Understanding Statistical Control

2019-07-05

Introduction and Research Question

In this set of notes, we will return again to examine the question of whether time spent on homework is related to GPA using the *keith-gpa.csv* data (see the data codebook). However this time, we will control for parent education level. To begin, we will load several libraries and import the data into an object called `keith`.

Preparation

```
# Load libraries
library(broom)
library(corrr)
library(dotwhisker)
library(dplyr)
library(ggplot2)
library(readr)
library(tidyr)

keith = read_csv(file = "~/Documents/github/epsy-8251/data/keith-gpa.csv")
head(keith)
```

```
# A tibble: 6 x 3
    gpa homework parent_ed
  <dbl>    <dbl>     <dbl>
1    78        2        13
2    79        6        14
3    79        1        13
4    89        5        13
5    82        3        16
6    77        4        13
```

Initial exploration (not shown) suggests that the distribution of parent education values is slightly right-skewed. The correlation matrix for the variables used in the analysis is shown in Table 1. The means and standard deviations for each variable are shown in the main diagonal.

Table 1
Correlations between Three Student Attributes. Means and Standard Deviations are Displayed on the Main Diagonal

Attribute	1	2	3
1. GPA	80.47 (7.62)	—	—
2. Time spent on homework	.33	5.09 (2.06)	—
3. Parent education level	.29	.28	14.03 (1.93)

Both time spent on homework and parent education level are moderately and positively correlated with student GPA in the sample. Moreover, time spent on homework is also moderately and positively correlated with parent education level.

```r
# Fit regression models to examine effect of time spent on homework
lm.1 = lm(gpa ~ 1 + homework, data = keith)
lm.2 = lm(gpa ~ 1 + homework + parent_ed, data = keith)
```

Table 2 shows the results from fitting a series of models to examine the effect of time spent on homework on student GPA. The results from Model 1 are consistent with time spent on homework having a positive association with GPA ($p < .001$). Each one hour difference in time spent on homework is associated with a 1.21-point difference in GPA, on average. This positive association is seen even after controlling for parent education level (see Model 2; $p = .026$), although the effect is somewhat smaller, with each one hour difference in time spent on homework is associated with a 0.98-point difference in GPA, on average.

Table 2
Taxonomy of OLS Regression Models Fitted to Explore the Effect of Time Spent on Homework on GPA for 100 8th-Grade Students

	Model 1	Model 2
Time spent on homework	1.21	0.99
	(0.35)	(0.36)
Parent education level		0.87
		(0.38)
Constant	74.29	63.2
	(1.94)	(5.24)
R^2	0.107	0.152
RMSE	7.24	7.09

Note. RMSE = Root mean squared error.

Understanding Statistical Control via Predicted Values

The fitted equation for Model 2 is,

$$\hat{\text{GPA}}_i = 63.22 + 0.99(\text{Homework}_i) + 0.87(\text{Parent Education}_i)$$

Let's predict the average GPA for students who spend differing amounts of time on homework,

- Time spent on homework = 1 hour
- Time spent on homework = 2 hours
- Time spent on homework = 3 hours

Let's also assume that these student all have parent education level of 12 years.

Table 3
Predicted GPA for Students with Parent Education of 12 and Spend Varying Amounts of Time on Homework

Homework	Parent Education	Model Predicted GPA
1	12	63.22 + 0.99(1) + 0.87(12) = 74.65
2	12	63.22 + 0.99(2) + 0.87(12) = 75.64
3	12	63.22 + 0.99(3) + 0.87(12) = 76.63

In this example, the value of `parent_ed` is "constant" across the three types of students. Time spent on homework differs by one-hour between each subsequent type of student. The difference in model predicted average GPA between these students is 0.99. When we hold level of parent education constant, the predicted difference in average GPA between students who spend an additional hour on homework is 0.99.

What if we had chosen a parent education level of 13 years instead?

Table 4
Predicted GPA for Students with Parent Education of 13 and Spend Varying Amounts of Time on Homework

Homework	Parent Education	Model Predicted GPA
1	13	63.22 + 0.99(1) + 0.87(13) = 75.52
2	13	63.22 + 0.99(2) + 0.87(13) = 76.51
3	13	63.22 + 0.99(3) + 0.87(13) = 77.50

The model predicted average GPAS are higher for these students because they have a higher parent education level. But, again, when we hold parent education level constant, the predicted difference in average GPA between students who spend an additional hour on homework is 0.99.

