INVESTIGATING FACTORS AFFECTING THE SMALLEST ANTRAL FOLLICLE COUNT IN ASSESSING FERTILITY

Aims

- To identify the relationship between *LowAFC* and *Age*.
- To determine the best single predictor of *LowAFC*.
- To figure out the best model for predicting *LowAFC*.

Background

The experiment was conducted by a medical doctor (Dr Priya Maseelall) and her team of researchers on 332 women who were having trouble getting pregnant. A key method used for fertility assessments is the smallest antral follicle count (*LowAFC*), obtained using noninvasive ultrasound. The collected data contains various features of fertility, with each being described as follows:

Variable	Description
Age	Age (in years)
LowAFC	Smallest antral follicle count
FSH	Maximum follicle stimulating hormone level
E2	Fertility level
MaxE2	Maximum fertility level
TotalGn	Total gonadotropin level
Embryos	Number of embryos

Findings

- There is a moderate, negative, linear relationship between the *LowAFC* and *Age* (*Figure 1, Figure 2, Table 1*).
- *TotalGn* is the best single predictor of the *LowAFC* (*Table 2, Figure 3).*

The best model for predicting LowAFC *(Table 3, Figure 4, Figure 5, Figure 6, Figure 7)*:

$$y_i = 1.928e{+}01 - 4.697e{-}01\, x_{i1} - 4.659e{-}02\, x_{i2} - 8.916e{-}04\, x_{i3} - 1.425e{-}03\, x_{i4} + 3.059e{-}02\, x_{i5} + 1.981e{-}04\, x_{i4}x_{i5}$$

where y_i is the value of the smallest antral follicle count (LowAFC) for the i^{th} woman,
x_{i1} is the value of maximum follicle-stimulating hormone level for the i^{th} woman,
x_{i2} is the value of the fertility level (E2) for the i^{th} woman,

x_{i3} is the value of the maximum fertility level (MaxE2) for the i^{th} woman,
x_{i4} is the value of the total gonadotropin level (TotalGn) for the i^{th} woman,
x_{i5} is the value of the number of embryos (Embryos) for the i^{th} woman.

- For a 35-year-old woman with the values of *FSH* = 0.5, *E2* = 48, *MaxE2* = 2527, *TotalGn* = 1388, and *Embryos* = 7, the *LowAFC* value is expected to fall within the range between 4.880 and 27.705 (*Table 4*).

Discussion

The fact that the proportion of variance explained for the LowAFC is relatively low might be attributed to the presence of outliers (*Figure 2*, *Figure 3*, *Figure 4*) and the lack of important data, which might exert undue influence on coefficient estimates, leading to models that might generalise to new data, significantly affecting the reliability of the results of the regression analyses. Confounding variables like *Age* might also lead to biased estimates and misinterpretations of the results. Therefore, it might be worth investigating the outliers, adding more explanatory variables to the initial dataset, and removing confounding variables from the dataset to enhance the generalizability and reliability of the results.

Statistical methods and results

Computational methods

Statistical analyses were performed using *R* and *RStudio* version 4.3.1 *"Beagle Scouts"*.

Results

Since our predictors are on different scales, we need to standardise the dataset to make it easier to interpret the coefficients of the linear models by using:

```
> # Standardization
> afc.std <- afc %>% mutate_all(~(scale(.) %>% as.vector))
```

Initial inspection of the data showed that there is a moderate, negative association between the age (in years) and the smallest antral follicle count of women with fertility issues. It is possible that the relationship is linear (*Figure 1*).

