

Homework 6

Denis Ostroushko

2022-10-21

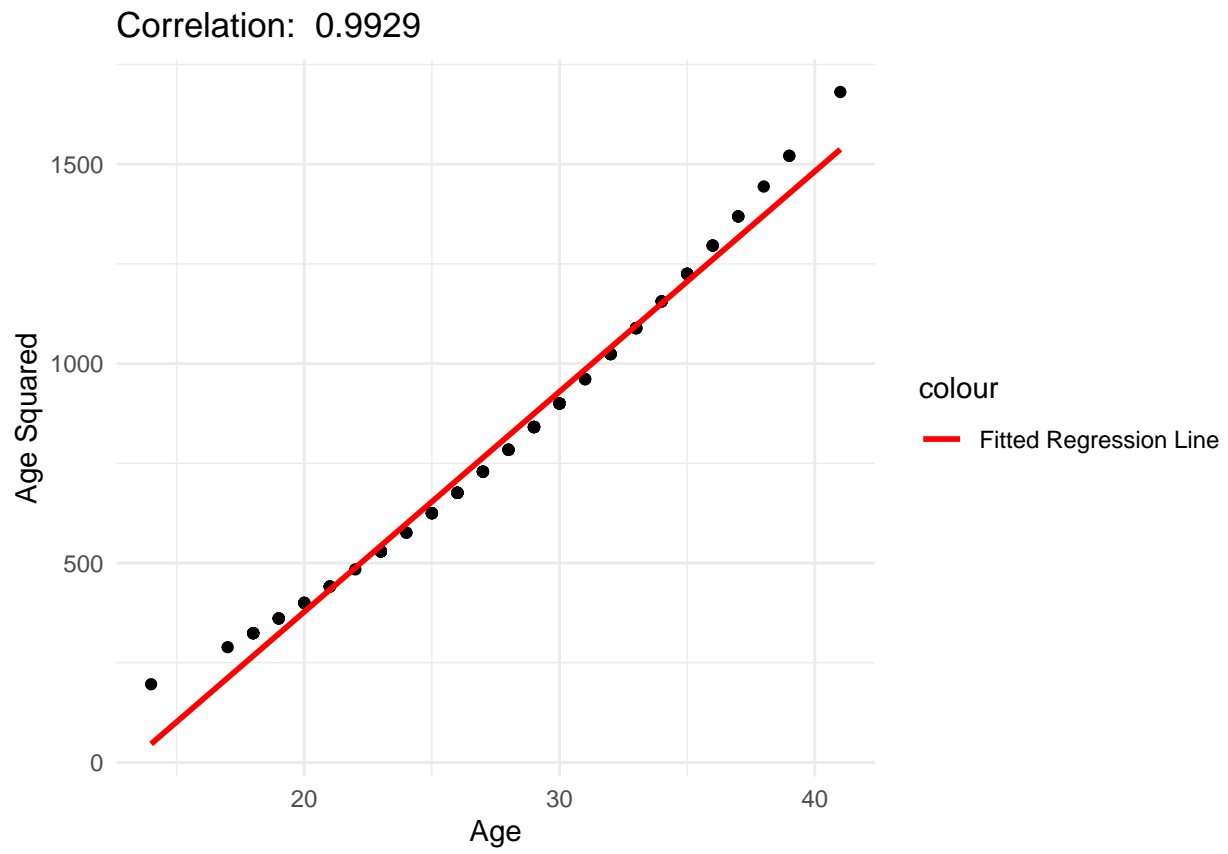
```
library(tidyverse)
library(kableExtra)
library(readxl)
library(gridExtra)
library(ggeffects)
library(mltools) # one hot encoding outside of caret package
```

