Essay Front Page

Please complete this sheet electronically and attach it as the front page

for each essay submitted.

Candidate Number:	VGGL5

****Please remember that you must use your **candidate number** as the Submission Title on Turnitin****

Essay Title:	FINAL ASSESSMENT 3,000 WORD ESSAY
Essay Number:	1
Module Code:	POLS0008
Module Title:	Introduction to Quantitative Research Methods
Submission Date:	03/05/2022
Word Count:	2,960

Please tick this box if you **do not** agree to your essay being used anonymously in future teaching ☐

****Your name must not appear anywhere in this submission****

Introduction:

This report is based off of data collected by the British National Surveys of Sexual Attitudes and Lifestyle (Natsal), 2011-12. In this report, I will be analysing the relationship between age at first birth and related variables, in a campaign aimed for new parents, as commissioned by the National Childbirth Trust (NCT). These variables include independent variables like sex, relationship status, education, government region, conservative score, and whether the responded had their first child by age 18. Dependent variables include age at first birth, and whether respondents have screened positive for depression. For the purposes of this commission, certain variables were found to be critical and will be discussed in the analysis.

Variables are also distinguished by order of measurement in table 1 below. Variables that qualify as numeric include age at birth of first child, body mass index (BMI), and the conservative attitudes scale score because we can discuss the numerical difference of the size of the responses and take averages with those values. Numerical values are summarized using the mean below as they follow normal distribution. On the other hand, the average, sum, and difference of categorical variables like relationship status does not have any clear meaning, and are therefore summarised using frequencies and proportions. Categorical variables can also be further specified into ordinal and nominal, depending on whether or not the levels have a natural ordering. These descriptive statistics are demonstrated in the table below.

Table 1

Characteristics of Participants (N = 8,381)

Numerical Values	**Mean (SD)**	**Missing Values**
Age at First Birth	25.07	0.05
Body Mass Index	26.69	0.05
Sexually Conservative Attitude Scale	0.14	0.08
Categorical Values	**Proportion (%)**	**Missing Values (%)**
Respondent’s Sex		
Male	34.42	
Female	65.58	
Screened Positive for Depression		
Yes	10.90	4.30
No	84.80	
Relationship Status		
Married or civil partnership	52.54	2.94
Living with a partner	13.69	
In a ‘steady’ relationship but not living together	9.13	
Not in a ‘steady relationship’	21.72	
Had Child Before 18		
No	88.93	5.33
Yes	5.74	
Highest Educational Attainment		
Degree	21.05	
Higher education	23.72	0.45
GCSE, O-level or equivalent	36.89	
Foreign or other	1.13	
None	16.75	
Government Office Region		
North East	5.24	
North West (incl. the old ‘Merseyside’ region)	13.17	
Yorkshire & Humber	8.65	
East Midlands	8.57	
West Midlands	9.12	
South West	8.30	
Eastern	10.52	
Inner London	3.05	
Outer London	6.37	
South East	13.28	
Wales	5.49	
Scotland	8.23	

As demonstrated above, more than half the participants were female. The mean age at first birth was 25 years old and just over half of the participants were married. The proportion of participants that were single, or not in a 'steady relationship' came to almost 22%. The majority of participants had their first child after the age of 18, while almost 6% of participants had their child before. Almost 11% of participants screened positive for depression. The mean conservative attitude scale response was 0.14 and leans slightly towards the lower end of the scale. The mean measure of body mass index (BMI) is 26.69 and leans slightly towards the higher end of the scale. Just over twenty one percent of participants have obtained a degree, while almost 17% have obtained no educational qualifications at all. The largest percentage of participants come from North West and South East regions of England.

Question 1:

Measures of central tendency are meant to indicate the location of the middle or centre of a distribution, and is the point at which the distribution is in balance. After plotting age at first birth in a histogram, the data distribution is skewed towards the left of the dataset and there are outliers present. Therefore, the median is the most appropriate measure to report as it is not effected by extreme outliers or skewed distribution. In table 2 below, the median has been used to describe variation between the measures of relationship status, educational achievement, and government region.