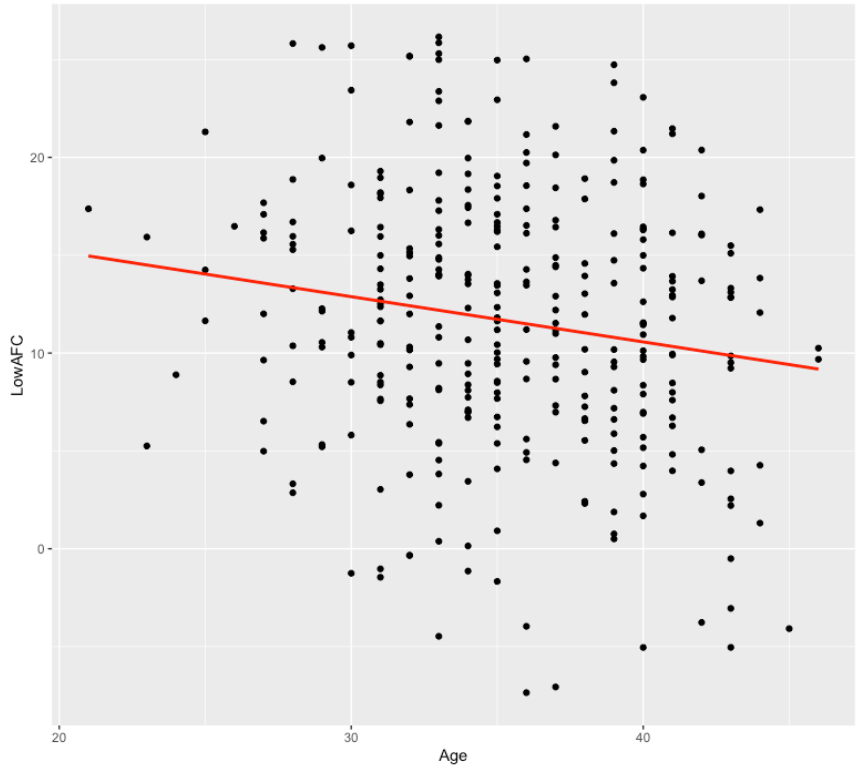

Figure 1. *Scatterplot showing a moderate negative linear relationship between LowAFC and Age.*

A simple linear regression analysis was then performed to quantify the linear relationship between *Age* and *LowAFC*. First, assumptions for a simple linear regression model were checked by evaluating the diagnostic plots:

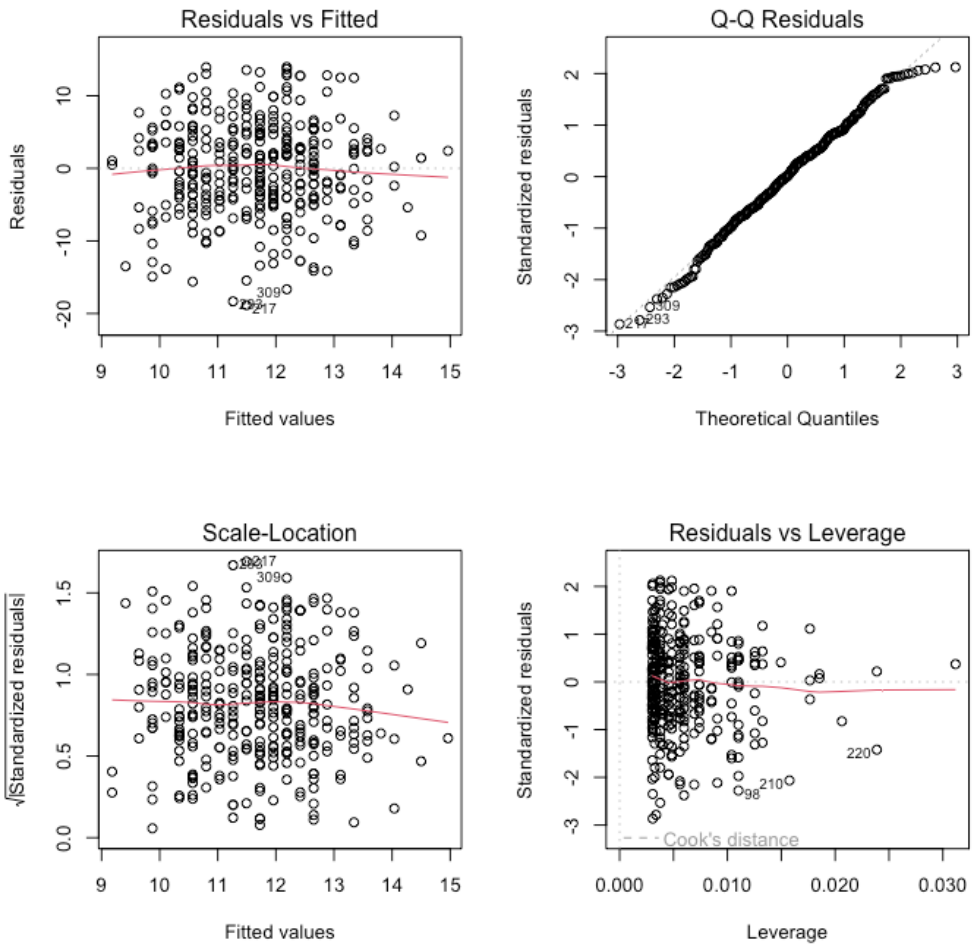

Figure 2. *Diagnostic plots for simple linear regression of LowAFC ~ Age showing that the assumptions about linearity, homoscedasticity, and normality of errors are met.*

The *Residuals vs. Fitted* plot in *Figure 1* shows that there are random scatters around the zero line, and the trend line is roughly flat, indicating that assumptions about the linearity and homoscedasticity have been met. The assumption about the normality of errors also appears reasonable by looking at the *Q-Q Residuals* plot, which shows points lying roughly on a straight line. The assumption about

independence is also reasonable since the age is independent among the women in the experiment. Besides, it is clear from the Residual *vs Leverage* plot that there are some outliers in our data.

Since the assumptions have been met, we define the null and alternative hypotheses:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0,$$

where β_1 is the estimate of the age (in years) of women having trouble getting pregnant.

```
Call:
lm(formula = LowAFC ~ Age, data = afc)

Residuals:
     Min       1Q   Median       3Q      Max
-18.8488  -4.2342   0.0838   4.5001  13.9702

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.81845    2.74827   7.211 3.81e-12 ***
Age         -0.23124    0.07708  -3.000   0.0029 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.59 on 330 degrees of freedom
Multiple R-squared:  0.02655,   Adjusted R-squared:  0.0236
F-statistic: 9.001 on 1 and 330 DF,  p-value: 0.002904
```

Table 1. *R output for simple linear regression analysis for LowAFC and Age, showing that the negative linear relationship between the two variables is statistically significant.*

From *Table 1*, we can obtain the line of best fit for this model:

$$\textit{LowAFC} = 19.818 - 0.231\ (\textit{Age})$$

From the simple linear regression model, the *t*-statistic is -3.000 on 330 degrees of freedom, and the *p*-value is 0.0029. Since the *p*-value is less than $\alpha = 0.05$, we can reject the null hypothesis at the 5% significance level and conclude that there is a statistically significant linear relationship between *Age* and *LowAFC*. On average, we would expect older women with fertility issues to have lower antral follicle count. Specifically, for every extra year in age, we would expect, on average, to see a decrease of approximately 0.231 units in the smallest antral follicle count.

To identify the best single predictor of *LowAFC,* similar simple linear regression analyses were performed on other variables, including *FSH*, *E2*, *MaxE2*, *TotalGn*, and *Embryos*. The test statistics and estimates of regression coefficients of these tests are summarised in *Table 2* below:

	Estimate		***R-squared***	***P-value***	***Residuals***
	Intercept	***Coefficient***			
Age	-1.738e-16	-1.629e-01	0.02655	0.0029	0.9881
FSH	3.692e-16	-3.279e-01	0.1075	9.2e-10	0.9461
E2	-1.058e-16	-1.174e-01	0.01377	0.0325	0.9946
MaxE2	-1.214e-16	2.681e-01	0.07189	7.12e-07	0.9648
TotalGn	-6.943e-17	-4.089e-01	0.1672	8.17e-15	0.914
Embryos	-1.998e-16	3.554e-01	0.1263	2.55e-11	0.9361

Table 2. *Summary of regression coefficients and test statistics of simple linear regression models for LowAFC and predictors (Age, FSH, E2, MaxE2, TotalGn, Embryos) showing that TotalGn has the largest r-squared, p-value and residuals, indicating that it is the best single predictor of LowAFC.*

It can be seen from *Table 1* that *TotalGn* is the variable with the highest r-squared (0.1672), indicating that it best explains the variability in *LowAFC*; the highest *p*-value (8.17e-15), showing that the linear relationship between *LowAFC* and *TotalGn* is the most statistically significant, and the smallest residual value (0.914), which means the differences between the observed values and the values predicted by the model is the smallest. The magnitude of the coefficient is also the largest (4.089e-01). All indicates that the linear relationship between *LowAFC* and *TotalGn* is the strongest, meaning that *TotalGn* is the variable that best predicts the smallest antral follicle count (*LowAFC*).