12.2

```
infants <- readxl::read_xls('/Users/denisostroushko/Desktop/UofM MS/MS Fall 2022/Puhb 7405/Data Sets/Infants.xlsx')
colnames(infants) <- c("head_c", "length", "gest_weeks", "birth_w", "m_age", "toxemia")

# process the data and keep variables for analysis

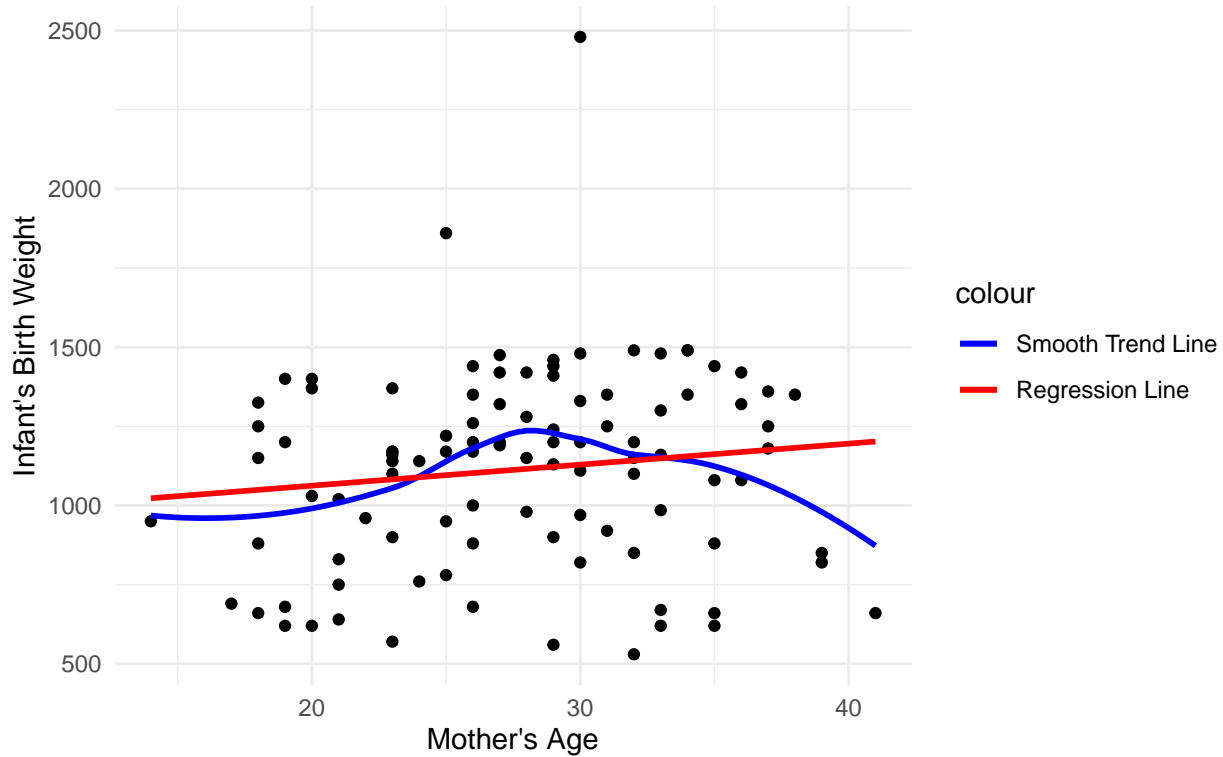
infants_f <- infants %>%
  select(birth_w, gest_weeks, m_age)
```



12.2 - A

Model Specifications and T-tests

Correlation Between Mother's Age
and Infant's Birth Weight: 0.1263



Model specification:

$$E[Y] = \hat{\beta}_0 + \hat{\beta}_1 * \text{Gestational Weeks} + \hat{\beta}_2 * \text{Mother's Age} + \hat{\beta}_3 * \text{Mother's Age}^2$$

Model Summary

Model Term	Estimate	Std. Error	T-value	P-value
Intercept	-1442.928	496.023	-2.909	0.005
Gestational Weeks	75.667	10.652	7.103	0.000
Mother's Age	30.252	36.813	0.822	0.413
Mother's Age Squared	-0.582	0.656	-0.887	0.377

Comments on Model summary:

- R and Adjusted R: 0.3904 0.3714
- Coefficients for Age and Age ²

Evaluate Extra Sum of Squares

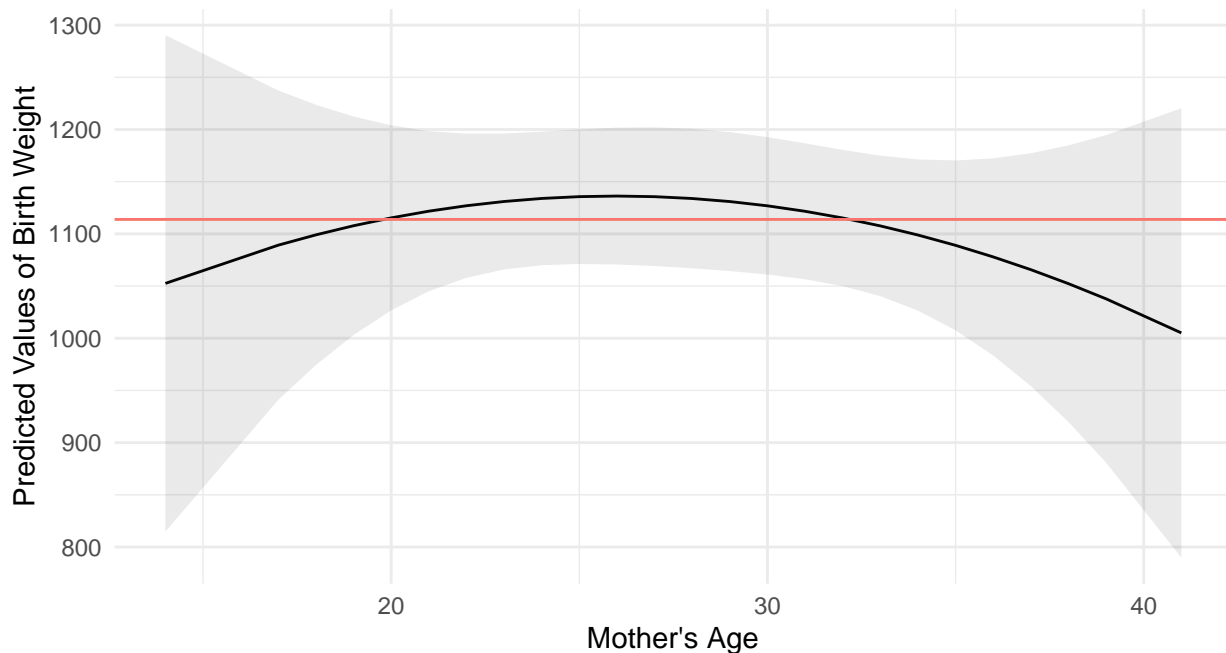
Focus: Evaluate SSR(Age² | Gest, Age)

Model Term	DF	SS	MS	F-statistic	P(F* > F)
Gestational Weeks	1	3755985.30	3755985.30	60.4451134	0.0000
Mother's Age	1	15505.20	15505.20	0.2495254	0.6186
Mother's Age Squared	1	48879.84	48879.84	0.7866239	0.3773
Residuals	96	5965322.40	62138.78	NA	NA

- Extra SS
- Extra R^2
- Connection with the t-test

Visualize Model Effects

Model Estimated Effects of Mother's Age on Infant's Birth Weight



Additional Elements: — Birth Weight Mean Value: 1114

- Comment on Standard Error and fit, we can fit a line with slope = +

Interpretation of Mother's Age Coefficients

From google, interpretation of the quadratic coefficient:

” A positive quadratic coefficient causes the ends of the parabola to point upward. A negative quadratic coefficient causes the ends of the parabola to point downward. The greater the quadratic coefficient, the narrower the parabola. The lesser the quadratic coefficient, the wider the parabola.”

<https://stats.stackexchange.com/questions/108657/how-to-interpret-coefficients-of-x-and-x2-in-same-regression>

It may be useful to describe the effect of a unit change at some low value, some high value and somewhere in between.