By fixing the value of parent level of education to a particular value (holding it constant) we can "fairly" compare the average predicted GPA for different values of time spent on homework. This gives us a "truer" picture of the association between time spent on homework and GPA since we don't have to worry that the GPAs we are comparing have different values for parent level of education. By holding that variable constant, we remove it as a potential confounding source and leave time spent on homework as the only reason (aside from random error) the GPAs vary. Moreover, this difference in average GPAs for any one hour difference in time spent on homework will be 0.99, regardless of the value we pick for parent level of education.

Understanding Statistical Control via the Fitted Model

Let us return to the fitted equation for Model 2,

$$\widehat{\text{GPA}}_i = 63.22 + 0.99(\text{Homework}_i) + 0.87(\text{Parent Education}_i)$$

But this time, instead of computing predicted values, let's focus on the fitted equation for students with a specified parent education level, say 12 years. We can substitute this value into the fitted equation and reduce the result.

$$\begin{aligned}\widehat{\text{GPA}}_i &= 63.22 + 0.99(\text{Homework}_i) + 0.87(12) \\ &= 63.22 + 0.99(\text{Homework}_i) + 10.44 \\ &= 73.66 + 0.99(\text{Homework}_i)\end{aligned}$$

By substituting in a constant value for parent education level, we can write the model so that GPA is a function of time spent on homework. Interpreting the coefficients,

- Students with a parent education level of 12 years and who spend 0 hours a week on homework are predicted to have a mean GPA of 73.66.
- For students with a parent education level of 12 years, each additional hour spent on homework is associated with a 0.99-pt difference in GPA, on average.

What about the students whose parent education level is 13? Substituting this value into the fitted equation and reducing the result, we get,

$$\begin{aligned}\hat{\text{GPA}}_i &= 63.22 + 0.99(\text{Homework}_i) + 0.87(13) \\ &= 63.22 + 0.99(\text{Homework}_i) + 11.31 \\ &= 74.53 + 0.99(\text{Homework}_i)\end{aligned}$$

Interpreting these coefficients,

- Students with a parent education level of 13 years and who spend 0 hours a week on homework are predicted to have a mean GPA of 74.53.
- For students with a parent education level of 13 years, each additional hour spent on homework is associated with a 0.99-pt difference in GPA, on average.

The key here is that the slope for these two sets of students is the same. The relationship between time spent on homework and GPA is exactly the same regardless of parent education level.

Understanding Statistical Control via the Plot of the Fitted Model

To create a plot that helps us interpret the results of a multiple regression analysis, we pick fixed values for all but one of the predictors and substitute those into the fitted equation. We can then rewrite the equation and use `geom_abline()` to draw the fitted line. Below I illustrate this by choosing three fixed values for parent level of education (namely 8, 12, and 16) and rewriting the three equations.

Parent education level = 8

$$\begin{aligned}\hat{\text{GPA}}_i &= 63.22 + 0.99(\text{Homework}_i) + 0.87(8) \\ &= 70.18 + 0.99(\text{Homework}_i)\end{aligned}$$

Parent education level = 12

$$\begin{aligned}\hat{\text{GPA}}_i &= 63.22 + 0.99(\text{Homework}_i) + 0.87(12) \\ &= 73.66 + 0.99(\text{Homework}_i)\end{aligned}$$

Parent education level = 16

$$\begin{aligned}\hat{\text{GPA}}_i &= 63.22 + 0.99(\text{Homework}_i) + 0.87(16) \\ &= 77.14 + 0.99(\text{Homework}_i)\end{aligned}$$

Now I will create a plot of the outcome versus the predictor we left as a variable in the three equations (time spent on homework) and use `geom_abline()` to include the line for each of the three rewritten equations. Note that since I have three different equations (one for each of the three parent education levels), I will need to include three layers of `geom_abline()` in the plot syntax.

```
ggplot(data = keith, aes(x = homework, y = gpa)) +
  geom_point() +
  theme_bw() +
  xlab("Time spent on homework") +
  ylab("Model Ppredicted GPA") +
  geom_abline(intercept = 70.18, slope = 0.99) +
  geom_abline(intercept = 73.66, slope = 0.99) +
  geom_abline(intercept = 77.14, slope = 0.99)
```