The assumptions for the simple linear regression for LowAFC ~ TotalGn can also be checked using the diagnostic plots:

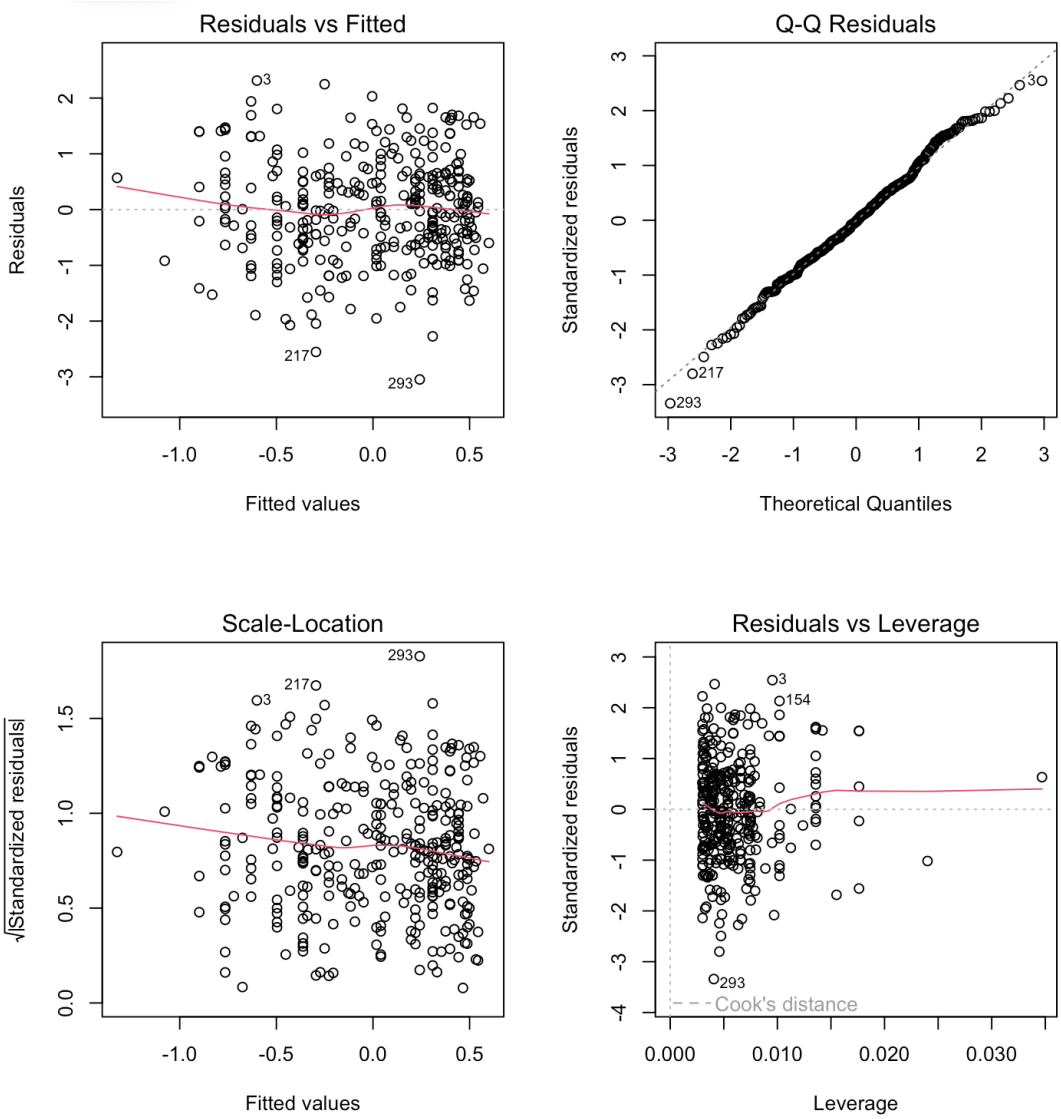

Figure 3. *Diagnostic plots for simple linear regression of LowAFC ~ TotalGn showing that the assumptions about linearity, homoscedasticity, and normality of errors are met.*

