0          0.701 0.458 0.018
## 10  12 offspring  568   0   1      1  0  1   1   0          0.722 0.448 0.019
## 11  13 offspring  494   0   1      1  0  1   1   0          0.686 0.464 0.021
## 12  14 offspring  424   0   1      1  0  1   1   0          0.660 0.474 0.023
## 13  15 offspring  343   0   1      1  0  1   1   0          0.694 0.462 0.025
## 14  16 offspring  290   0   1      1  0  1   1   0          0.603 0.490 0.029
## 15  17 offspring  222   0   1      1  0  1   1   0          0.545 0.499 0.033
## 16  18 offspring  173   0   1      1  0  1   1   0          0.566 0.497 0.038
## 17  19 offspring  146   0   1      0  0  1   1   0          0.473 0.501 0.041
## 18  20 offspring  101   0   1      0  0  1   1   0          0.475 0.502 0.050
## 19  21 offspring   79   0   1      0  0  1   1   0          0.405 0.494 0.056
## 20  22 offspring   57   0   1      0  0  1   1   0          0.298 0.462 0.061
## 21  23 offspring   37   0   1      0  0  0   0   0          0.135 0.347 0.057
## 22  24 offspring   26   0   1      0  0  0   0   0          0.115 0.326 0.064
##       ci
## 1  0.019
## 2  0.022
## 3  0.023
## 4  0.024
## 5  0.025
## 6  0.027
## 7  0.028
## 8  0.031
```

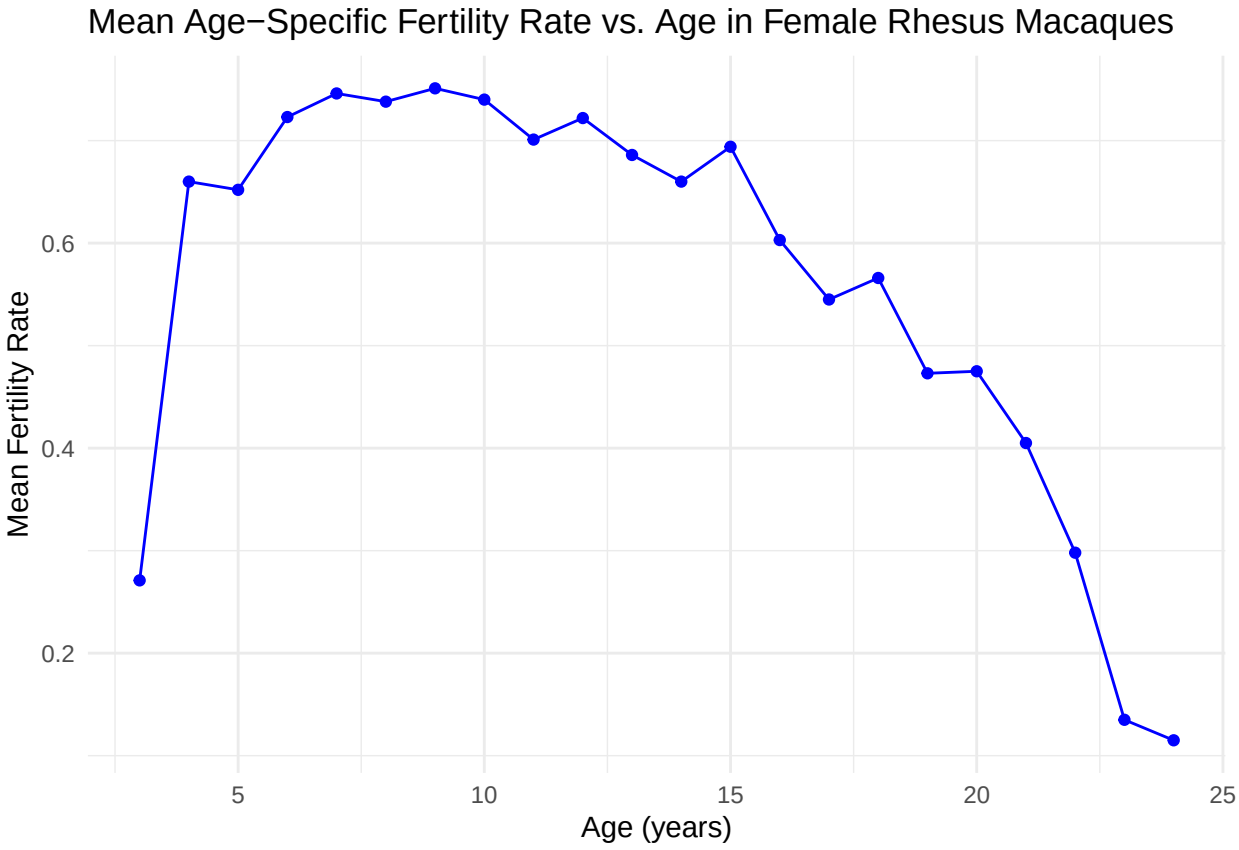
```
## 9  0.035
## 10 0.037
## 11 0.041
## 12 0.045
## 13 0.049
## 14 0.057
## 15 0.066
## 16 0.075
## 17 0.082
## 18 0.099
## 19 0.111
## 20 0.122
## 21 0.116
## 22 0.132
```