Table 2:

Age at first birth varied by relationship status, educational attainment, and government region

Categorical Variable	Age at First Child		
	N =	Median =	IQR =
Educational Achievement			
Degree	1693	28	7
Higher Education	1921	25	7
GCSE	2967	23	7
Foreign	85	25	7
None	1245	21	6
Missing	18	27	8
Relationship Status			
Married/Civil Partnership	4212	26	8
Living with Partner	1119	23	8
In a 'Steady Ongoing Relationship but Not Living Together'	760	22	7
Not in a 'Steady Relationship'			
Missing	1791	23	7
	47	21	8
Government Region			
North East	429	23	7
North West (incl. the old 'Merseyside' region)	1047	24	8

Yorkshire & Humber	684	24	8
East Midlands	674	24	7
West Midlands	711	24	8
South West	651	25	8
Eastern	851	25	8
Inner London	213	25	9
Outer London	489	26	9
South East	1072	26	9
Wales	440	24	8
Scotland	668	24	7

Measures of dispersion describe how the data is clustered or dispersed around the middle of the distribution. In other words, this describes how age at first birth varies within the measures listed above. Seeing as we have chosen to report the median, the inter quartile ranges (IQR) are most appropriate, and describes the middle 50% of values when ordered from lowest to highest. Larger IQR values indicate the central portion of data is more spread out, while smaller values show that the middle values cluster together.

Based on the table, we can infer that the median age at first age is highest among parents with a degree and lowest among parents with no degree. The IQR is also lowest among parents with no degree, meaning the values are less varied. Median age at first birth is highest among parents who are married, and lowest among those who are in a 'Steady Ongoing Relationship but Not Living Together'. The government region where median age at first birth is highest is in the Southeast region and outer London. The IQR is highest among these two regions as well, describing more variation of the variables.

Question 2:

In order to determine if our variables for age at first birth for women in our sample are different to the national average (27.9 years), we must run a one-sample t-test to determine statistical significance. We select a one sample t-test as the appropriate test, because our test variable (mean age at first birth among women) is continuous, the sample is large (N = 4,596), and scores on the test variable are independent of one another. A plotted histogram for age at first birth among women demonstrates a roughly bell-shaped curve. Because of the large and random sample, any issues with normality and outliers are relaxed

The alpha-level was set to 0.05 and the p-value was returned. When the test was run on the mean age a first birth for women (24.03) with a mu value set to the national average (27.9), a p-value of 2.2e-16 was returned. This gives the probability associated with observing the difference in our sample if the null hypothesis were true. Because the value is lower than our alpha-level, we can reject our null hypothesis that there is no difference between the sample and the population. We can infer from this that the mean age at first birth for women in our sample is not different to the national average.

Question 3:

Table 3:

Cross-tabulation between whether parents had child before 18 and whether they screened positive for depression

Had Child Before 18	**Screened Positive for Depression**		
	No	**Yes**	**Row Total**
No	6599 88.90%	824 11.10%	7423 93.95%
Yes	400 83.68%	78 16.32%	478 6.05%
Column Total	6999	902	7901

In the cross-tabulation in table 3 above, whether or not parents had their child before the age of 18 was compared with whether or not they screened positive for depression. In looking at the outcomes, the proportion of parents who did have their child at before 18 and screened positive for depression was 16%. The proportion of parents who didn't have their first child by 18 and also screened positive for depression was 11.10%. It's worth mentioning that missing values were omitted from this table as the information is not relevant to the relationship we are trying to assess. Seeing as our outcome (screening for depression) is defining the columns, we have shown only row percentages in the table.

In order to measure association, we have decided to use Cramer's V. Considering we are using at least two categorical variables, and they are both nominal, it is an appropriate measure of association. Opposed to the Pearson's coefficient test alone, the Cramer's v test also demonstrates the strength of the relationship. The reported value for Cramer's V is 0.03. This falls between 0 to 0.2, which can be defined as a weak level of association. Using this standard we can say there is a weak association between having a child before 18 and screening positive for depression. Therefore, based on our sample data, parents who have had their child before the age of 18 are not very likely to screen positive for depression.

'Question 4:

To determine whether men or women are more likely to have their first child at a younger age, we will use a linear regression model to determine the relationship between the intercept and slope coefficient. Our hypothesis is that women are more likely to give birth at a younger age, as it optimal for both the health of the mother and baby. For the purposes of this question, we assigned dummy variables to the categorical variables of male and female (male=0, female=1). This was in order to meet the requirements of a linear regression model, in which both variables must be numerical. The findings of the regression model are listed in table below.