The null and alternative hypotheses are:

$$H_0: \ \beta_1 = 0$$

$$H_a: \ \beta_1 \neq 0,$$

where β_1 is the estimate of the total gonadotropin level of women facing fertility issues.

From *Table 2*, we can obtain the line of best fit for the model for *LowAFC ~ TotalGn*:

LowAFC = – 6.943e-17 – 4.089e-01 (TotalGn)

The *p*-value for this model is 8.17e-15 < α = 0.05. Hence, we can reject the null hypothesis at the 5% significance level and conclude that there is a statistically significant linear relationship between *LowAFC* and *TotalGn*. On average, we would expect women with higher total gonadotropin levels to have lower antral follicle count. Specifically, for every increased unit in the total gonadotropin level, we expect, on average, to see a decrease of 4.089e-01 units in the smallest antral follicle count.

To find out the best model for predicting *LowAFC,* a multiple linear regression analysis was performed to investigate how multiple independent variables are simultaneously related to a dependent variable (*LowAFC*).

A stepwise regression is used to obtain the best model:

```
> #best model
> best_model <- step(afc.lm2,direction='backward',trace=0)
> summary(best_model)
```

```
Call:
lm(formula = LowAFC ~ FSH + E2 + MaxE2 + TotalGn + Embryos +
    MaxE2:Embryos, data = afc)

Residuals:
     Min       1Q   Median       3Q      Max
-17.8967  -3.8325  -0.1604   4.1832  13.8957

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.928e+01  2.140e+00   9.009  < 2e-16 ***
FSH           -4.697e-01  1.882e-01  -2.495   0.0131 *
E2            -4.659e-02  2.082e-02  -2.238   0.0259 *
MaxE2         -8.916e-04  8.263e-04  -1.079   0.2814
TotalGn       -1.425e-03  2.657e-04  -5.362 1.56e-07 ***
Embryos        3.059e-02  1.885e-01   0.162   0.8712
MaxE2:Embryos  1.981e-04  9.408e-05   2.106   0.0360 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.708 on 325 degrees of freedom
Multiple R-squared:  0.2805,    Adjusted R-squared:  0.2672
F-statistic: 21.12 on 6 and 325 DF,  p-value: < 2.2e-16
```

Table 3*. R output of the stepwise regression showing the best multiple linear regression model for predicting LowAFC*

From the estimated coefficients obtained from *Table 3*, the multiple linear regression model is:

$$y_i = 1.928e{+}01 - 4.697e{-}01\, x_{i1} - 4.659e{-}02\, x_{i2} - 8.916e{-}04\, x_{i3} - 1.425e{-}03\, x_{i4} + 3.059e{-}02\, x_{i5} + 1.981e{-}04\, x_{i4}x_{i5}$$

where y_i is the value of the smallest antral follicle count (LowAFC) for the i^{th} woman,

x_{i1} is the value of maximum follicle-stimulating hormone level for the i^{th} woman,

x_{i2} is the value of the fertility level (E2) for the i^{th} woman,

x_{i3} is the value of the maximum fertility level (MaxE2) for the i^{th} woman,

x_{i4} is the value of the total gonadotropin level (TotalGn) for the i^{th} woman,

x_{i5} is the value of the number of embryos (Embryos) for the i^{th} woman.

The F-statistic from the multiple linear regression model is 21.12 with a *p*-value of < 2.2e-16. Since the *p*-value is less than 0.05, the data provide sufficient evidence to conclude that our regression model fits the data better than the model with no independent variables, indicating that the overall model is statistically significant for predicting *LowAFC*. The model has an adjusted R-squared value of 0.2672 (*Table 3*), indicating that 26.72% of the variation in the dependent variable (*LowAFC*) is explained by the independent variables (*FSH, E2, MaxE2, TotalGn, Embryos, MaxE2: Embryos*).