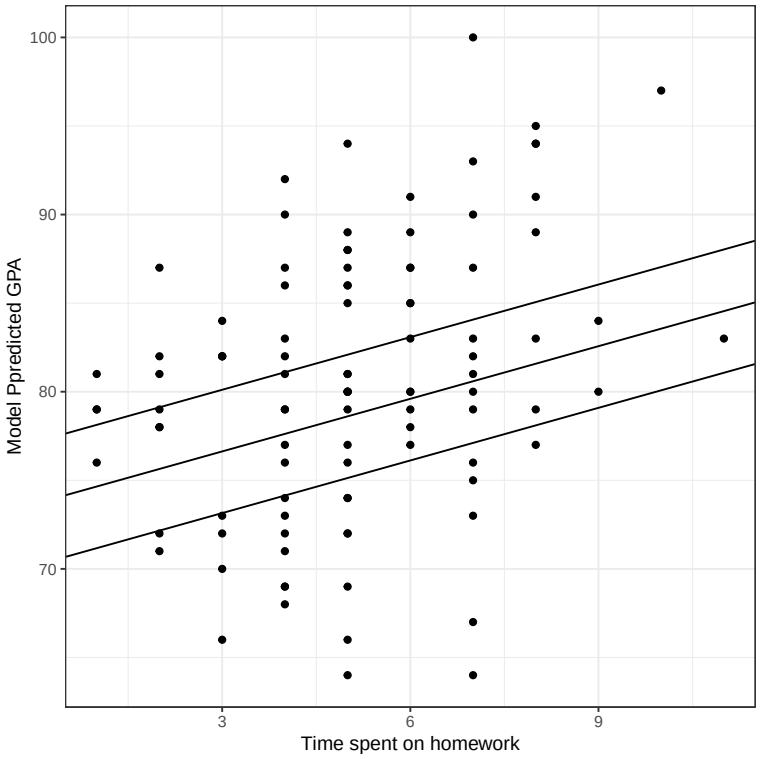

Figure 1. Model predicted GPA as a function of time spent on homework for students with a parent education level of 8, 12, and 16 years.

This plot is a good start, but has a couple issues. First, since the focus is on the fitted regression lines and not the observations, including the points may clutter the plot and obscure the message. Therefore we may want to make the observations much more transparent or remove them altogether. Second, although we know which line corresponds to which parent education level, others will not. Thus, we need to differentiate between the three lines.

```
ggplot(data = keith, aes(x = homework, y = gpa)) +
  geom_point(alpha = 0) +
  theme_bw() +
  xlab("Time spent on homework") +
  ylab("Model predicted GPA") +
  geom_abline(intercept = 70.18, slope = 0.99, color = "#46ACC8", linetype = "dotdash") +
  geom_abline(intercept = 73.66, slope = 0.99, color = "#E58601", linetype = "solid") +
  geom_abline(intercept = 77.14, slope = 0.99, color = "#B40F20", linetype = "dashed")
```

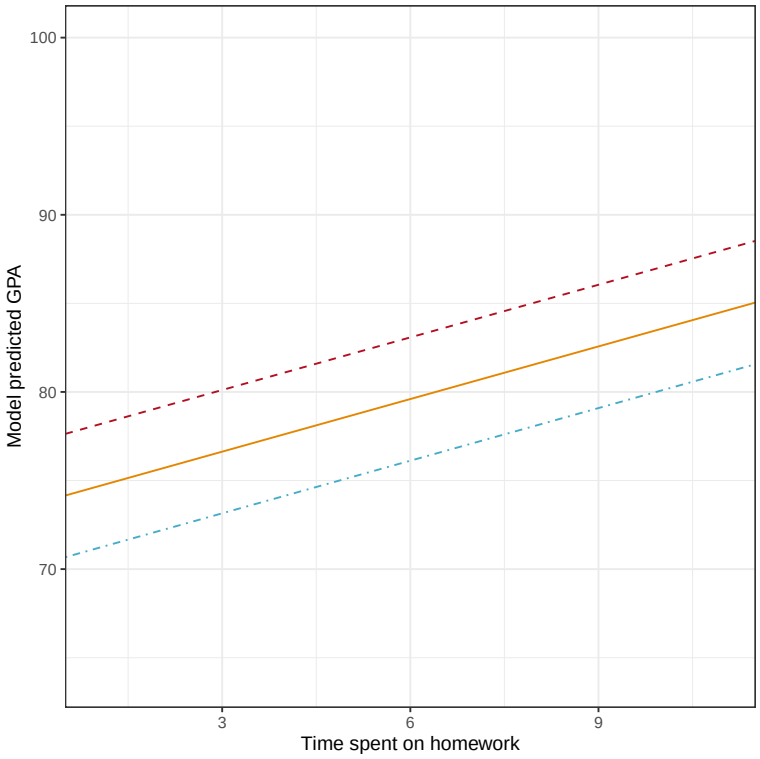

Figure 2. Model predicted GPA as a function of time spent on homework for students with a parent education level of 8 years (blue, dot-dashed line), 12 years (orange, solid line), and 16 years (red, dashed line).

