Homework 6

Denis Ostroushko

2022-10-24

```
library(MASS)
require(tidyverse) # require instead of library to make sure that other packages do not overwrite tidyv
library(kableExtra)
library(readxl)
library(gridExtra)
library(ggeffects)
library(mltools) # one hot encoding outside of caret package
library(data.table) # need this for mltools to work
library(olsrr) # a better package for stepwise regression
```

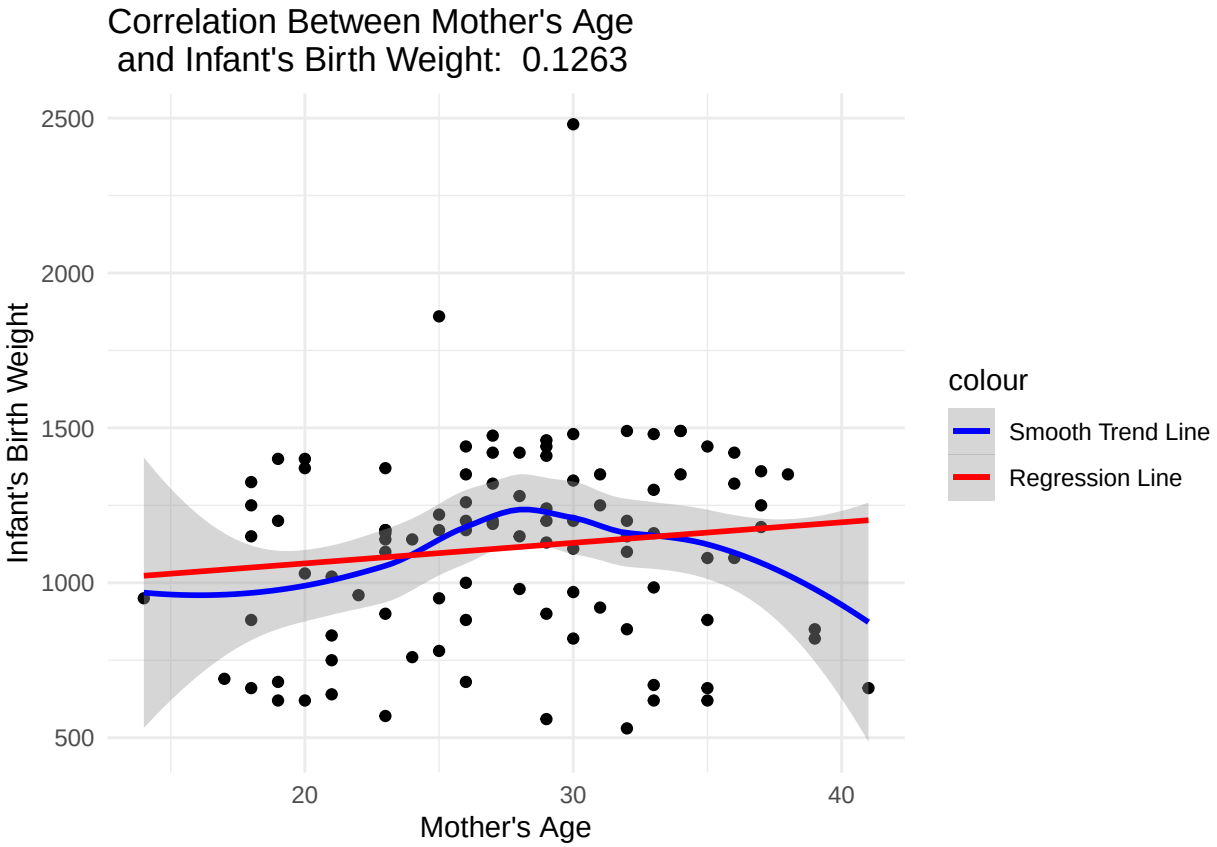
12.2

```
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

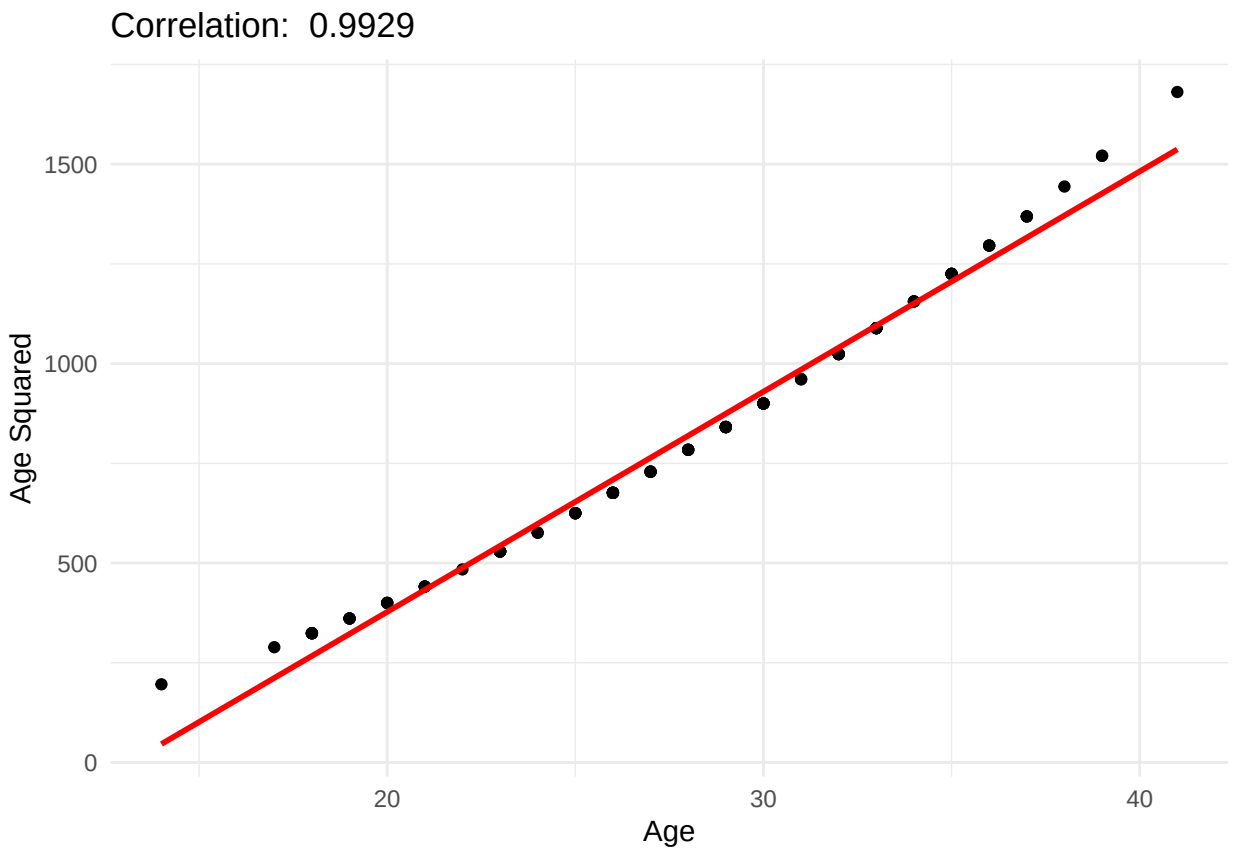
Before fitting the model, we wish to investigate the relationship between mother's age and infant's birth weight. Since the problem asks us to fit the model with age squared, we will have a second order polynomial relationship. We will look at the scatter plot to find any visual evidence that such model is justifiable.



We can see that we should fit the plynomial regression model because the smooth line shows a curved relationship between the two variables. However, the confidence bound around the smooth line suggest that potentially we may be able to fit a straight, first order, line in order to predict infant's birth weight. Overall, it is not very clear to what the verdict is, so we will fit the model with a higher order term and will use statistical tests to verify contribution of the squared term. Pearson's linear correlation estimate is low, so we should not expect to see string statistical evidence that mother's age is a strong predictor for infant's birth weight.

It is known that inclusion of higher order terms introduces multicollinearity issue to the model, which is hard to handle, and affects confidence intervals for predictors. Normally, we wish to perform another transformation of variables called *centering* in order to reduce the degree of linear correlation between the linear and higher order terms, however, I decided to include that into the appendix.

The plot below shows correlation between age and age squared.



Since the two variables are almost perfectly correlated, we expect that estimate for the standard error of $\hat{\beta}_i$ are higher in the model with no centering transformation applied. We verify it in the appendix section.

We are now ready to fit the model and explore the contribution of age-squared term. Model specification:

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

We obtain model the estimates from the model and present them in the table below:

Table 1: Polymonial 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

Comments:

- R-squared is 0.3904 and Adjusted R-squared is 0.3714
- The number of gestational weeks is an extremely strong predictor of the infant's birth weight. Each additional week add an average of 75.667 units of measurement (not sure what they are in this problem) to infant's birth weight

- Both linear are quadratic terms for mother's age are not statistically significant, and therefore we do not have enough evidence to reject the null hypothesis and conclude that the coefficients for these predictors are statistically different from zero.