Table 4:

Data output on linear regression model

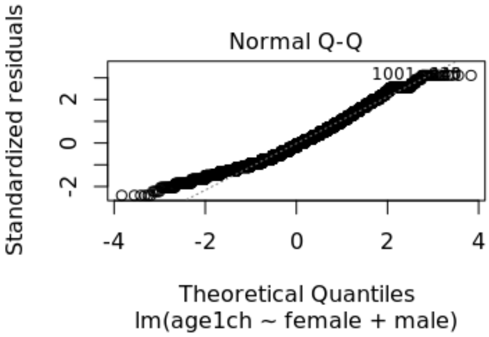
	Estimate	**P-Value**	**Multiple R-squared:**	**Durbin Watson Test:**
Male (Intercept)	27.13	< 2.2e-16	0.067	1.85
Female	-3.10	< 2.2e-16		

In our code, being male was assigned the value of 0, so when our linear regression output references the intercept, it was actually the age at birth of first child for men (27.13). The slope is -3.1008, and references how much the y variable (age at first variable) increases of one unit increase in x (in females). Because the coefficient is negative, this resembles a decrease in age at birth of first child for females. Taking the p-value into account, it smaller than our alpha value and therefore we can reject the null hypothesis and assure statistical significance. This information affirms our hypothesis that women are more likely to have their first child at a younger age, and is not a relationship by chance.

In terms of whether our data meets the four assumptions required to use a linear regression test, we address each of the fours assumptions separately. In order to test the independence of observations, the Durbin Watson test was completed on our model, and the value is included in table 4 above. We can interpret this value to mean that the assumption has been met, as the value (1.85) is extremely close to 2. In order to test normality, we utilised a normal probability plot to assess normality of residuals for the model. As seen in figure 1 below, the plot in both diagonal and linear and runs along the reference line, so therefore we can assume the residuals are normally distributed. Lastly, we used a plot of residuals versus predicted values to establish homoscedascity and linearity. As there was no clear pattern in the distribution and way the residuals appear randomly around the 0 line, we can assume both homoscedascity and a linear relationship.

<u>Figure 1:</u>

Normal probability plot for linear regression model



Question 5:

The last objective of this essay is to determine whether sexual conservative attitude scores predict age at birth of first child. We have chosen sexual conservative attitude score as a numeric variable to explain variation in age at first birth, because we believe those who are more sexually conservative prioritize marriage and childbirth at a young age. The hypothesis is that there is a negative relationship between sexual conservative score and age at first birth, and this relationship will be tested separately in both men and women. Two statistical models have been chosen, and the results posted in table 5 below.

Table 5:

Data output for correlation values

Sex	Spearman's Correlation Coefficient	P-Value
Female	-0.048	0.001
Male	-0.028	0.160

Histograms demonstrated that there was a negative skew in the distribution of age at first birth towards younger age in women with outliers giving birth at much older ages, whereas age at first birth was more normally distributed across men. Conservative attitude scores skewed slightly higher in women than in men, but both were relatively normally distributed. Scatterplots demonstrated that in both men and women, the relationship between sexual conservative score and age at first birth are non-linear, and therefore a Spearman's test was conducted to measure the strength and direction of the relationship. Interpretation of magnitudes and direction are where -1 indicates a perfect negative relationship and 1 indicates a perfect positive. The coefficient value found that there was a negative weak relationship between conservative score and age at first birth for both men and women.

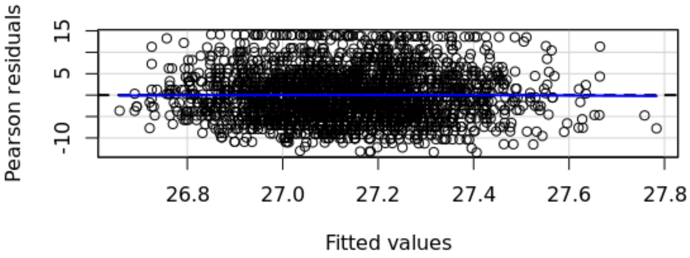
While this aligns with our hypothesis, the p-value was tested to determine the chance of getting these results if indeed there was no relationship between our variables (i.e. due to sampling error) at an alpha level of 0.05. For women, the p-value was less than 0.05, meaning we can reject the null hypothesis that there is no relationship and conclude that the findings are statistically significant. In men, on the other hand, the value is greater than 0.05 and therefore we cannot reject the null hypothesis and cannot conclude that there is a relationship. In summary, these findings can mean that while there is a statistically significant relationship between age at first birth and conservative attitude score for women, it is still quite weak. Regardless, of this we will continue to report on model fit and residuals of both men and women.

