More Interaction Models

2019-11-22

Preparation

In this set of notes, you will continue to learn about interaction models—twith emphasis on interactions between two continuous predictors. To do so, we will revisit the *fertility.csv* (see the data codebook) dataset you explored in Assignment 05. In that assignment, we found that contraception use in a country was negatively associated with fertility rates, even after controlling for differences in the country's female education level and infant mortality rate. Now we will explore whether the effect of contraception on fertility rates is a function of female's education level (i.e., Is there an interaction between contraception use and female education level?).

```
# Load libraries
library(broom)
library(corrr)
library(gridExtra)
library(tidyverse)

# Read in data
fert = read_csv(file = "~/Documents/github/epsy-8251/data/fertility.csv")
```

Exploring the Data for an Interaction between Contracpetion Use and Female Education Level

Ideally, we would only include an interaction effect in the model if there is support for this in the theoretical/substantive literature. However, barring this support, we might explore the sample data for empirical evidence of the interaction (generally via plots of the data). To explore an interaction effect between two quantitative variables poses some unique challenges.

To understand those challenges consider how we explored the interaction between sex and boundary-spanning work on guilt in a previous set of notes. We created a separate plot of the effect of boundary-spanning work on guilt for males and females, and asked whether the fitted line for males and females were parallel. In other words, we need to examine the relationship between $X1$ and Y for different levels of $X2$.

If we are examining an interaction between contraception use and female education level on fertility rates, we need to examine the effect of contraception use on fertility rates at different levels of female education. But, since female education is continuous (there are a lot of levels!), to explore this, we have to choose a small number of levels of female education. Another challenge is that in continuous variables, there are typically very few observations at a specific value of that variable, so instead of selecting specific values, we typically cut the variable into distinct ranges of values.

Here we empirically identify distinct ranges for female education, by examining the `summary()` output associated with that variable.

```
summary(fert)
```

```
   country             region          fertility_rate   educ_female
 Length:124         Length:124         Min.   :1.240   Min.   : 0.600
 Class :character   Class :character   1st Qu.:1.657   1st Qu.: 4.800
 Mode  :character   Mode  :character   Median :2.205   Median : 8.000
                                       Mean   :2.703   Mean   : 7.494
                                       3rd Qu.:3.085   3rd Qu.:10.250
                                       Max.   :7.030   Max.   :13.000
 infant_mortality  contraceptive    gni_class             high_gni
 Min.   :  2.200   Min.   : 6.00   Length:124         Min.   :0.0000
 1st Qu.:  7.025   1st Qu.:42.00   Class :character   1st Qu.:0.0000
 Median : 16.950   Median :59.50   Mode  :character   Median :1.0000
 Mean   : 26.050   Mean   :56.04                      Mean   :0.5887
 3rd Qu.: 40.925   3rd Qu.:71.25                      3rd Qu.:1.0000
 Max.   :117.400   Max.   :86.00                      Max.   :1.0000
```

Here, based on the quartiles, we might choose female education ranges of:

- Female education level < 4.8
- $4.8 \leq$ Female education level < 8
- $8 \leq$ Female education level < 10.25
- $10.25 \leq$ Female education level

Note that you will want to use the entire range of data to explore effects, otherwise, we might see a spurious relationship. Also, and this is VERY IMPORTANT, *This discretizing is ONLY carried out to create the plot. When we fit the actual interaction in the regression model, we use the continuous predictor.*

There are several R functions that can be employed to dicretize a continuous variable. We will use the `case_when()` function from **dplyr** to create a new variable that has four different categories, one for each quartile.