- An addition of a quadratic term turns the effect of age on birth weight from a straight line to the parabola. We can use estimates of the linear and quadratic terms to describe the shape of this parabola.
  - 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.
  - In our case the coefficient is -0.582, so the effect can be visualized as a wide downward-pointing parabola.
  - A very wide parabola usualyy does not indicate a strong effect, and visually it should apprear closer to a straight line with a zero linear coefficient.

Evaluate Extra Sum of Squares

We already know that the linear and quadratic terms of mother's age are not strong predictors of birth weight. In this section we will specifically investigate the extra sum of squares and partial coefficient of determination to desribe their predictive power in more depth.

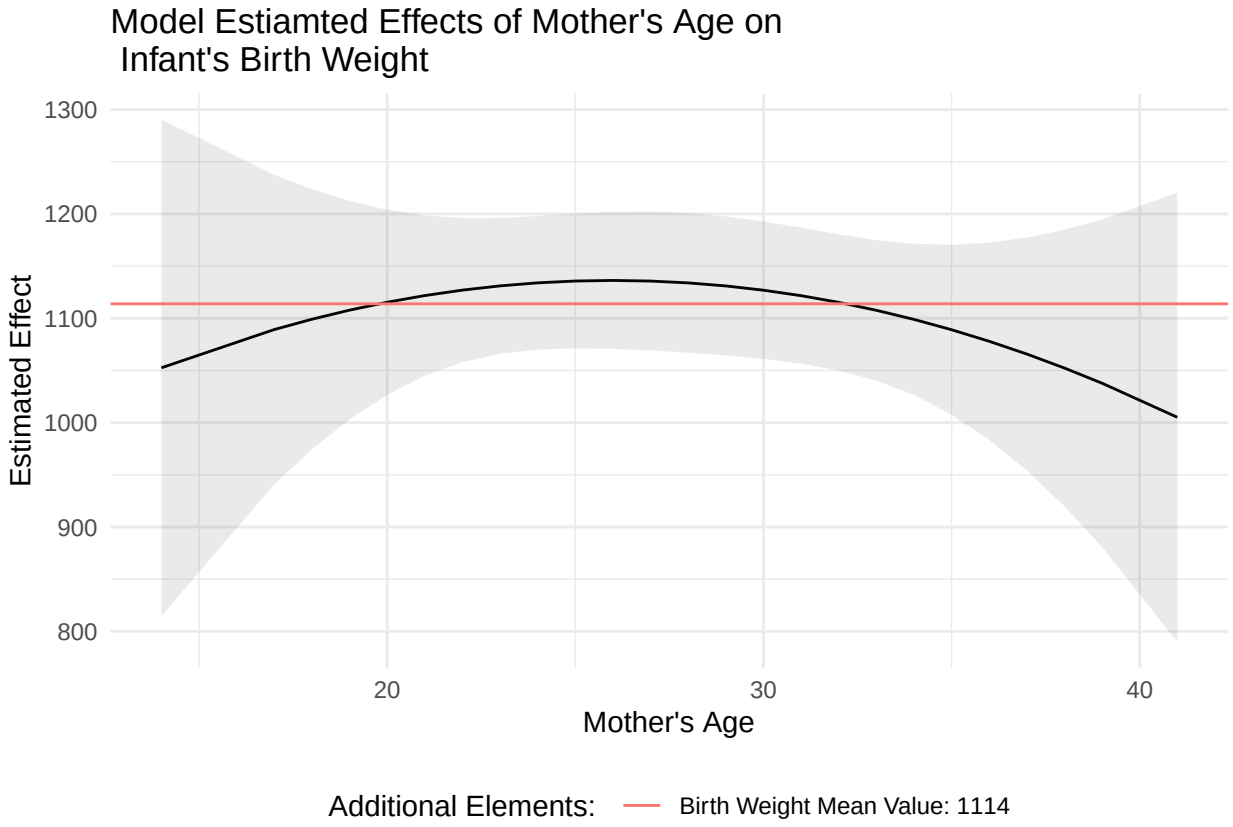
I used built in `R` functions to create an ANOVA table to decompose SSR and obtain partial coefficients of determination.

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 SSR for mother's age squared is: 48879.8443669
- Extra R^2 of mother's age squared after the addition of gestational weeks and linear term for mother's age is: 0.0081, which is extremely low
- We can see once again that these F tests are directly related to the T test we obtain from the model summary. The connection can be seen by observing that the p-values are the same, and F-test statistics is the square of the T-test statistic.
- The p-value of Extra SSR for the squared mother's age term is 0.3773, so we can't reject the null hypothesis. Therefore, there is no statistical evidence that the squared mother's age term helps to meaningfully explain variation in birth weight.
- Both linear and quadratic terms for mother's age can be removed.

Visualize Model Effects

We conclude this section by visualizing the effects both linear and quadratic terms on the birth weight of an infant. As we can see, the fitted effect is a downward-facing parabola, as we expected by looking at the coefficients.



This plot visually confirms all inferences and conclusions we have made thus far about the linear and quadratic terms of mother's are.

- We can see that the confidence bound around fitted effect line includes a flat straight line with the zero-coefficient. So, a 95% confidence interval includes a 'scenario' where the predictors show no effect on the response variables
- We have a very wide parabola with a downward-facing ends, which is the result of a very small absolute values of a coefficient and a negative value of a coefficient.

Interpretation of Mother's Age Coefficients

It does not make sense to interpret the linear term while "holding the square term constant", because the two are a function of the same measurement. This conclusion also follows from the fact that the two variables are correlated both mathematically and physically.

12.2 - B

Correlation Transformation for variables $Y, X_1, ..., X_{p-1}$, denoted by V, is given by this formula:

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

We use the code below to transform the variables