```r
nrow(repro_data)
```

```
## [1] 22
```

a

```r
ggplot(repro_data, aes(x = age, y = mean_fertility)) +
  geom_point(color = "blue") +
  geom_line(color = "blue") +   #
  labs(
    title = "Mean Age-Specific Fertility Rate vs. Age in Female Rhesus Macaques",
    x = "Age (years)",
    y = "Mean Fertility Rate"
  ) +
  theme_minimal()
```



Per the graph above; we see female rhesus macaques from Cayo Santiago exhibit a sharp rise in fertitility in youth (until around 5 years), stead growth and an ultimately decreasingly severe dropoff post around 10 years. We peak at around 70% fertility; the drop off is somewhat linear with age from around 10; although it better resembles some non linear function

b

```
fert_model <- lm(mean_fertility ~ age, data = repro_data)


summary(fert_model)
```

```
##
## Call:
## lm(formula = mean_fertility ~ age, data = repro_data)
##
## 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
```

```
intercept <- round(coef(fert_model)[1], 3)
grad <- round(coef(fert_model)[2], 3)

# answer to q


cat("The linear regression equation is: \nmean_fertility =", intercept, "+", grad, "* age")
```

```
## The linear regression equation is:
## mean_fertility = 0.822 + -0.019 * age
```

Per above; per each entire year of increase, the fertility rate decreases by -0.019 (overall) This is sensible as we expect older (living) things to be less fertile. Note that the intercept @ 0.822 indicates fertility in the group $<$ 1y old. 1y is a very broad time period, obviously new born (newly hatched?) should be at zero fertility. This error is due to the sampling and 'compression' into year groups. Finer units for age would provide a more useful interpretation. However, given the domain, this is not a significant factor to consider; as we are concerned with the overall trend. While the grouping causes some distortion in the early age categories, the overall negative relationship between age and fertility remains clear and biologically plausible.

C

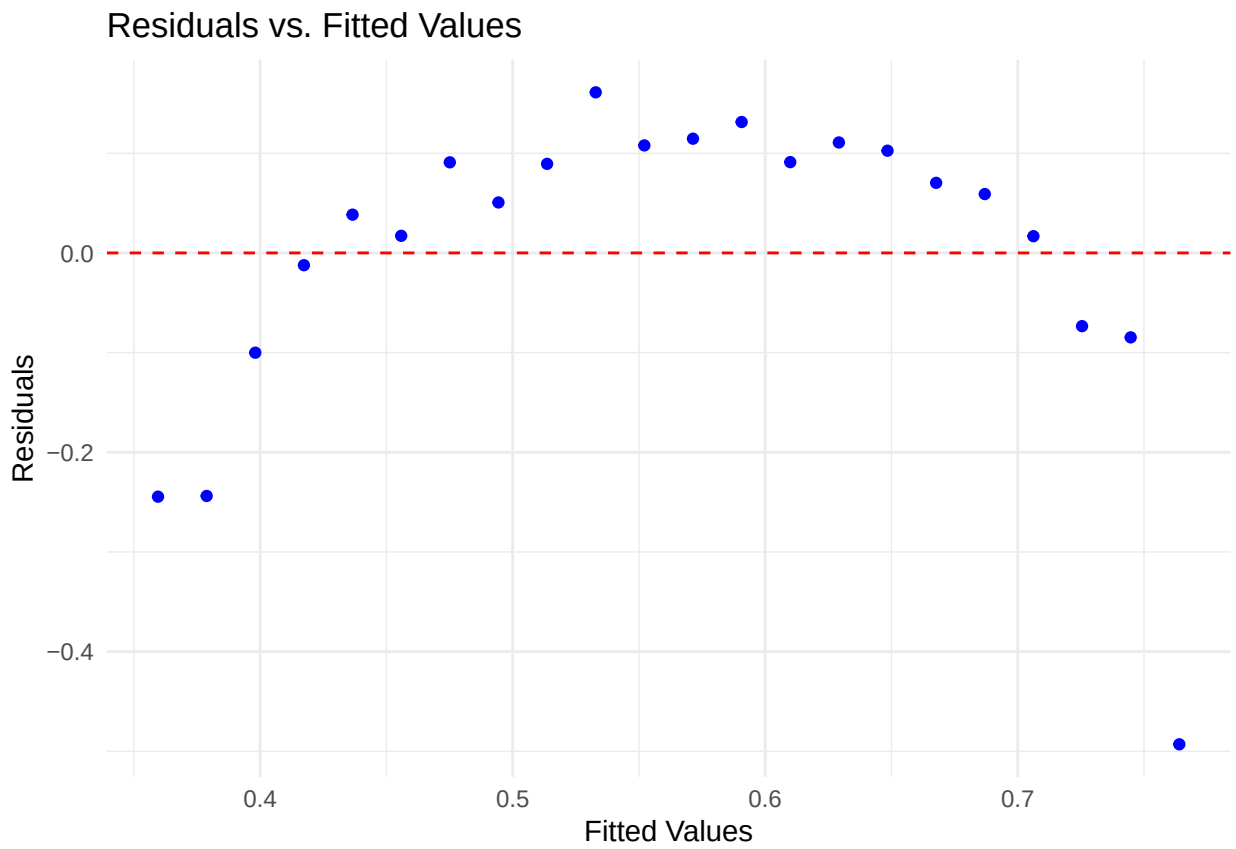
In order to determine the 'goodness of fit' we consider some of the residuals of our data; Note we're not using any kind of test / train set in this dataset. We consider residuals of all data.