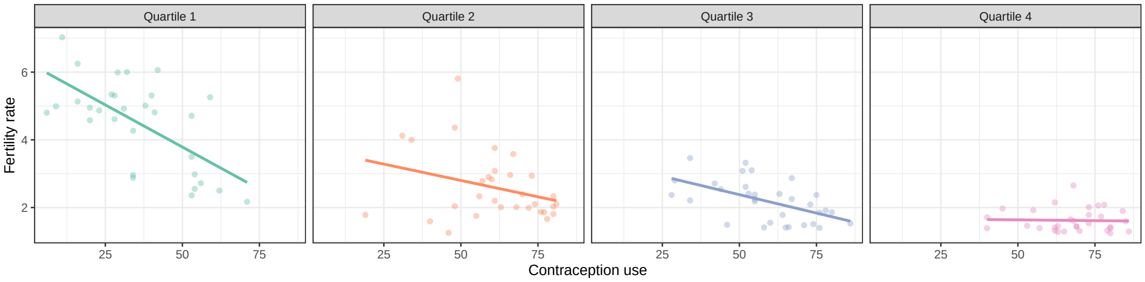
```
# Discretize female education level
fert = fert %>%
  mutate(
    female_educ_discrete = case_when(
      educ_female < 4.8 ~ "Quartile 1",
      educ_female >= 4.8 & educ_female < 8 ~ "Quartile 2",
      educ_female >= 8 & educ_female < 10.25 ~ "Quartile 3",
      educ_female >= 10.25 ~ "Quartile 4"
      )
  )

head(fert)
```

```
# A tibble: 6 x 9
  country region fertility_rate educ_female infant_mortality contraceptive
  <chr>   <chr>           <dbl>       <dbl>            <dbl>         <dbl>
1 Albania Europ~           1.49         9.1             15              46
2 Algeria Middl~           2.78         5.9             17.2            57
3 Armenia Europ~           1.39        10.8             14.7            57
4 Austria Europ~           1.42         8.9              3.3            66
5 Azerba~ Europ~           1.92        10.5             30.8            55
6 Bahama~ Latin~           1.97        11.1             13.9            45
# ... with 3 more variables: gni_class <chr>, high_gni <dbl>,
#   female_educ_discrete <chr>
```

Now we have discretized female education level, we can use our new discretized variable to examine the potential interaction with contraception use.

```
ggplot(data = fert, aes(x = contraceptive, y = fertility_rate, color = female_educ_discrete)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
    theme_bw() +
    xlab("Contraception use") +
    ylab("Fertility rate") +
  scale_color_brewer(palette = "Set2") +
  facet_wrap(~female_educ_discrete, nrow = 1) +
  guides(color = FALSE)
```



The empirical evidence is consistent with there being an interaction effect between contraception use and female education level on fertility rate in the sample; the effect of contraception use on fertility rate differs depending on the level of female education. It looks like the absolute magnitude of the effect of contraception use on fertility rate decreases for higher levels of female edcuation.

Fit the Interaction Model

To fit the interaction model, use the consituent main effects and the interaction term to predict fertility rates. VERY IMPORTANT—Use the original quantitative female education level predictor, not the discretized variable in the model. We will also use the colon (`:`) notation to include the interaction term in the model. The colon implicitly creates the product term and includes it in the model.

```
# Fit model
lm.1 = lm(fertility_rate ~ 1 + educ_female + contraceptive + educ_female:contraceptive, data = fert)

# Model-level output
glance(lm.1)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
1     0.709         0.701 0.758      97.2 5.59e-32     4  -139.  289.  303.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