```
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))
```

Using transformed variables, we use this model specification and provide a summary for this model below:

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

In addition to the correlation transformed coefficient, we provide a summary table for the model of unstransformed variables, which we saw in part (A)

Table 2: 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 3: 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

Comments:

- Intercept is zero when we fit the model to the correlation transformed variables, which is to be expected.
- T and P values are the same for the covariates in both models, which suggests that the scale of the predictors should not affect the statistical tests and results for the inference of coefficients and standard errors.
- Therefore, same conclusions apply for the correaltion transformed regression model and orignal scale model.

12.2 - C

Transformation back to the original scale is given by this formula:

For variables $X_1, ..., X_{p-1}$:

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

Once we obtain the original scale estimates, we can calculate the intercept term. We omit this calculation here. If we verify that $\hat{\beta}_1, ..., \hat{\beta}_{p-1}$ match the original estimates after the transformation back, we know that we also will obtain the proper intercept term.

Back transformation will be done using code in the chunk below:

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

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

I hid the code that constructs a data frame for presentation. In addition to the $\hat{\beta}_i$ estimates I included their confidence intervals. It was of interest to me to verify that linear transformation applies to both the coefficient and the standard errors.

Recall the the original model with the transformed variables was called `inf_lm`. I used it for Extra SS, t-tests and model effects in the previous sections.

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

We can obtain standard errors and confidence intervals for the estimates to compare with the transformation back from the correlation transformation procedure.

Table 4: Original Model Estiamtes 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 5: Estimaes and C.I. obtained via Back-Trnasformation

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

As we can see, the results matched.

13.4

```
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"
```

13.4 - A

We can summarize each variable that we consider for analysis and use this table to find the number of predictors that we will have:

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

cig$arm <- as.factor(cig$arm)

cig <- data.frame(one_hot(as.data.table(cig))) %>% select(-arm_5)

cig[,4:(length(cig)-1)] <- lapply(cig[,4:(length(cig)-1)], as.factor)
```

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

meta_data <-

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

Table 6: Sumamry of Covariates

Predictors	Assigned Class	N of Unique Values
age	numeric	51
arm_6	factor	2
arm_7	factor	2
arm_8	factor	2
gender	factor	2
educ2	factor	2
income30	factor	2
ftnd	numeric	8

After consideration of all variables that we need for analysis, we know that this is the final set of covariates.

- Arm will result in $4 - 1 = 3$ variables
- Age is untouched
- FTND is treated as continuous

So, the total number of predictors is 3 for Arm indicators, plus other covariates, which results in 8 total predictors.

13.4 - B

In this section we will create a regression model for $Y1$, provide a summary table for model estimates, and look at three different ways to adjust p-values for multiple comparisons.

We then will repeat this process for $Y2$.

Regression on Y1

First, we specify the model in the code chunk below. Regression model has 8 predictors, so we will avoid writing the entire expression for $E[Y]$.

```
y1_lm1 <- lm(Y1 ~ . - Y2, data = cig )
```

Summary of coefficients and tests statistics is given below:

Table 7: Original Scale Regression Estimates

Model Term	Estimate	Std. Error	T-value	P-value
Intercept	0.027	0.281	0.094	0.925
Age	-0.003	0.004	-0.701	0.484
Arm 6	-0.690	0.175	-3.940	0.000
Arm 7	-0.068	0.174	-0.392	0.696
Arm 8	-0.426	0.179	-2.380	0.018
Gender	-0.112	0.109	-1.031	0.304
Education	-0.066	0.112	-0.588	0.557
Income >= $30K	-0.229	0.119	-1.922	0.056
FTND	0.046	0.042	1.093	0.276

- It appears that at the $\alpha = 0.05$ significance level, participants from group 6 and higher levels of income saw their measurements consistently reduced, after adjusting for other predictors. Other variables do not show a statistically significant relationship with the response variable.

Since we have 8 tests for each predictor variable, we have $1 - (1 - 0.05)^8 = 0.3366$ probability of at least one test being a false positive. This is not an incredibly high probability, but it definitely raises a cause for concern. Therefore, we need to make an adjustment for performing multiple comparisons at the same time.

We begin with the **Bonferroni Adjustment**. This is the simplest, and most conservative adjustment. Given our desired significance level, $\alpha = 0.05$, the new level at which we declare significance is $Bonferroni\ adjusted\ P - value = 0.0063$.