The assumptions of the best model were checked by evaluating the diagnostic plots:

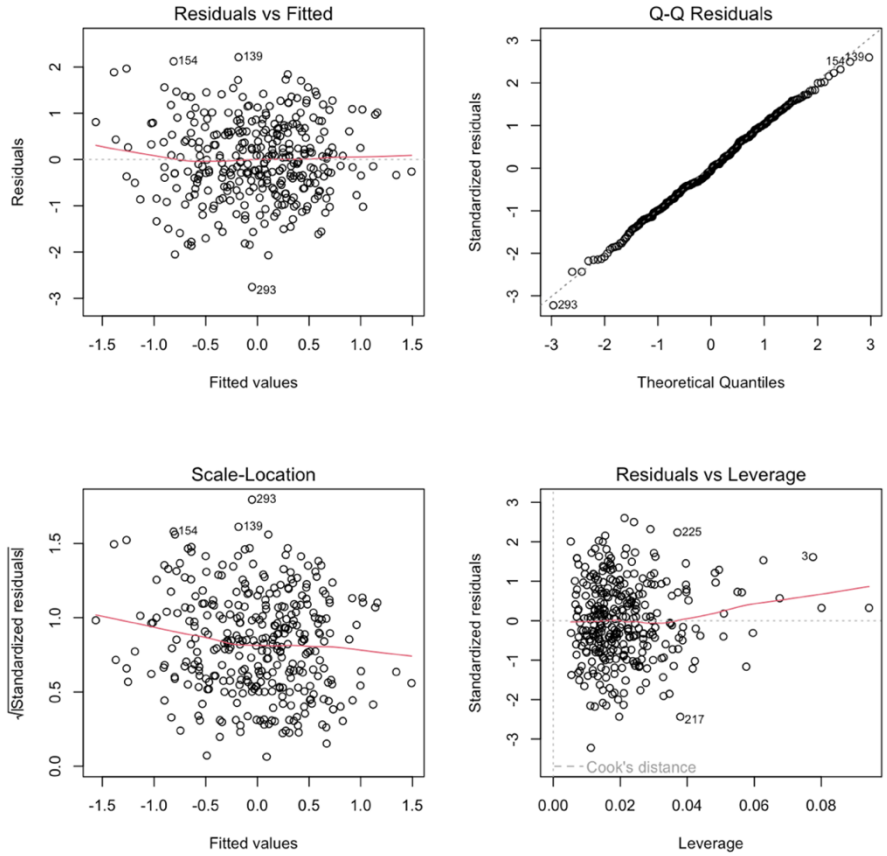

Figure 4. *Multiple linear regression diagnostic plots for LowAFC ~ predictors showing that the assumptions of linearity, homoscedasticity, and normality of errors have been met.*