```
# Coefficient-level output
tidy(lm.1)
```

```
# A tibble: 4 x 5
  term                      estimate std.error statistic  p.value
  <chr>                        <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)                6.84      0.354       19.3  1.27e-38
2 educ_female               -0.408     0.0580      -7.04 1.29e-10
3 contraceptive             -0.0479    0.00768     -6.23 7.16e- 9
4 educ_female:contraceptive  0.00350   0.00100      3.50 6.63e- 4
```

The interaction model explains 70.9% of the variation in fertility rates which is not very consistent with the null hypothesis that the model explains no variation ($F_{3,120} = 97.23$, $p < .001$).

To examine whether there is an interaction effect, we evaluate the evidence from the coefficeint-level output. The *p*-value associated with the interaction term ($p = .0007$) suggests that the empirical evidence is inconsistent with the hypothesis of no interaction effect. We do believe that there is an interaction between contraceptive use and female level of education. Because the interaction effect is statistically relevant, we do not interpret any of the constituent main-effects in the model.

Plot of the Interaction Model

To further understand the nature of the interaction, we will create a plot of the effect of contraception use on fertility rates for different levels of female education. (Or, you could create a plot of the effect of female education level on fertility rates for different values of contraception use.) Here I will choose female education values of 5, 8, and 10 (the nearest integer values to the 25th, 50th, and 75th percentile values) to create my plot. First we will substitute these values into the fitted equation to find the partial equation for each selected female education level.

Female education level of 5

$$
\begin{aligned}
\hat{\text{Fertility rate}}_i &= 6.84 - 0.41(5) - 0.05(\text{Contraceptive use}_i) + 0.004(5)(\text{Contraceptive use}_i) \\
&= 6.84 - 2.05 - 0.05(\text{Contraceptive use}_i) + 0.02(\text{Contraceptive use}_i) \\
&= 4.79 - 0.03(\text{Contraceptive use}_i)
\end{aligned}
$$

Female education level of 8

$$
\begin{aligned}
\hat{\text{Fertility rate}}_i &= 6.84 - 0.41(8) - 0.05(\text{Contraceptive use}_i) + 0.004(8)(\text{Contraceptive use}_i) \\
&= 6.84 - 3.28 - 0.05(\text{Contraceptive use}_i) + 0.032(\text{Contraceptive use}_i) \\
&= 3.56 - 0.018(\text{Contraceptive use}_i)
\end{aligned}
$$

Female education level of 10

$$
\begin{aligned}
\widehat{\text{Fertility rate}}_i &= 6.84 - 0.41(10) - 0.05(\text{Contraceptive use}_i) + 0.004(10)(\text{Contraceptive use}_i) \\
&= 6.84 - 4.1 - 0.05(\text{Contraceptive use}_i) + 0.04(\text{Contraceptive use}_i) \\
&= 2.74 - 0.01(\text{Contraceptive use}_i)
\end{aligned}
$$

Now we can create out plot of fertility rates versus contraception use and add the three partial regression lines.