Thus, we present the table with predictors and corresponding p-values from the multiple regression model, and denote predcitors which are statistically significant at the new level with a "*".

```
sum_bonf_adj <- sum2 %>% select(`Model Term`, `P-value`)
sum_bonf_adj$`Significant at Adj. Level` =
  with(sum_bonf_adj,
       ifelse(`P-value` < 0.05 / n_predictors , "*", "")
       )

sum_bonf_adj %>%
  kbl( booktabs = T, caption = "Regression of Y1 Bonferroni Adjusted Comparison") %>%
```

```
    kable_styling(latex_options = c("striped", "HOLD_position")) %>%
    column_spec(3, width = "2cm")
```

Table 8: Regression of Y1 Bonferroni Adjusted Comparison

Model Term	P-value	Significant at Adj. Level
Intercept	0.925	
Age	0.484	
Arm 6	0.000	*
Arm 7	0.696	
Arm 8	0.018	
Gender	0.304	
Education	0.557	
Income >= $30K	0.056	
FTND	0.276	

it appears that still only an indicator for Arm 6 group is a statistically significant predictor.

Next we would like to implement **HOLM Adjustments**. The procedure involves iterrative evaluation of p-values using these steps:

- order p-values smallest to largest
  - if first p-value if smaller than $0.05/8 = 0.0063$ then conclude significance, and move to next predictor, otherwise stop, none are significant
- next predictor will be tested at $0.05/7 = 0.0071$.
  - So, each time we declare a predictor significant, we increase the denominator for the new cutoff by $+1$.
- The process stops when we find the first predictor that is not statistically significant

We use the code below to evaluate p-values according to the adjustments and present the table with the results. Note that the predictors are sorted in the table from the smallest to the highest original p-value.

```
holm_data <-
  sum2 %>% select(`Model Term`, `P-value`) %>% arrange(`P-value`) %>%
  filter(`Model Term` != "Intercept")

holm_data$`Comparison P-value` <- 1
holm_data$`Significant at Adj. Level` <- ""

cur_adj_n <- n_predictors

for(i in 1:nrow(holm_data)){

  cur_level <- 0.05 / cur_adj_n
  holm_data[i,3] <- cur_level

  if(holm_data[i,2] <= cur_level ){
    cur_adj_n <- cur_adj_n - 1
    holm_data[i,3] <- cur_level
    holm_data[i,4] <- "*"
  }
}
```

```
holm_data[,2:3] <- lapply(holm_data[,2:3], round_3)

holm_data %>%
  kbl( booktabs = T, caption = "Regression of Y1 HOLM Adjusted Comparison") %>%
    kable_styling(latex_options = c("striped", "HOLD_position")) %>%
    column_spec(c(3,4), width = "2cm")
```

Table 9: Regression of Y1 HOLM Adjusted Comparison

Model Term	P-value	Comparison P-value	Significant at Adj. Level
Arm 6	0.000	0.006	*
Arm 8	0.018	0.007	
Income >= $30K	0.056	0.007	
FTND	0.276	0.007	
Gender	0.304	0.007	
Age	0.484	0.007	
Education	0.557	0.007	
Arm 7	0.696	0.007	

Once again, Arm 6 indicator variable is the only statistically signifcant predictor.

And finally we also implement the **Hochberg Adjustments**. The process is somewhat similar with the HOLM adjustment, but it is different enough that we should implement it. The process is given by:

- Sort P-values largest to smallest
- Compare the largest to 0.05, if significant, declare all significant
- Otherwise, compare the next one to $0.05/2 = 0.025$
- Keep comparing to $0.05/3$, $0.05/4$, etc.. until we find a comparison where the predictor is not statistically significant, the first such predictor terminates the process.

```
hoch_data <-
  sum2 %>% select(`Model Term`, `P-value`) %>% arrange(-`P-value`) %>%
  filter(`Model Term` != "Intercept")

hoch_data$`Comparison P-value` <- 0.05
hoch_data$`Significant at Adj. Level` <- ""

cur_adj_n <- 1

for(i in 1:nrow(hoch_data)){

  cur_level <- 0.05 / cur_adj_n
  hoch_data[i,3] <- cur_level

  if(hoch_data[i,2] > cur_level){
    cur_adj_n <- cur_adj_n + 1

    holm_data[i,3] <- cur_level
  }
}
```

```r
hoch_data[,4] <- ifelse(hoch_data[,2] < hoch_data[,3], "*", "")

hoch_data[,2:3] <- lapply(hoch_data[,2:3], round_3)

hoch_data %>%
  kbl( booktabs = T, caption = "Regression of Y1 HOCHBERG Adjusted Comparison") %>%
    kable_styling(latex_options = c("striped", "HOLD_position")) %>%
    column_spec(c(3,4),  width = "2cm")
```

Table 10: Regression of Y1 HOCHBERG Adjusted Comparison

Model Term	P-value	Comparison P-value	Significant at Adj. Level
Arm 7	0.696	0.050	
Education	0.557	0.025	
Age	0.484	0.017	
Gender	0.304	0.013	
FTND	0.276	0.010	
Income >= $30K	0.056	0.008	
Arm 8	0.018	0.007	
Arm 6	0.000	0.006	*

Once again, Arm 6 indicator variable is the only statistically signifcant predictor.

Therefore, after a series of p-value adjustemtns we can conclude that we do not have any false positive statsistically signifiacnt predictors, and we can state that Arm 6 is the only statistically significant predcitor after adjusting for other covaraites.

Regression on Y2

Similarly, we will create the same set of summary and adjustment table for the regression model for Y2 variable, but will omit most of the commentary due to similarities with the Y1 regression model.

```r
y2_lm1 <- lm(Y2 ~ . - Y1, data = cig )
```

Table 11: Original Scale Regression Estimates

Model Term	Estimate	Std. Error	T-value	P-value
Intercept	-0.183	0.438	-0.418	0.677
Age	-0.002	0.006	-0.243	0.808
Arm 6	-0.278	0.273	-1.017	0.310
Arm 7	0.195	0.272	0.718	0.474
Arm 8	-0.095	0.279	-0.341	0.734
Gender	-0.096	0.170	-0.567	0.572
Education	-0.198	0.175	-1.129	0.260
Income >= $30K	-0.218	0.186	-1.176	0.241
FTND	0.056	0.066	0.855	0.394

- Bonferroni Adjustments

Table 12: Regression of Y2 Bonferroni Adjusted Comparison

Model Term	P-value	Significant at Adj. Level
Intercept	0.677	
Age	0.808	
Arm 6	0.310	
Arm 7	0.474	
Arm 8	0.734	
Gender	0.572	
Education	0.260	
Income >= $30K	0.241	
FTND	0.394	