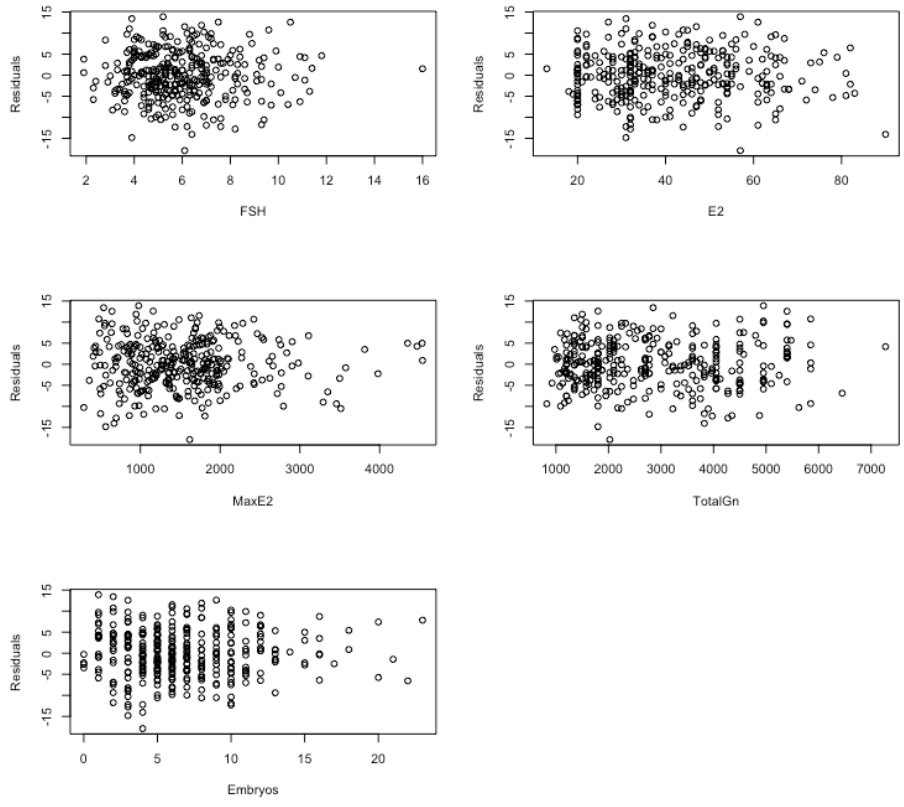

Figure 5. *Residual vs. Predictors plots showing that the assumption of linearity has been met.*

The *Residuals vs. Fitted* plot in *Figure 4* and *Figure 5* show that there are random scatters around the zero line, and the trend line is roughly flat, indicating that the assumptions about the linearity and constant variance have been met. The assumption about the normality of errors also appears reasonable by looking at the *Q-Q Residuals* plot, which shows points lying roughly on a straight line. Besides, it is clear from the *Residual vs Leverage* plot that there are some outliers in our data. The Residuals vs. Predictors plots show no signs of non-linearity so polynomial terms are not warranted.

The inclusion of interaction terms in the model can be checked by looking at the matrix of scatterplots demonstrating the pairwise interactions of the variables:

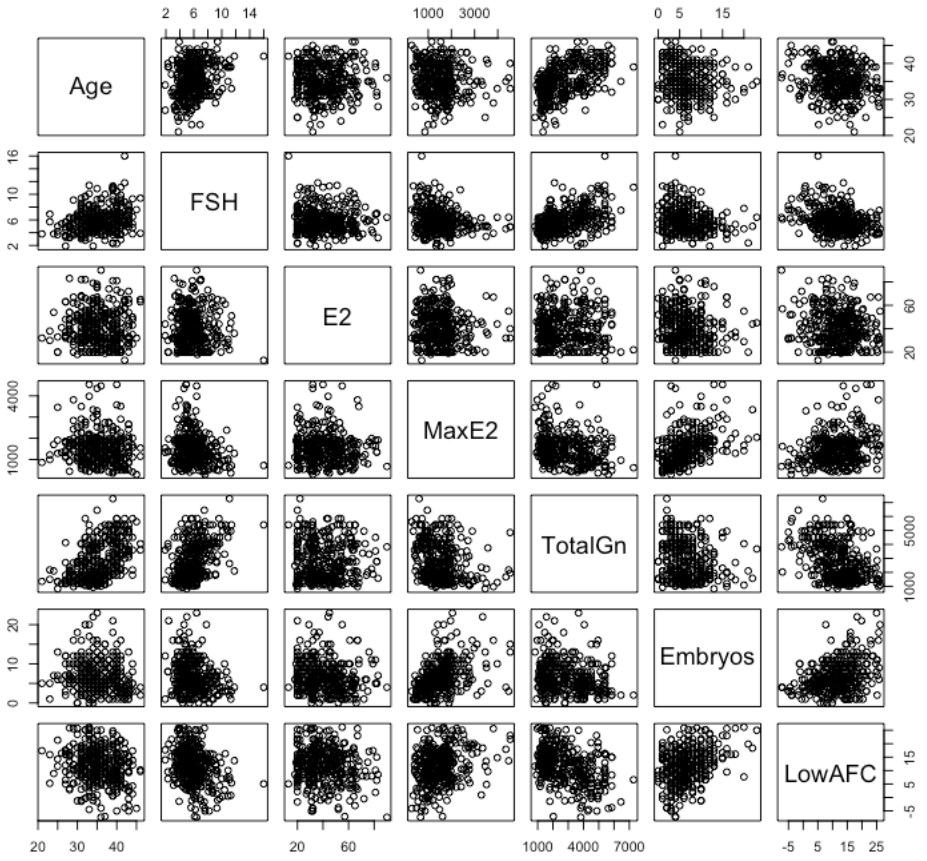

Figure 6. *Matrix of scatterplots to demonstrate the pairwise interactions between LowAFC and predictors showing that there seems to be an interaction between MaxE2 and Embryos.*

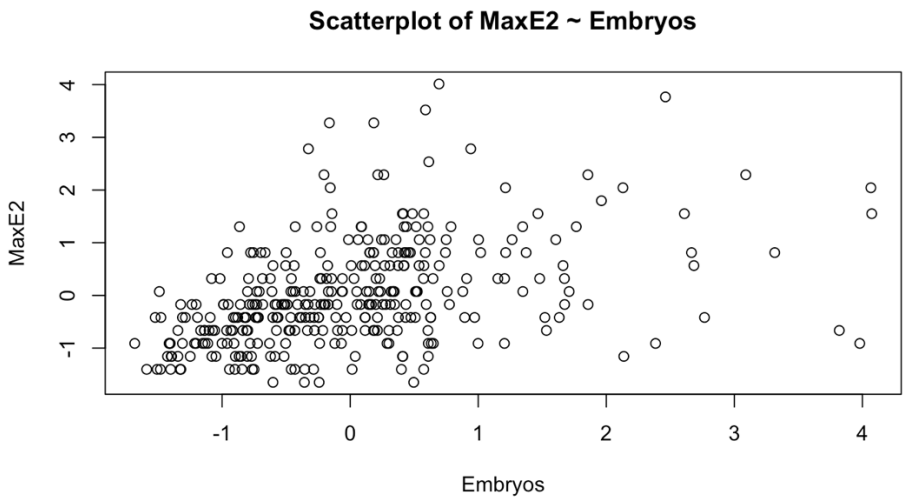

Figure 7. *Scatterplot showing a moderate, positive relationship between MaxE2 and Embryos.*

From *Figure 6*, there seems to be an interaction between the *MaxE2* and *Embryos* variables, with the *MaxE2* increasing when the *Embryos* increases and vice versa. This interaction is then checked using the scatterplot in *Figure 7*, which shows a moderate, positive relationship between *MaxE2* and *Embryos*. Because of this interaction, *MaxE2* and *Embryos* are still included in our best model, although their *p*-values are 0.367 and 0.8712 > 0.05. The interaction term *MaxE2: Embryos* in the model captures the joint effects of the independent variables (*MaxE2* and *Embryos*) on the dependent variable (*LowAFC)* over their individual effects.

We consider the following data to check the generalizability of your model:

Age	35
FSH	0.5
E2	48
MaxE2	2527
TotalGn	1388
Embryos	7

This woman has a LowAFC of 6. We can check if the model would categorise this woman as having an extreme LowAFC value or not by using:

```
+ newdata <- data.frame(FSH=0.5,E2=48,MaxE2=2527,TotalGn=1388,Embryos=7)
> predict(best_model, newdata, interval="prediction")
       fit      lwr      upr
1 16.29281 4.880363 27.70525
```

Fit	***Lower bound***	***Upper bound***
16.293	4.880	27.705

Table 4. *R output for the 95% prediction interval for LowAFC for a 35-year-old woman with the values of FSH = 0.5, E2 = 48, MaxE2 = 2527, TotalGn = 1388, and Embryos = 7.*

From *Table 4*, the prediction interval obtained for the given data is (4.880, 27.705), meaning that for that prediction interval, we can be 95% confident that our observation will fall within the range between 4.880 and 27.705. The *LowAFC* value of the woman is 6 which falls within the prediction interval. Therefore, the woman does not have an extreme LowAFC value.