```r
# Plot the fitted model
ggplot(data = fert, aes(x = contraceptive, y = fertility_rate)) +
  geom_point(alpha = 0) +
  geom_abline(intercept = 4.79, slope = -0.03, color = "#095b67", linetype = "dotted") +
  geom_abline(intercept = 3.56, slope = -0.018, color = "#d82f5a", linetype = "dashed") +
  geom_abline(intercept = 2.74, slope = -0.01, color = "#8dd444", linetype = "solid") +
  theme_bw() +
  xlab("Contraceptive use") +
  ylab("Fertility rate")
```

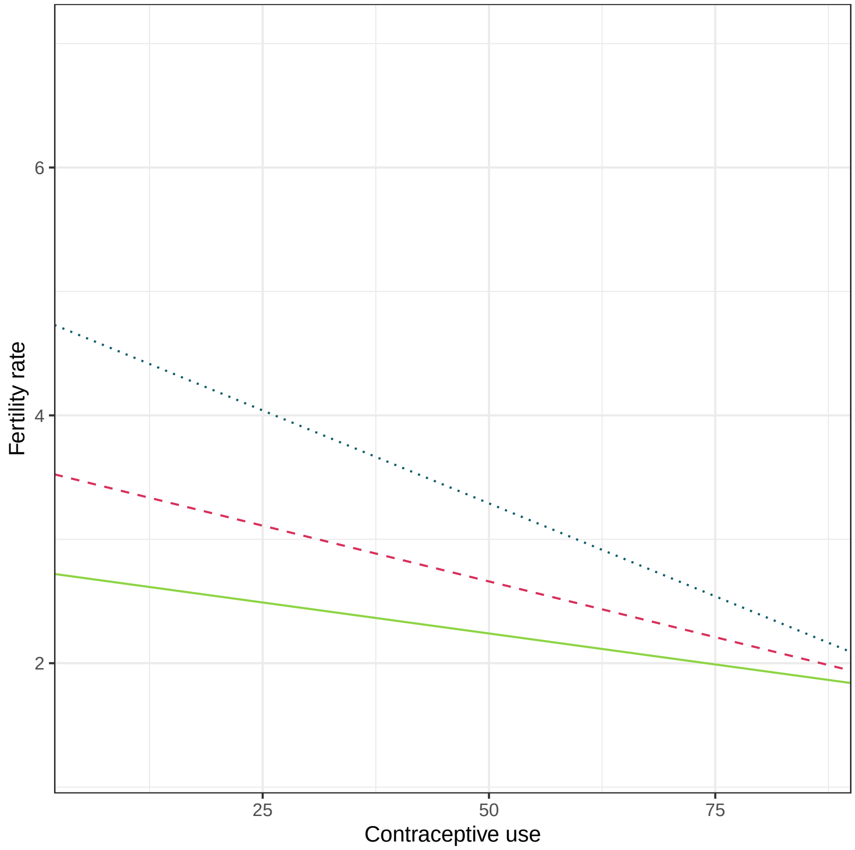


Figure 1. Plot of fertility rate as a function of contraceptive use and female education level. Partial regression lines are displayed for female education levels of 5th grade (blue, dotted line), 8th grade (red, dashed line) and 10th grade (green, solid line).

Based on the plot, we can see there is an *ordinal interaction* between contraceptive use and female education level (the lines do not cross in our plot). The effect of contraceptive use on fertility rate varies by level of female education. The largest effect of contraceptive use on fertility rate (highest absolute slope) is for countries that have the lowest female education level. This effect diminishes for countries with higher levels of female education (the magnitude of the slopes get smaller).

Similarly, the effect of female education level on fertility rate varies for different levels of contraceptive use. For countries with low rates of contraceptive use, there are large differences between the predicted average fertility rates by female education level (the distance between the lines is large). These differences diminish for countries with higher levels of contraceptive use.

Interpreting the Individual Effects from the tidy() Output

In practice, it is enough to say there is an interaction, and to use the plot of the results to interpret the nature of the interaction effects rather than to interpret the actual coefficient values estimated by the `lm()` function. This being said, in simple models, we can actually interpret the coefficients more directly. To do this, write out the fitted equations for countries that differ in female education level by 1 year. We will write the fitted equations for countries that have a female education level of 0 years and those with a female education level of 1 year. (Do the substitution yourself to verify these equations.)

$$\begin{aligned}
\textbf{0 years}: &\quad \hat{\text{Fertility rate}}_i = 6.84 - 0.048(\text{Contraceptive use}_i) \\
\textbf{1 year}: &\quad \hat{\text{Fertility rate}}_i = [6.84 - 0.408] + [-0.0479 + 0.00350]\,(\text{Contraceptive use}_i)
\end{aligned}$$

- The intercept (6.84) is the average fertility rate for countries with a female education level of 0 years and contraception use of 0. (extrapolation).
- The coefficient associated with contraceptive use (-0.048) is the effect of contraceptive use on fertility rate for countries with a female education level of 0 years.
- The coefficient associated with female education (-0.408) is the difference in average fertility rates between countries with contraceptive use = 0 and countries with contraceptive use = 1. Alternatively, it is the difference in intercepts between countries whose female education level differs by one year.
- The coefficient associated with the interaction term (0.00350) is the difference in slopes (effect of contraceptive use on fertility rate) between countries whose female education level differs by one year.

It cannot be iterated enough that although we can interpret the coefficients directly, in practice, the plot of the interaction model is much more informative and far less complicated for readers to understand.

Adding Covariates

Is there an interaction between contraception use and female education level on fertility rates after we control for differences in infant mortality rate?