- HOLM Adjustments

Table 13: Regression of Y2 HOLM Adjusted Comparison

Model Term	P-value	Comparison P-value	Significant at Adj. Level
Income >= $30K	0.241	0.006	
Education	0.260	0.006	
Arm 6	0.310	0.006	
FTND	0.394	0.006	
Arm 7	0.474	0.006	
Gender	0.572	0.006	
Arm 8	0.734	0.006	
Age	0.808	0.006	

- Hochberg Adjustments

Table 14: Regression of Y2 HOCHBERG Adjusted Comparison

Model Term	P-value	Comparison P-value	Significant at Adj. Level
Age	0.808	0.050	
Arm 8	0.734	0.025	
Gender	0.572	0.017	
Arm 7	0.474	0.013	
FTND	0.394	0.010	
Arm 6	0.310	0.008	
Education	0.260	0.007	
Income >= $30K	0.241	0.006	

As a result of the review, there are no statistically signifacnt predictors of Y2 when all variables are included.

13.4 - C

In this section we will again do the two different models for Y1 and Y2, will look at the model selection plots, and summarize the final selected model.

Step Wise Regression on Y1

Since we have 8 predictors, and for each predictor we have an option to include or not include it into the model, we have a total of $2^8 = 256$ possible models we can create. Therefore, using built in fucntions, we ontain a subset of candidate models. We will evaluate these candidate models using adjusted R-squared and AIC metrics. There two metrics should provide us with the simplest and most effective models that explain the higest proportion of Y1 variance, after adjusting for the number of predictors in the model.

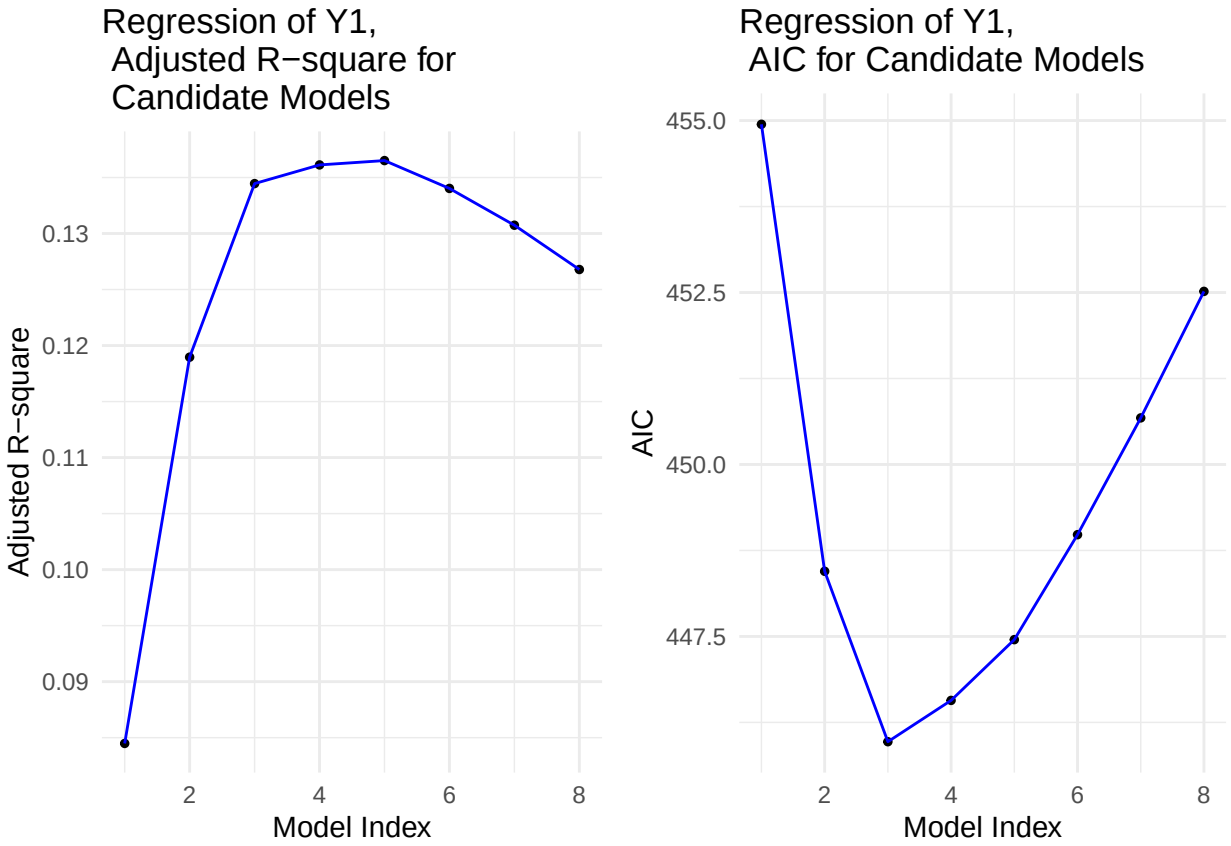
Table below lists 8 candidate models with the best R^2 and AIC scores.

```
k <- ols_step_best_subset(y1_lm1)

k %>% dplyr::select(n, predictors) %>%
  kbl(booktabs = T,
      caption = "Regression of Y1, Best Candidate Models") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Table 15: Regression of Y1, Best Candidate Models

	n	predictors
2	1	arm_6
17	2	arm_6 arm_8
65	3	arm_6 arm_8 income30
139	4	arm_6 arm_8 gender income30
210	5	arm_6 arm_8 gender income30 ftnd
231	6	age arm_6 arm_8 gender income30 ftnd
252	7	age arm_6 arm_8 gender educ2 income30 ftnd
255	8	age arm_6 arm_7 arm_8 gender educ2 income30 ftnd

Values of metrics are presented on the plots below, each model is indexed according to the table above.



It apprears that model 3 is perhaps the best possible model that we can employ to explain variation in Y1. Metrics for this model are given below:

Table 16: Regression of Y1, Parameters of Selected Model

Predictors	R-squared	Adj. R-squared	AIC
arm_6 arm_8 income30	0.148	0.134	445.969

Therefore, we fit this model, and provide a summary table for its coefficients and other statistics:

Table 17: Regression of Y1,

Model Term	Estimate	Std. Error	T-value	P-value
Intercept	-0.077	0.086	-0.897	0.371
Arm 6	-0.645	0.128	-5.033	0.000
Arm 8	-0.403	0.132	-3.045	0.003
Income >= $30K	-0.246	0.117	-2.106	0.036

After selecting the best possible predictors we observe different results:

- All predictors selected are now statistically significantly related to the outcome variable.
- The fact that p-values and coefficients changed implies that we had a multicollinearity problem in the previously stated regression model with all possible predictors.

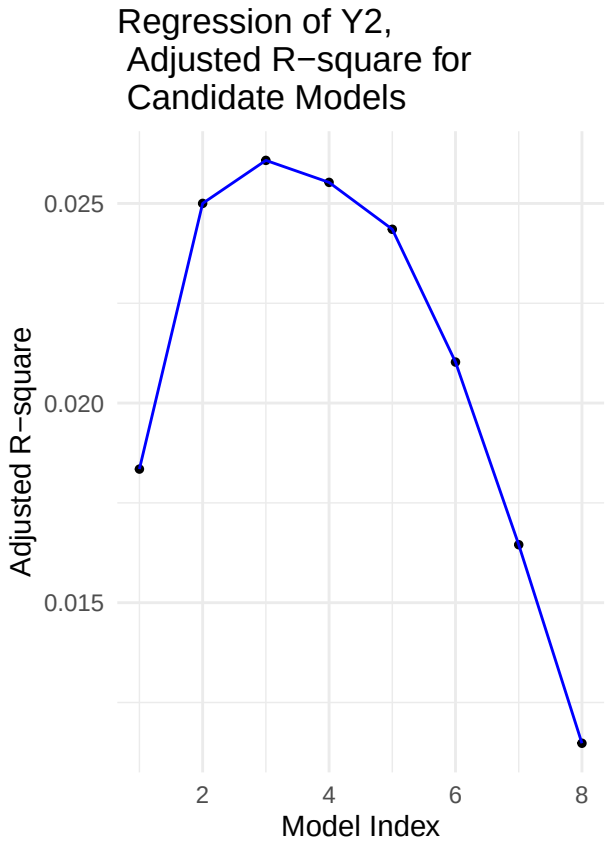
- For example, coefficient for Arm 8 indicator variable changed from -0.095 to -0.403
  –

Step Wise Regression on Y2