12.2 - B

Correlation Transformation for variables Y, X_1, \dots, X_{p-1} , denoted by V :

$$V^* = \frac{1}{\sqrt{n-1}} \times \left(\frac{V - \bar{V}}{sd(V)} \right)$$

```
correlation_transformation <-  
  function(X, n = nrow(infants_f_cor_tr)){  
    1/(sqrt(n - 1)) * (X - mean(X))/sd(X)  
  }  
  
infants_f$m_age_sq <- infants_f$m_age^2  
infants_f_cor_tr <- infants_f  
  
infants_f_cor_tr <- data.frame(lapply(infants_f_cor_tr, correlation_transformation))
```

Table 1: Original Scale Regression Estimates

Model Term	Estimate	Std. Error	T-value	P-value
Intercept	-1442.928	496.023	-2.909	0.005
Gestational Weeks	75.667	10.652	7.103	0.000
Mother's Age	30.252	36.813	0.822	0.413
Mother's Age Squared	-0.582	0.656	-0.887	0.377

Table 2: Correlation Transformation Regression Estimates

Model Term	Estimate	Std. Error	T-value	P-value
Intercept	0.000	0.008	0.000	1.000
Gestational Weeks	0.610	0.086	7.103	0.000
Mother's Age	0.576	0.701	0.822	0.413
Mother's Age Squared	-0.616	0.694	-0.887	0.377

- intercept is zero as expected in corr transformed
- P-values are different for m age
- Same conclusions apply

12.2 - C

Transformation back to the original scale:

For variables X_1, \dots, X_{p-1} :

$$\hat{\beta}_i = \hat{\beta}_i^* \times \frac{sd(Y)}{sd(X_i)}$$

Table 3: Original Model Estimates and C.I.

Model Term	Coefficient	95% C.I. Lower Bound	95% C.I. Upper Bound
Gestation Weeks	75.667	54.522	96.811
Mother's Age	30.252	-42.821	103.324
Mother's Age Squared	-0.582	-1.884	0.721

Table 4: Estimates obtained via Back-Transformation and C.I.

Model Term	Coefficient	95% C.I. Lower Bound	95% C.I. Upper Bound
Gestation Weeks	75.678	54.522	96.811
Mother's Age	30.268	-42.821	103.324
Mother's Age Squared	-0.582	-1.884	0.721

```
transform_back <-
  function(Beta_star, s_x, s_y){
    Beta_star * (s_y / s_x)
  }

S_Y <- sd(infants_f$birth_w)
```

Hid code to prepare the table.

recall the the original model with the transformed variables was called `inf_lm`. Used it for Extra SS, t-tests and model effects. We can obtain standard errors and confidence intervals for the estimates to compare with the transformation back from the correlation transformation procedure.

```
conf <- data.frame(confint(inf_lm)) # just the confidence intervals
conf <- cbind(coefficients(inf_lm), conf )
```

so we can use linear transformations good to know

13.4

```
cig <- read_xlsx('/Users/denisostroushko/Desktop/UofM MS/MS Fall 2022/Puhs 7405/Data Sets/E-CID-3.xlsx')

cig$Y1 <- with(cig, log(NNAL_vt4_creat / NNAL_vt0_creat))
cig$Y2 <- with(cig, log(TNE_vt4_creat / TNE_vt0_creat))

cig <- cig %>%
  select(Y1, Y2, arm, age, gender, white, educ2, income30, FTND)

colnames(cig)[length(cig)] <- "ftnd"

n_unique <- function(x){length(unique(x))}

meta_data <-

  data.frame(
    class = sapply(cig, class),
    n_unique = sapply(cig, n_unique)
```

)

Table 5: Sumamry of Covariates

Variables	Assigned Class	N of Unique Values
arm	numeric	4
age	numeric	51
gender	numeric	2
white	numeric	2
educ2	numeric	2
income30	numeric	2
ftnd	numeric	8

- Arm will result in 4 variables
- Age is untouched
- FTND is treated as continuous
- Others need to be converted to factor variables

```
cig$arm <- as.factor(cig$arm)
cig <- one_hot(as.data.table(cig))
length(cig[,3:length(cig)])
## [1] 10
```