```
# Fit model
lm.2 = lm(fertility_rate ~ 1 + educ_female + contraceptive + infant_mortality + educ_female:contraceptive, data =
# Coefficient-level output
tidy(lm.2)
```

```
# A tibble: 5 x 5
  term                      estimate std.error statistic  p.value
  <chr>                        <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)                4.82     0.574         8.41 1.05e-13
2 educ_female               -0.271    0.0629       -4.31 3.45e- 5
3 contraceptive             -0.0317   0.00811      -3.91 1.55e- 4
4 infant_mortality           0.0212   0.00494       4.30 3.55e- 5
5 educ_female:contraceptive  0.00246  0.000966      2.55 1.20e- 2
```

To examine whether there is an interaction effect, we again evaluate the evidence from the coefficient-level output. Based on the *p*-value associated with the interaction term ($p = .012$), the empirical evidence is consistent with there being an interaction effect between contraceptive use and female level of education after controlling for differences in countrys' infant mortality rates.

How would we interpret the effects in the model?

- **Female education level:** Since this effect is part of an interaction term, we would interpret the interaction, namely, controlling for differences in infant mortality rates, the effect of female education level on fertility rates depends on contraceptive use.
- **Contraceptive use:** Since this effect is part of an interaction term, we would only interpret the interaction, namely, controlling for differences in infant mortality rates, the effect of contraceptive use on fertility rates depends on female education level.
- **Infant mortality:** Since this is not part of any interaction term, we interpret this as we would any other effect included in a multiple regression model, anmely, controlling for differences in female education level and contraceptive use, each one-perecentage point difference in infant mortality rate is associated with a 0.02-unit difference in fertility rates, on average.

To better understand the nature of the effects, especially the interaction, we will plot the fitted model. We choose female education values of 5, and 10 (the nearest integer values to the 25th and 75th percentile values) to create my plot. We also choose infant mortality values of 7 and 41 (the nearest integer values to the 25th and 75th percentile values) to show the effect of infant mortality rate.

Female education level of 5; Infant mortality rate of 7

$$\begin{aligned}
\hat{\text{Fertility rate}}_i &= 4.82 - 0.271(5) - 0.03(\text{Contraceptive use}_i) + 0.02(7) + 0.002(5)(\text{Contraceptive use}_i) \\
&= 3.61 - 0.02(\text{Contraceptive use}_i)
\end{aligned}$$

Female education level of 5; Infant mortality rate of 41

$$\begin{aligned}
\hat{\text{Fertility rate}}_i &= 4.82 - 0.271(5) - 0.03(\text{Contraceptive use}_i) + 0.02(41) + 0.002(5)(\text{Contraceptive use}_i) \\
&= 4.29 - 0.02(\text{Contraceptive use}_i)
\end{aligned}$$

Female education level of 10; Infant mortality rate of 7

$$\begin{aligned}
\hat{\text{Fertility rate}}_i &= 4.82 - 0.271(10) - 0.03(\text{Contraceptive use}_i) + 0.02(7) + 0.002(10)(\text{Contraceptive use}_i) \\
&= 2.25 - 0.01(\text{Contraceptive use}_i)
\end{aligned}$$

Female education level of 10; Infant mortality rate of 41

$$\begin{aligned}\widehat{\text{Fertility rate}}_i &= 4.82 - 0.271(10) - 0.03(\text{Contraceptive use}_i) + 0.02(41) + 0.002(10)(\text{Contraceptive use}_i) \\ &= 2.93 - 0.01(\text{Contraceptive use}_i)\end{aligned}$$

Now we can create out plot of fertility rates versus contraception use and add the four partial regression lines.