```
comb_df <- repro_data
comb_df$diffs <- resid(fert_model)
comb_df$preds <- fitted(fert_model)
#just combining outputs and preds
```

c

As we did earlier; we graph the residuals per fitted values below.

```
ggplot(comb_df, aes(x = preds, y = diffs)) +
 geom_point(color = "blue") +  # Plot the residuals
 geom_hline(yintercept = 0, color = "red", linetype = "dashed") +  # Add a horizontal line at y=0
 labs(
   title = "Residuals vs. Fitted Values",
   x = "Fitted Values",
   y = "Residuals"
 ) +
 theme_minimal()
```



Unlike in our previous exploration; there is a *pattern* in our residual values; that is, they look eerily similar to the data points themselves This implies that the relationship between the predictor and response might be nonlinear, and a linear model may not be appropriate.

We therefore can state that a better model exists.

However; we do note that the dataset is very small. Small datasets can lead to variability in residuals and may not fully capture the complexity of the relationship between variables. We may wish to use some fucntion to transform the data to create a better fit

Any consistent pattern in residuals, as we see above as a curve indicates that the current model may not be capturing all relevant aspects of the data. The systematic pattern implies that the model is oversimplified; see our small data issue.

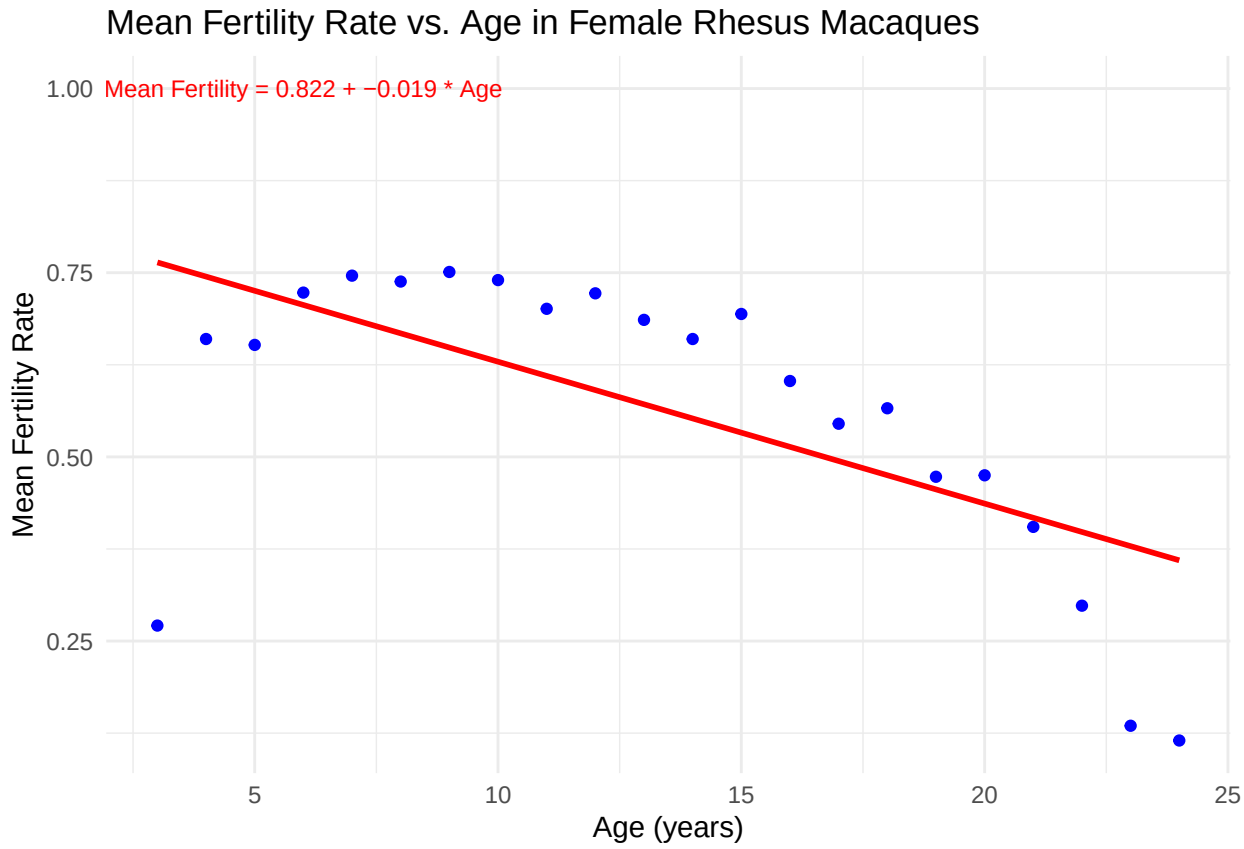
```
equation_text <- paste("Mean Fertility = ", intercept, " + ", grad, " * Age", sep = "")