13.4 - A

Encoding Variables

13.4 - B

Regression on Y1

Regression on Y2

13.4 - C

Step Wise Regression on Y1

Step Wise Regression on Y1

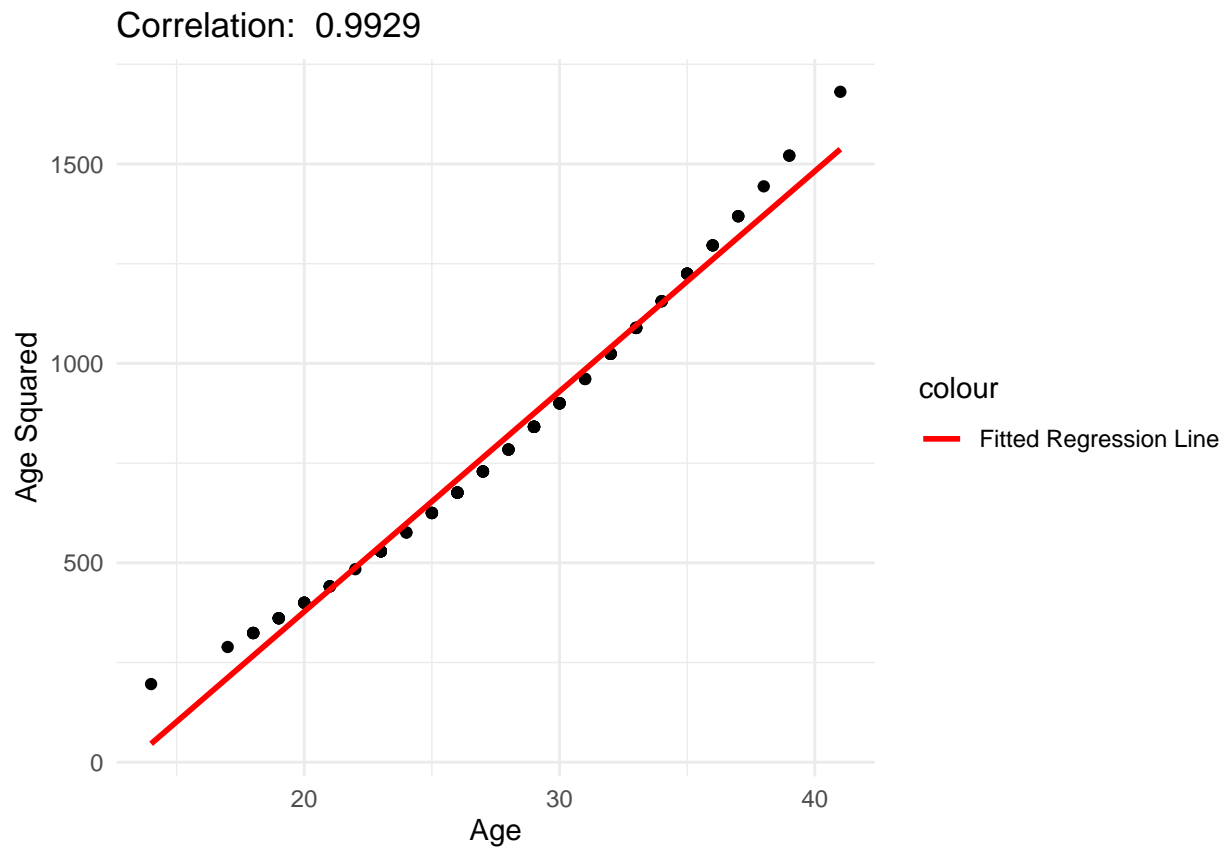
Appendix: 12.2

```
# look at the correlation between age and age^2
ggplot(data = infants_f,
       aes(x = m_age,
           y = m_age^2)) + geom_point() +

  stat_smooth(method = "lm", se = F, aes(color = "Fitted Regression Line")) +
  scale_color_manual(values = c("Fitted Regression Line" = "red")) +

  xlab("Age") +
```

```
ylab("Age Squared") +
ggtitle(paste("Correlation: ", round(cor(infants_f$m_age, infants_f$m_age^2),4))) +
theme_minimal()
```