```
# Plot the fitted model (infant mortality rate = 7)
p1 = ggplot(data = fert, aes(x = contraceptive, y = fertility_rate)) +
  geom_point(alpha = 0) +
  geom_abline(intercept = 3.61, slope = -0.02, color = "#095b67", linetype = "solid") +
  geom_abline(intercept = 2.25, slope = -0.01, color = "#d82f5a", linetype = "dashed") +
  theme_bw() +
  xlab("Contraceptive use") +
  ylab("Fertility rate") +
  ggtitle("Infant mortality rate at the first quartile (7%)")

# Plot the fitted model (infant mortality rate = 41)
p2 = ggplot(data = fert, aes(x = contraceptive, y = fertility_rate)) +
  geom_point(alpha = 0) +
  geom_abline(intercept = 4.29, slope = -0.02, color = "#095b67", linetype = "solid") +
  geom_abline(intercept = 2.93, slope = -0.01, color = "#d82f5a", linetype = "dashed") +
  theme_bw() +
  xlab("Contraceptive use") +
  ylab("Fertility rate") +
  ggtitle("Infant mortality rate at the third quartile (41%)")

# Layout side-by-side plot
grid.arrange(p1, p2, nrow = 1)
```

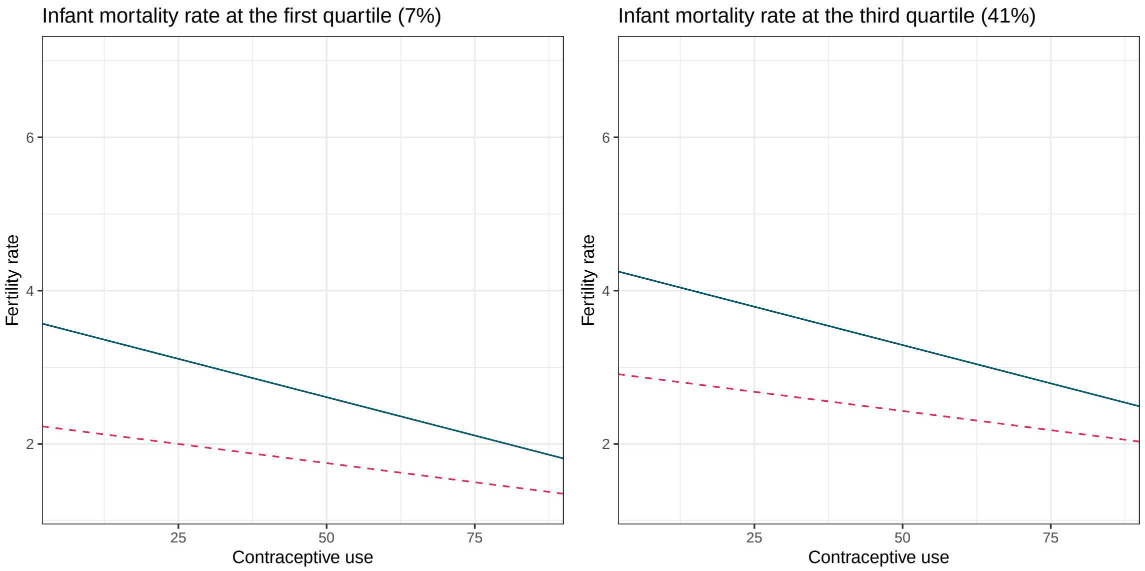


Figure 2. Plot of fertility rate as a function of contraceptive use, female education level, and infant mortality rate. Partial regression lines are displayed for female education levels of 5th grade (blue, solid line), 8th grade (red, dashed line) for countries with infant mortality rates at the first (7%) and third (41%) quartile values.

Higher Order Interactions

Interactions between two predictors (e.g., female education level and contraceptive use) are referred to as *first order* interactions. In the previous section, the model we fitted included a main-effect of infant mortality rate and a first order interaction between female education level and contraceptive use. The main-effect of infant mortality rate in this model suggested that the first order interaction between female education level and contraceptive use was THE SAME for every level of infant mortality rate.

We could also fit a model that posits that the first order interaction between female education level and contraceptive use IS DIFFERENT for different levels of infant mortality rate. This is technically an interaction between infant mortality rate and the first order interaction between female education level and contraceptive use. It is an interaction of an interaction. This is called a *second order* interaction.

To fit such a model, we would need to include the second order interaction between infant mortality rate, female education level and contraceptive use; the product of the three main effects. Since we are including an interaction, we need to include all three constituent main effects AND since it is a higher order interaction, we need to include all constituent lower order interactions; in this case all constituent first order interactions. As such the predictors would include:

- **Main-Effects:**
  - `educ_female`
  - `contraceptive`
  - `infant_mortality`
- **First Order Interactions:**
  - `educ_female:contraceptive`
  - `educ_female:infant_mortality`
  - `contraceptive:infant_mortality`
- **Second Order Interaction:**
  - `educ_female:contraceptive:infant_mortality`

We fit the model below.