# Plot with regression line and equation
ggplot(repro_data, aes(x = age, y = mean_fertility)) +
  geom_point(color = "blue") +  # Plot the data points
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # Add the regression line
  labs(
    title = "Mean Fertility Rate vs. Age in Female Rhesus Macaques",
    x = "Age (years)",
    y = "Mean Fertility Rate"
  ) +
```

```
    annotate("text", x = 6, y = 1, label = equation_text, size = 3, color = "red") + #
  theme_minimal()
```

d

```
## `geom_smooth()` using formula = 'y ~ x'
```



e

Earlier, we discussed how the datapoints are not particularly precise (i.e. years as units is quite large in the life of a macquae). We can, however, use our regression model to determine unseen values between these integer values..

The regression line slopes downward, indicating a negative association between age and mean fertility rate. As age increases, the mean fertility rate decreases. This suggests that older female rhesus macaques tend to have a lower fertility rate compared to younger macaques.

The negative slope shows a clear trend in the data, however; it does not well reflect the curved nature of the data (graph). It is important to note that the data points exhibit some scatter around the regression line. While the linear model captures the general trend, the presence of patterns in the residuals suggests that a more complex (non linear) model might better fit this data. For instance.

Here, the negative slope (-0.019) suggests that for each additional year of age, the mean fertility rate decreases by 0.019 units.

The same zero value as mentioned above applies.

However, as before; thee observed negative trend aligns with biological expectations that fertility may decline with age in female rhesus macaques.

We therefore consider some bounds as below for the data points 6.95,14.85 and 19.45

```
test_data <- data.frame(age = c(6.95, 14.85, 19.45))
predictions <- predict(fert_model, newdata = test_data, interval = "prediction")
print(predictions)
```

```
##         fit        lwr       upr
## 1 0.6879166 0.33836689 1.0374664
## 2 0.5357736 0.19371626 0.8778309
## 3 0.4471840 0.09899011 0.7953779
```

We expect 6.95,14.85 and 19.45

Age -> Mean Fertility

6.95 -> 0.688

14.85 -> 0.536

19.45 -> 0.448

with certainties that they are within the upper,lwr bounds as above.