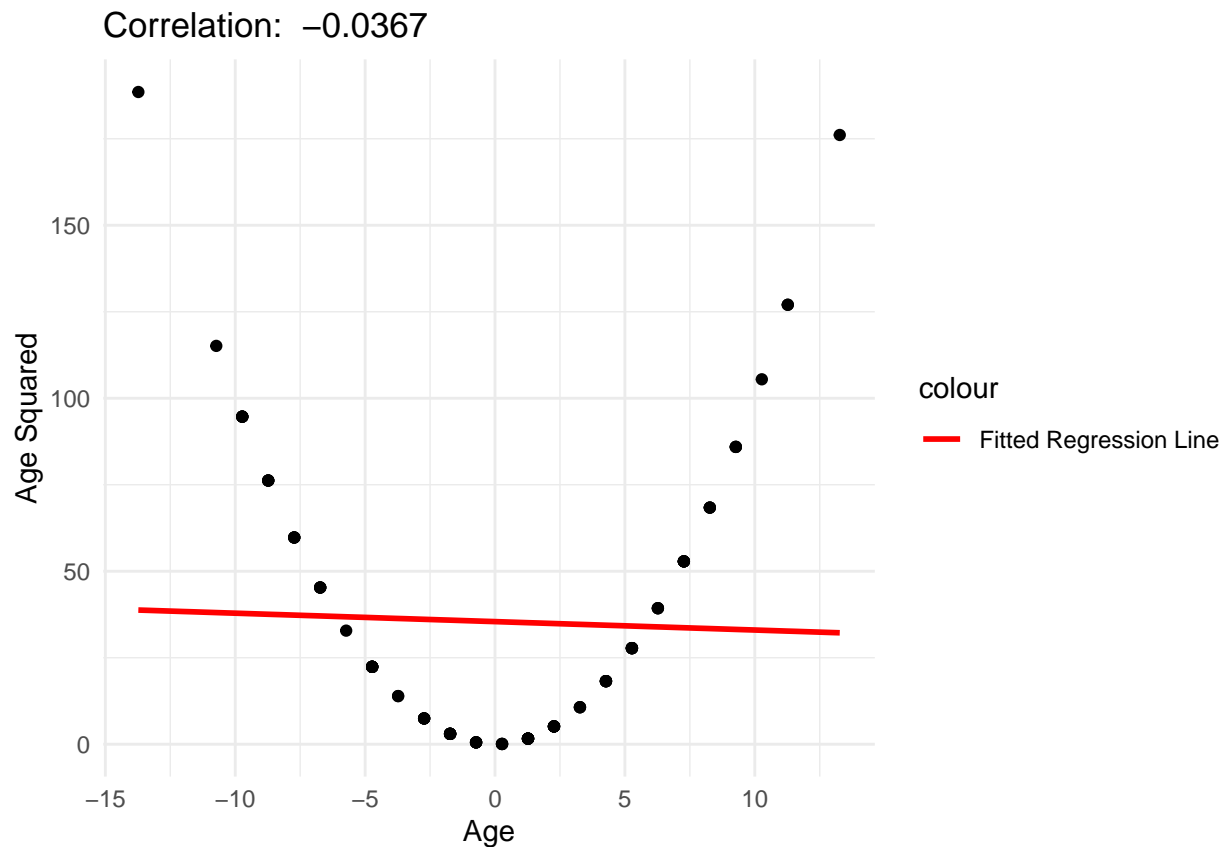
now apply centering:

```
infants_f$m_age_centered <- with(infants_f, m_age - mean(m_age))
infants_f$gest_weeks_centered <- with(infants_f, gest_weeks - mean(gest_weeks))

ggplot(data = infants_f,
       aes(x = m_age_centered,
           y = m_age_centered^2 )) + geom_point() +

  stat_smooth(method = "lm", se = F, aes(color = "Fitted Regression Line")) +
  scale_color_manual(values = c("Fitted Regression Line" = "red")) +

  xlab("Age") +
  ylab("Age Squared") +
  ggtitle(paste("Correlation: ", round(cor(infants_f$m_age_centered, infants_f$m_age_centered^2),4))) +
  theme_minimal()
```