```r
# Fit model
lm.3 = lm(fertility_rate ~ 1 + educ_female + contraceptive + infant_mortality +
            educ_female:contraceptive + infant_mortality:contraceptive + educ_female:infant_mortality +
            educ_female:contraceptive:infant_mortality, data = fert)

# Coefficient-level output
tidy(lm.3)
```

```
# A tibble: 8 x 5
  term                                     estimate std.error statistic  p.value
  <chr>                                       <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)                               7.02e+0 1.02           6.86 3.52e-10
2 educ_female                              -5.12e-1 0.124         -4.14 6.51e- 5
3 contraceptive                            -7.54e-2 0.0168        -4.47 1.80e- 5
4 infant_mortality                         -1.82e-2 0.0138        -1.32 1.90e- 1
5 educ_female:contraceptive                 7.06e-3 0.00185        3.81 2.21e- 4
6 contraceptive:infant_mortality            9.62e-4 0.000325       2.96 3.74e- 3
7 educ_female:infant_mortality              4.55e-3 0.00285        1.59 1.14e- 1
8 educ_female:contraceptive:infant_mortal~ -1.04e-4 0.0000516     -2.01 4.69e- 2
```

The fitted equation is:

$$
\begin{aligned}
\hat{\text{Fertility rate}}_i = 7.02 &- 0.512(\text{Female education}_i) - 0.08(\text{Contraceptive use}_i) - 0.02(\text{Infant mortality}_i) + \\
&0.007(\text{Female education}_i)(\text{Contraceptive use}_i) + 0.001(\text{Contraceptive use}_i)(\text{Infant mortality}_i) + \\
&0.005(\text{Female education}_i)(\text{Infant mortality}_i) \\
&- 0.0001(\text{Female education}_i)(\text{Contraceptive use}_i)(\text{Infant mortality}_i)
\end{aligned}
$$

The *p*-value associated with the second order interaction term ($p = .0469$) suggests that there is a second order intraction effect between female education level, contraceptive use, and infant mortality rate on fertility rates. To interpret this, plot the model results. Again, pick values for female education level and infant mortality rate, substitute them into the fitted equation, and reduce it. I again used 5 and 10 for female education level and 7 and 41 for infant mortality rate. (Note: Algebra not shown.)