```
k <- ols_step_best_subset(y2_lm1)

k %>% dplyr::select(n, predictors) %>%
  kbl(booktabs = T,
      caption = "Regression of Y2, Best Candidate Models") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Table 18: Regression of Y2, Best Candidate Models

	n	predictors
3	1	arm_7
25	2	arm_7 income30
80	3	arm_7 educ2 income30
135	4	arm_6 arm_7 educ2 income30
207	5	arm_6 arm_7 educ2 income30 ftnd
244	6	arm_6 arm_7 gender educ2 income30 ftnd
254	7	arm_6 arm_7 arm_8 gender educ2 income30 ftnd
255	8	age arm_6 arm_7 arm_8 gender educ2 income30 ftnd



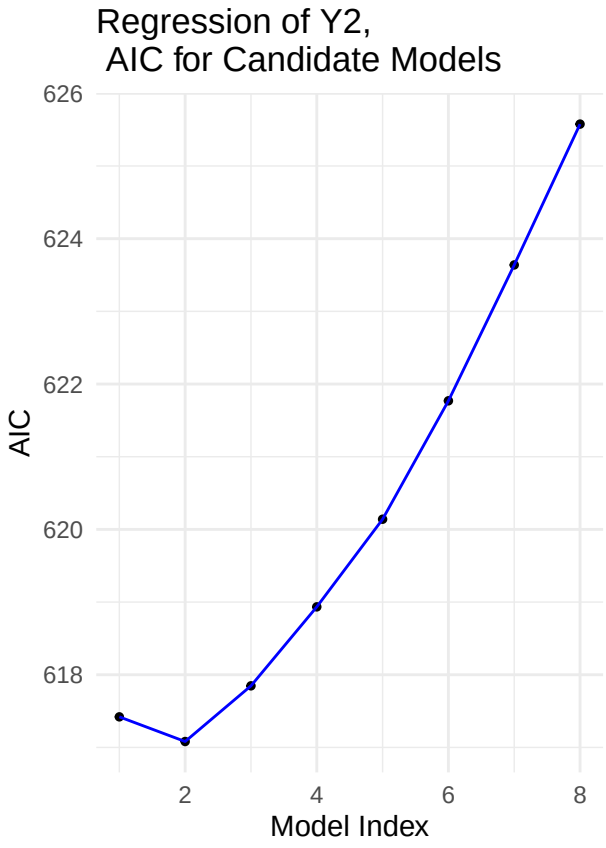


Table 19: Regression of Y1, Parameters of Selected Model

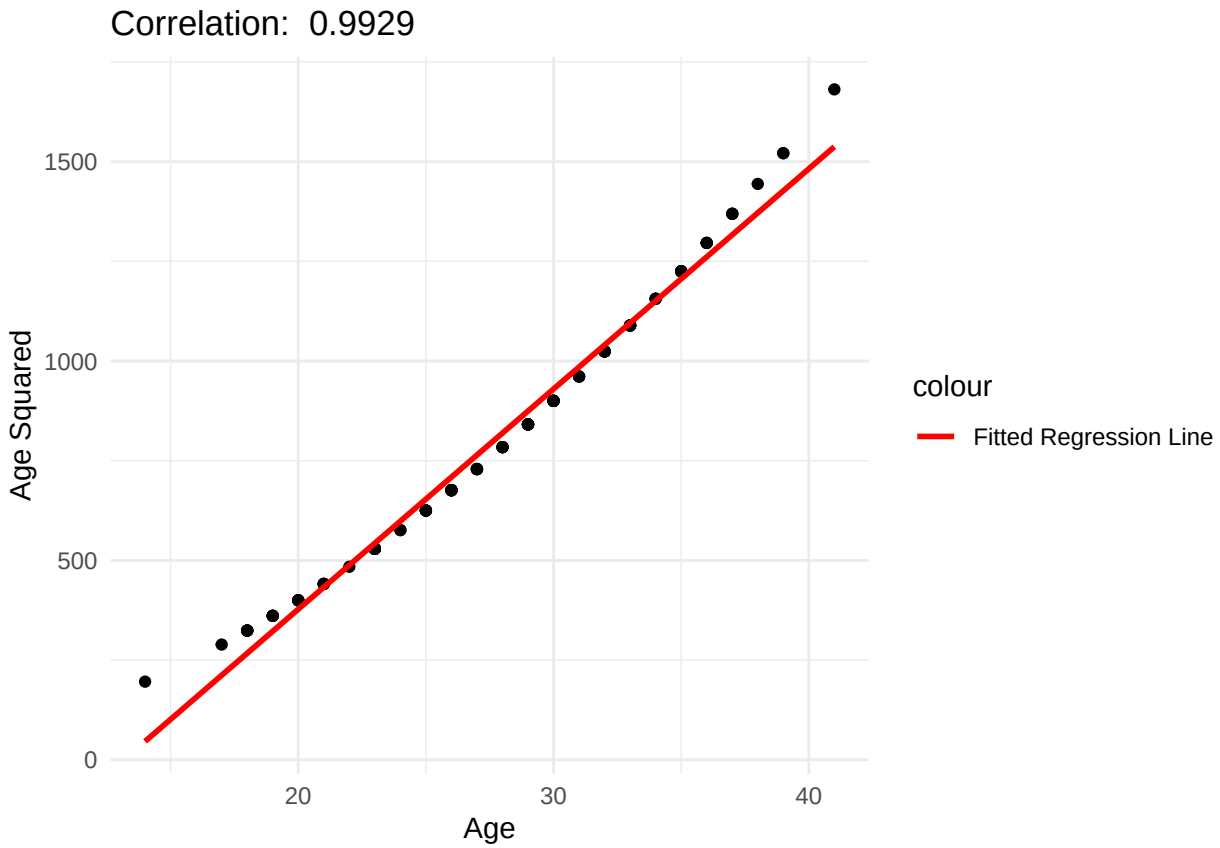
Predictors	R-squared	Adj. R-squared	AIC
arm_7 income30	0.035	0.025	617.082

Table 20:

Model Term	Estimate	Std. Error	T-value	P-value
Intercept	-0.372	0.115	-3.224	0.001
Arm 7	0.382	0.180	2.130	0.034
Income >= $30K	-0.275	0.180	-1.522	0.130

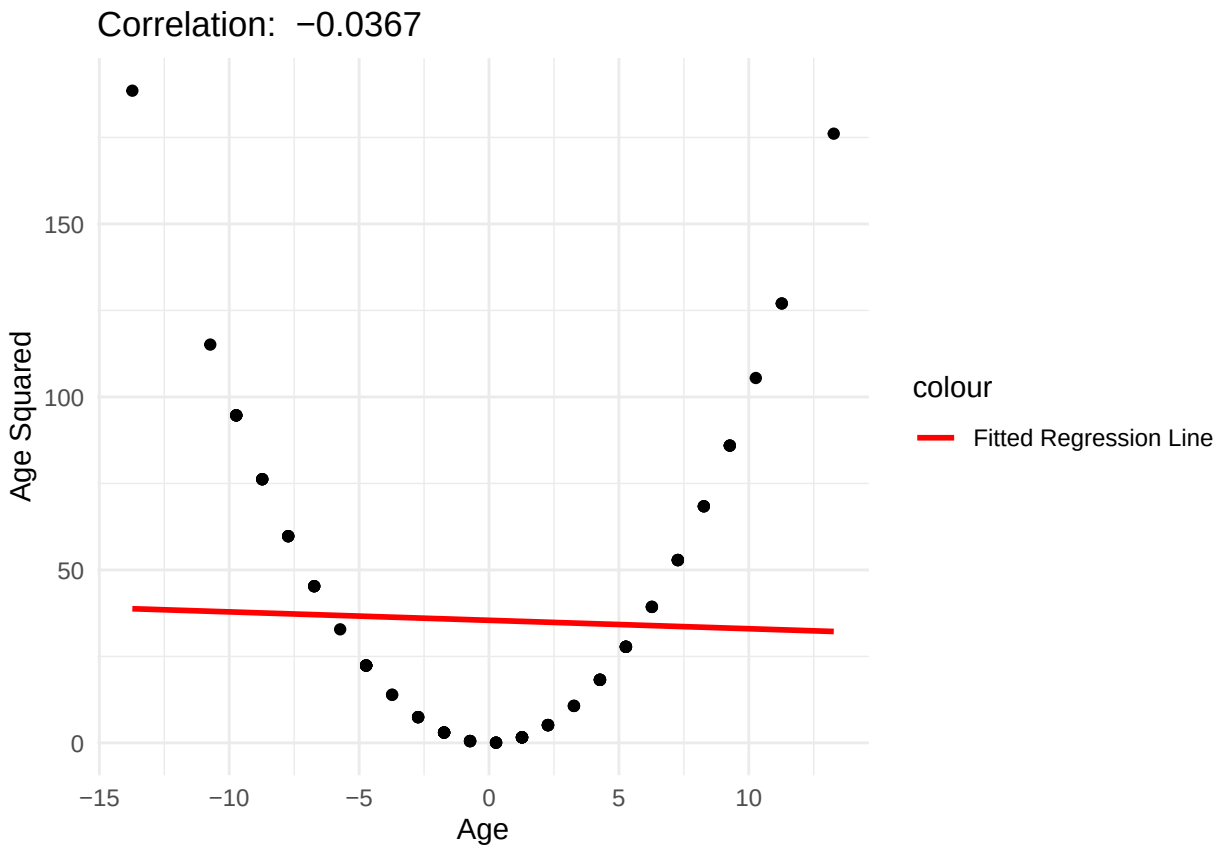
Appendix: 12.2

```
# 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()
```