For women, the coefficient for slope is -0.23 and the intercept is 24.13. In men, the slope coefficient is -0.17 and intercept is 27.10. We can interpret this to mean that the predicted value of age at first birth when conservative scale score is zero for women is around 24 years old for women and around 27 for men. A score of zero on the conservative attitudes scale is within the bounds of the scale score and so this has meaningful interpretation. For the slope coefficient, since both values are negative, we can assume that for every one unit increase in X, there is a decrease on Y. We can assume that for every one unit increase in conservative scale score, there is a -0.30 unit decrease in age at first birth in women and -0.17 unit decrease in age at first birth in men. Our R squared value assumes the goodness-of-fit for our model. Based on our output, we can say that our model explains 0.2% % of the variance in age at first child by attitude scores in women, and 0.08% in men which is relatively low. Regardless, we can still draw conclusions on our statistically significant predictors (in this case, conservative scale scores for women) because that coefficient holds other predictors in the model constant.

Next we checked our models for two assumptions of our residual values: homoskedascity and normality. A plot of residuals against predicted (fitted) values alongside the non-constant variance score was used to test normality for both plots (see figure 4 below). Although at first glance the relationship seems linear, the output of the non-constant variance score for both men and women were statistically significant, so we can therefore suspect there is an issue with non-constant variance (aka heteroskedascity). A normal probability plot and computed a Bonferroni-corrected t test were used to assess normality of residuals (see figure 5 below). Seeing as plots for both men and women follow a diagonal linear line, we can assume the residuals are nearly normally distributed. However, because the t-test ran values larger than 0.05 for both men and women, we can conclude that the largest studentised residuals are indeed outliers. Residuals with only a few outliers is not a severe issue because in practice, we will never get perfectly predicted regression lines. It is worth remembering that we can have a good model and use it to draw conclusions about your sample, even if some assumptions are violated. This simply means we cannot generalize our findings beyond our sample.

Figure 4:

Plot of residuals against predicted (fitted) values for men (top) and women (bottom)



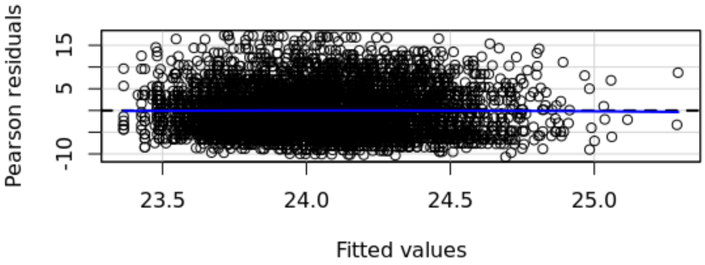
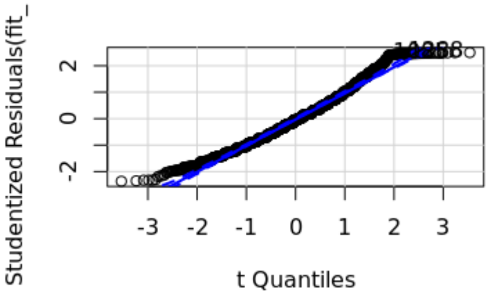
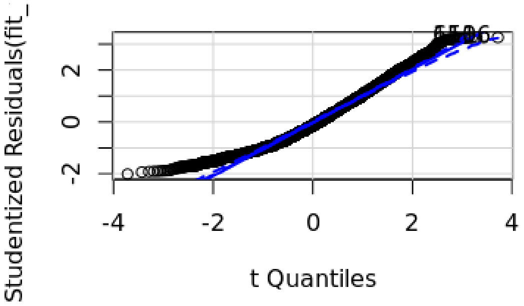


Figure 5:

Normal probability plot of residuals/errors for men (left) and women (right)





Conclusion:

In conclusion, findings from our data highlight populations from which the NCT should aim their campaign. It has been affirmed that the mean age of women at first birth is 27, however it's important to take into consideration that more conservative women are likely to have their first child younger. Men are likely to have their children older than women, and this is regardless of conservativity. Variance in distribution of age at first birth has been demonstrated across categories of relationship status, educational attainment, and government region, however because we have not conducted tests to determine statistical significance on these variables and BMI we cannot draw conclusions to the wider population. Non-critical variables include screening for depression, as tests have confirmed that there is a statistically non-significant relationship the two in our sample. We suggest then, that the NCT use the critical variables of conservative scale score and sex to target new parents by age of birth of first child and keeping this in mind when recruiting members of religious or conservative groups.

Appendix

```
#load data
library(haven)
data<-read_stata("natsal_3_teaching.dta")
View(data)

#introduction
#begin cleaning data
data$child18<-as_factor(data$child18, levels = "labels")
data$relstatgp2<-as_factor(data$relstatgp2, levels = "labels")
data$depscr<-as_factor(data$depscr, levels = "labels")
data$rsex<-as_factor(data$rsex, levels = "labels")

data$educ3
data$attconservative
data$gor_l

# select variables into a new dataset
myvars <- c("rsex", "bodymass", "depscr", "relstatgp2", "age1ch", "child18", "attconservative", "educ3", "gor_l")
newdata <- data[myvars]

#make gor_l and educ3 factors
newdata$educ3<-as_factor(newdata$educ3, levels = "labels")
newdata$gor_l<-as_factor(newdata$gor_l, levels = "labels")

#look at summary statistics of data
summary(newdata$rsex)
summary(newdata$bodymass)
summary(newdata$depscr)
summary(newdata$relstatgp2)
summary(newdata$age1ch)
summary(newdata$child18)
summary(newdata$attconservative)
summary(newdata$educ3)
summary(newdata$gor_l)

dim(newdata)
sapply(newdata[1, ], class)

newdata.v2<-subset(newdata, age1ch!=-1 & depscr!="Not Applicable" & child18!="Not Applicable")
table(newdata.v2$age1ch)
summary(newdata.v2$age1ch)

#look for missing variables now
is.na(newdata.v2$educ3)

newdata.v2$age1ch[newdata.v2$age1ch==99] <- NA
newdata.v2$bodymass[newdata.v2$bodymass==99] <- NA

#summary statistics
summary(newdata.v2$age1ch)
summary(newdata.v2$bodymass)
summary(newdata.v2$attconservative)
summary(newdata.v2$rsex)
summary(newdata.v2$depscr)
summary(newdata.v2$relstatgp2)
summary(newdata.v2$child18)
summary(newdata.v2$educ3)
summary(newdata.v2$gor_l)

#question 1
q1foreal <- ddply(newdata.v2, c("educ3"), summarise,
        N    = sum(!is.na(age1ch)),
        median = median(age1ch, na.rm=TRUE),
        IQR   = IQR(age1ch, na.rm=TRUE),
)
library(plyr)
q1foreal <- ddply(newdata.v2, c("educ3"), summarise,
                N    = sum(!is.na(age1ch)),
                median = median(age1ch, na.rm=TRUE),
                IQR   = IQR(age1ch, na.rm=TRUE)
)

q1foreal

q2foreal <- ddply(newdata.v2, c("relstatgp2"), summarise,
          N    = sum(!is.na(age1ch)),
          median = median(age1ch, na.rm=TRUE),
          IQR   = IQR(age1ch, na.rm=TRUE)
)
q2foreal

q3foreal <- ddply(newdata.v2, c("gor_l"), summarise,
          N    = sum(!is.na(age1ch)),
          median = median(age1ch, na.rm=TRUE),
          IQR   = IQR(age1ch, na.rm=TRUE)
)
q3foreal

summary(newdata.v2$age1ch)

#make a barplot to see how our findings for age1ch are distributed
install.packages('ggplot2')
library(ggplot2)
```

```
p <- ggplot(newdata.v2, aes(x=age1ch))
p
p + geom_histogram()



#question 2
#chi-squared test for age at first birth for women

#subset new data set for age at first birth specifically for women