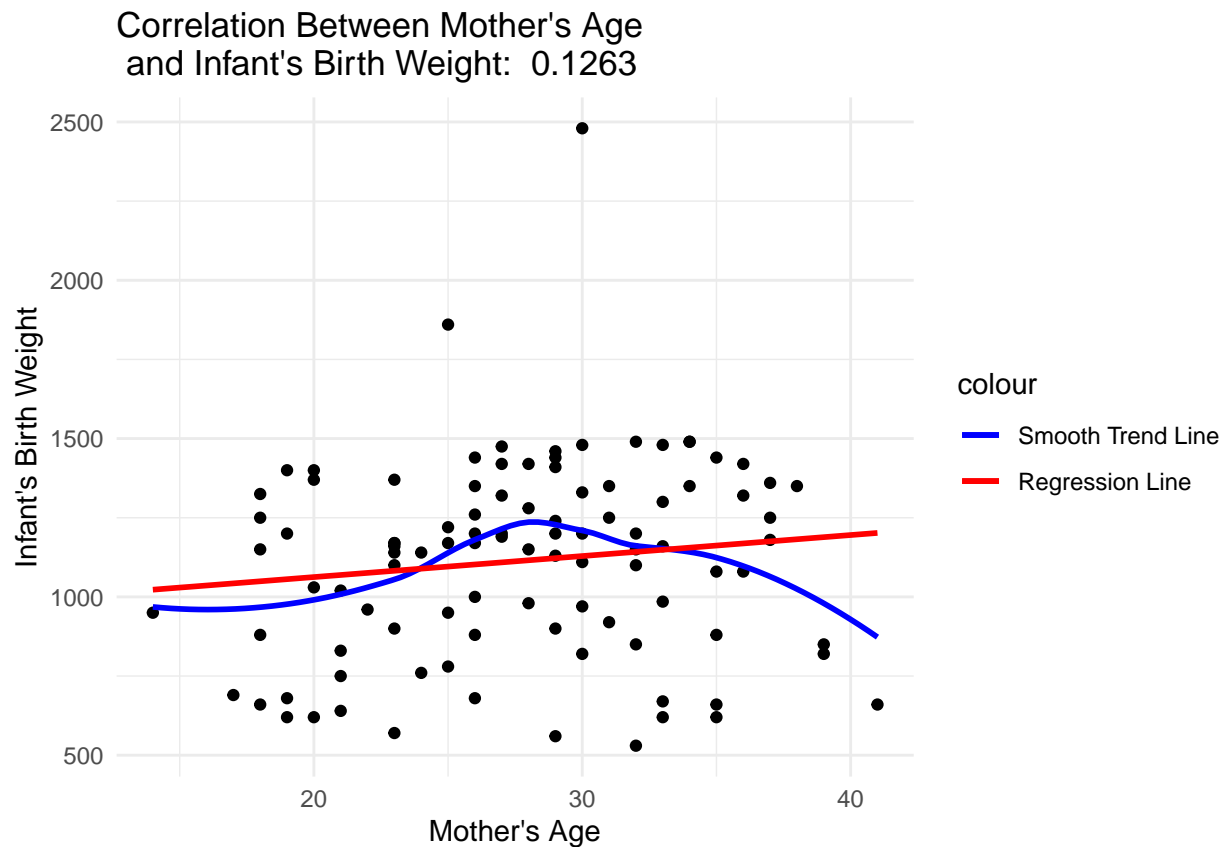



```
ggplot(data = infants_f,
       aes(x = m_age,
           y = birth_w)) + geom_point() +

  stat_smooth(se = F, aes(color = "Smooth Trend Line")) +
  stat_smooth(se = F, method = "lm", aes(color = "Regression Line")) +

  scale_color_manual(values = c("Smooth Trend Line" = "blue",
                                "Regression Line" = "red")) +

  xlab("Mother's Age") +
  ylab("Infant's Birth Weight") +
  ggtitle(paste("Correlation Between Mother's Age \n and Infant's Birth Weight: ",
                round(cor(infants_f$m_age,
                          infants_f$birth_w), 4))) +
  theme_minimal()
```



```
ggplot(data = infants_f,
       aes(x = m_age_centered,
           y = birth_w)) + geom_point() +

  stat_smooth(se = F, aes(color = "Smooth Trend Line")) +
  stat_smooth(se = F, method = "lm", aes(color = "Regression Line")) +

  scale_color_manual(values = c("Smooth Trend Line" = "blue",
                                "Regression Line" = "red")) +

  xlab("Mother's Age") +
  ylab("Infant's Birth Weight") +
  ggtitle(paste("Correlation Between Centered Mother's Age \n and Infant's Birth Weight: ",
                round(cor(infants_f$m_age_centered,
                          infants_f$birth_w), 4))) +
  theme_minimal()
```

Correlation Between Centered Mother's Age
and Infant's Birth Weight: 0.1263