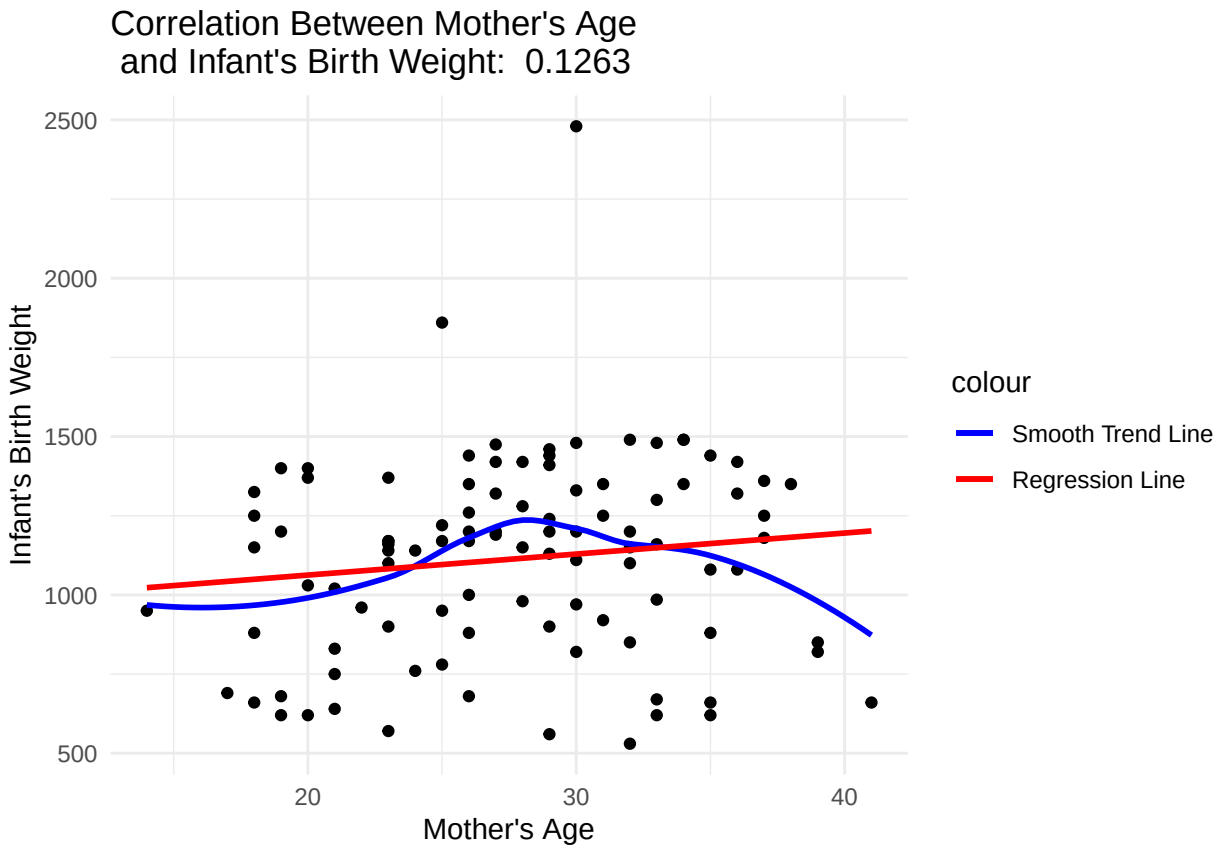



```
# now apply centering:

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")) +
```

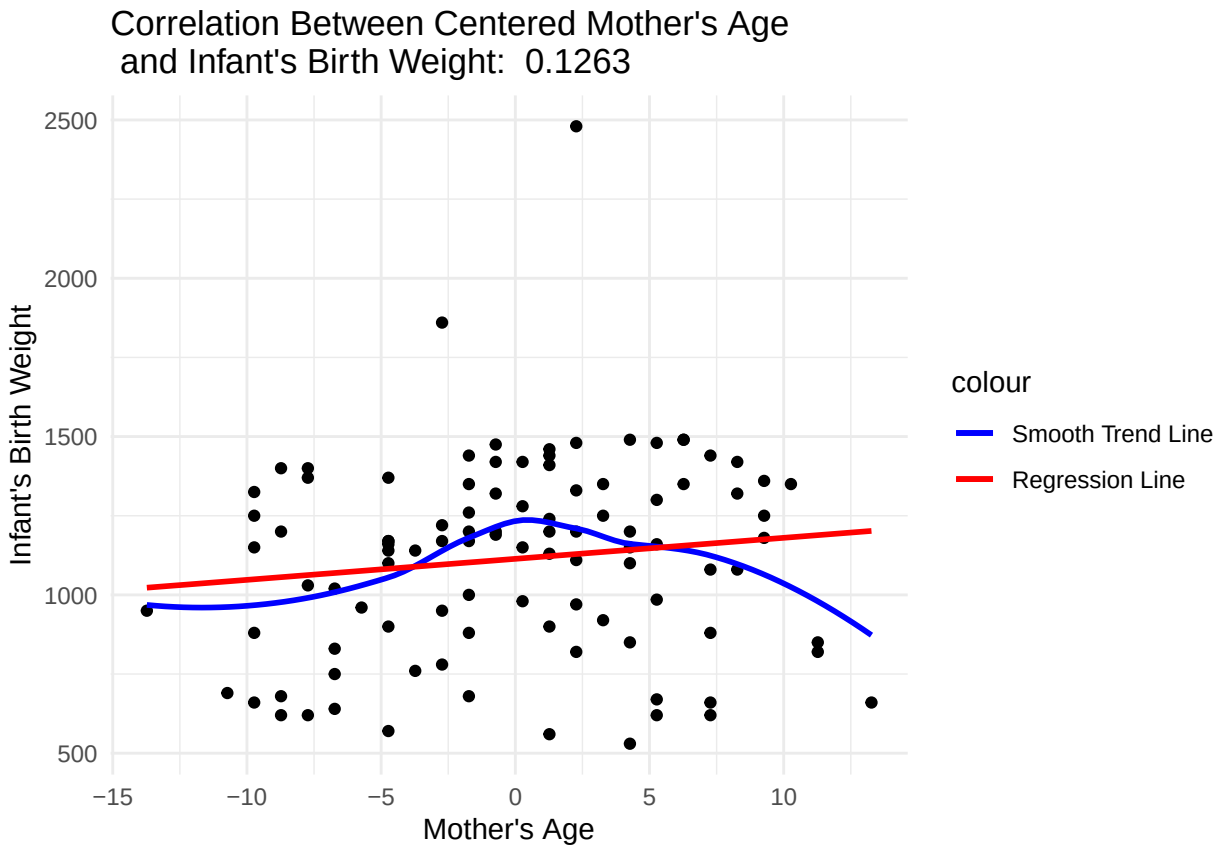
```
  xlab("Age") +
  ylab("Age Squared") +
  ggtitle(paste("Correlation: ", round(cor(infants_f$m_age_centered, infants_f$m_age_centered^2),4))) +
  theme_minimal()
```




```
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
Infant's Birth Weight
2500
2000
1500
1000
500
−15
−10
−5
0
5
10
Mother's Age
colour
Smooth Trend Line
Regression Line