question2set<- subset(newdata.v2, rsex=='Female', selec=c(age1ch))
question2set
chisq.test(question2set)
t.test(question2set$age1ch)

#run t test with mu set to national average
res <- t.test(question2set$age1ch, mu = 27.9)
res

#question 3
#subset a new dataset without missing or non applicable variables
sub_dataq3 <- subset (newdata.v2, child18 %in% c("Yes", "No") & depscr %in% c("Yes","No"))
summary(sub_dataq3)
table(sub_dataq3$depscr)
table(sub_dataq3$child18)

#create crosstab
library(gmodels)
child18table <- table(sub_dataq3$child18)
prop.table(child18table)
with(sub_dataq3, CrossTable(child18, format = c("SPSS")))
with(sub_dataq3, CrossTable(child18, depscr, prop.chisq = FALSE, format = c("SPSS")))
```

```
#look for only row percentages

with(sub_dataq3, CrossTable(child18, depscr, prop.chisq=FALSE, prop.c=FALSE, prop.t=FALSE, format=c("SPSS")))


#look for cramers v as measure of association

mytable.2<-table(sub_dataq3$child18, sub_dataq3$depscr)

print(mytable.2)


#redefine variables as factors o get rid of zeros

sub_dataq3$child18<-factor(sub_dataq3$child18)

sub_dataq3$depscr<-factor(sub_dataq3$depscr)

mytable.2<-table(sub_dataq3$child18, sub_dataq3$depscr)

print(mytable.2)


#find cramers v

library(vcd)

assocstats(mytable.2)



#question 4

library(tidyverse)

library(ggplot2)


#try creating new dataframe

question4set<- subset(newdata.v2, selec=c(rsex, age1ch))

question4set


#create dummy variables to chnage female and male into numerical values to fit regression

female<-ifelse(question4set$rsex == 'Female', 1, 0)

male<-ifelse(question4set$rsex == 'Male', 1, 0)


#create data frame to use for regression

df_reg<- data.frame(age1ch = question4set$age1ch,
          female = female,
          male = male)
df_reg

#create regression model
model <- lm(age1ch ~ female + male, data = df_reg)
plot(model)

#view regression model output
summary(model)

#durbin watson test to interpret correlation
dwt(model)

#plot model
plot(model)




#question5
#subset dataframes for women and men
q5women<-newdata.v2[ which(newdata.v2$rsex=='Female'),]
q5men<-newdata.v2[ which(newdata.v2$rsex=='Male'),]
q5women
#check assumptions for linear regression
#histogram to check for normality
library(ggplot2)
qplot(x=attconservative, data=q5women )
qplot(x=age1ch, data=q5women)

qplot(x=attconservative, data=q5men )
qplot(x=age1ch, data=q5men)

#scatterplot for women and men
ggplot(q5women, aes(x = attconservative, y = age1ch)) +
  geom_point(alpha=.2, position="jitter")

ggplot(q5men, aes(x = attconservative, y = age1ch)) +
  geom_point(alpha=.2, position="jitter")



#test for correlation coefficient
cor(q5women$attconservative, q5women$age1ch, use="complete.obs")
cor(q5men$attconservative, q5men$age1ch, use="complete.obs")

#spearman's correlation correlation coeff test with more output, p value etc
cor.test(q5women$attconservative, q5women$age1ch, method="spearman")
cor.test(q5men$attconservative, q5men$age1ch, method="spearman")

#fitting a linear model
fit_women <- lm(age1ch ~ attconservative, data=q5women)
fit_men<-lm(age1ch ~ attconservative, data=q5men)

#plot line of regression

library(ggplot2)
ggplot(q5women, aes(x = attconservative, y = age1ch)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red")

ggplot(q5men, aes(x = attconservative, y = age1ch)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red")

summary(fit_women)
summary(fit_men)

coef(fit_women)

#regression assumption tests for residuals
#plotting residuals

#testing for linearity
library(car)
residualPlot(fit_women)
residualPlot(fit_men)

#testing for homoskedasticity
library(lmtest)
coeftest(fit_women, vcov=hccm)
summary(fit_women)

qqPlot(fit_women)
qqPlot(fit_men)

#checking durbin watson test for independence of variables

ncvTest(fit_women)
ncvTest(fit_men)
library(lmtest)

outlierTest(fit_women)
outlierTest(fit_men)
```

```
dwt(fit_women)

dwt(fit_men)
```