```r
# Plot the fitted model (infant mortality rate = 7)
p1 = ggplot(data = fert, aes(x = contraceptive, y = fertility_rate)) +
  geom_point(alpha = 0) +
  geom_abline(intercept = 4.49, slope = -0.04, color = "#095b67", linetype = "solid") +
  geom_abline(intercept = 2.09, slope = -0.005, color = "#d82f5a", linetype = "dashed") +
  theme_bw() +
  xlab("Contraceptive use") +
  ylab("Fertility rate") +
  ggtitle("Infant mortality rate at the first quartile (7%)")

# Plot the fitted model (infant mortality rate = 41)
p2 = ggplot(data = fert, aes(x = contraceptive, y = fertility_rate)) +
  geom_point(alpha = 0) +
  geom_abline(intercept = 4.65, slope = -0.02, color = "#095b67", linetype = "solid") +
  geom_abline(intercept = 3.02, slope = -0.008, color = "#d82f5a", linetype = "dashed") +
  theme_bw() +
  xlab("Contraceptive use") +
  ylab("Fertility rate") +
  ggtitle("Infant mortality rate at the third quartile (41%)")

# Layout side-by-side plot
grid.arrange(p1, p2, nrow = 1)
```

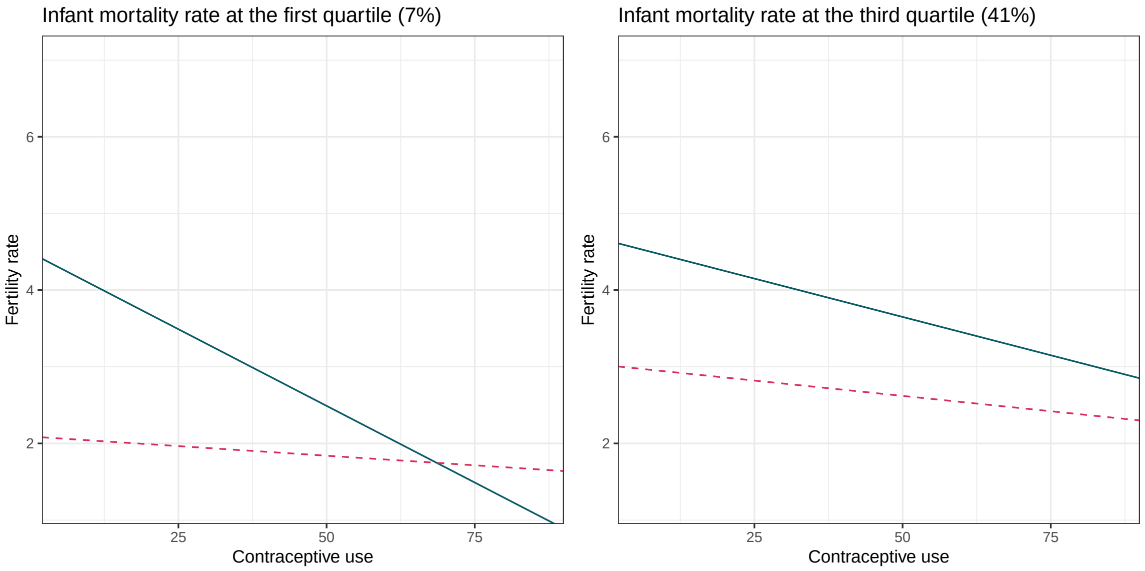


Figure 3. Plot of fertility rate as a function of contraceptive use, female education level, and infant mortality rate. Partial regression lines are displayed for female education levels of 5th grade (blue, solid line), 8th grade (red, dashed line) for countries with infant mortality rates at the first (7%) and third (41%) quartile values.

- The plots show that the interaction between contraceptive use and female education level on fertility rates DIFFERS by infant mortality rate.
- This also suggests that the interaction between contraceptive use and infant mortality rate on fertility rates DIFFERS by level of female education.
- Finally, it also implies that the interaction between female education level and infant mortality rate on fertility rates DIFFERS by level of contraceptive use.

Some Advice for Fitting Interaction Models

In general, only fit interaction terms that include focal predictors. Do not fit interaction terms that are composed of all control predictors. This has the implication that if you do not have a focal predictor (i.e., the analysis is purely exploratory) you should probably not fit interaction terms.

A second piece of advice is that unless there is specific theoretical reason to fit higher order interactions with your focal predictors, avoid them. This also is good advice for first order interaction terms as well.