From the plot we can see the effect of time spent on homework in the slopes of the fitted lines. Regardless of the level of parent education (8, 12, or 16), the slope of the line is 0.99, which means the three lines are parallel. The intercepts of these three lines vary reflecting the different level of parent education.

We can interpret the effect of parent level of education by fixing time spent on homework to a particular value on the same plot. For example, fixing time spent on homework to 6, we see that the average GPA varies for the three levels of parent education displayed in the plot.

```
ggplot(data = keith, aes(x = homework, y = gpa)) +
  geom_point(alpha = 0) +
  theme_bw() +
  xlab("Time spent on homework") +
  ylab("Model predicted GPA") +
  geom_abline(intercept = 70.18, slope = 0.99, color = "#46ACC8", linetype = "dotdash") +
  geom_abline(intercept = 73.66, slope = 0.99, color = "#E58601", linetype = "solid") +
  geom_abline(intercept = 77.14, slope = 0.99, color = "#B40F20", linetype = "dashed") +
  geom_point(x = 6, y = 76.11908, color = "#46ACC8", size = 2) +
  geom_point(x = 6, y = 79.60157, color = "#E58601", size = 2) +
  geom_point(x = 6, y = 83.08407, color = "#B40F20", size = 2)
```

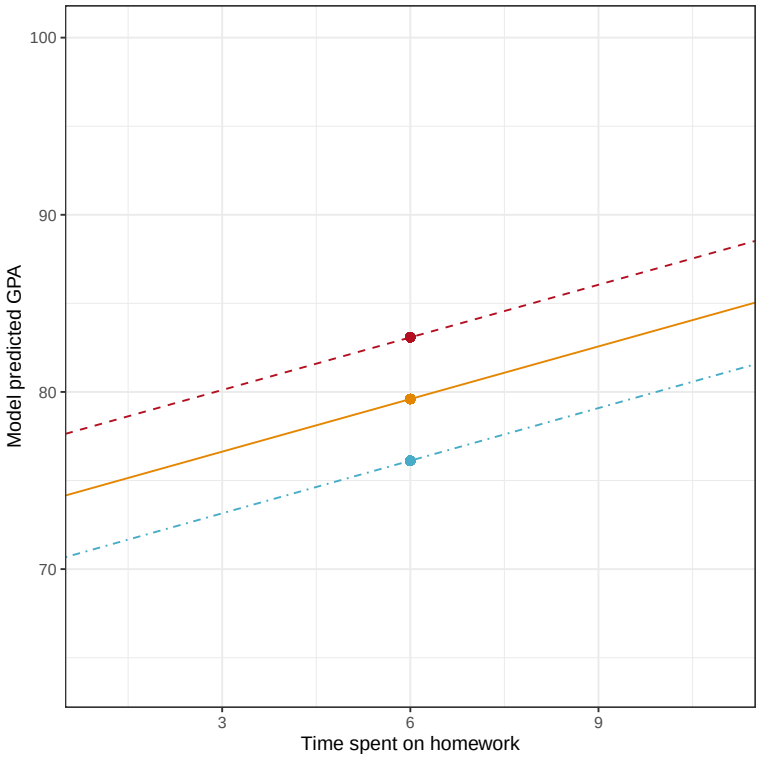

Figure 3. Model predicted GPA as a function of time spent on homework for students with a parent education level of 8 years (blue, dot-dashed line), 12 years (orange, solid line), and 16 years (red, dashed line). The model predicted GPAs for students who spend six hours a week on homework are also displayed.

How much do the model predicted GPAs vary for these three parent education levels?

Table 5
Predicted GPA for Students who spend six hours a week on homework with varying Parent Education Levels

Homework	Parent Education	Model Predicted GPA
6	8	76.12
6	12	79.60
6	16	83.08

The difference between each of these subsequent model predicted GPA values is 3.48. This is constant because we chose parent education levels that differ by the same amount, in this case each value of parent education differs by four years.

Effect of Parent Education Level

What if we would have chosen parent education levels that differed by one year rather than by four years?

Table 6
Predicted GPA for Students who spend six hours a week on homework with varying Parent Education Levels

Homework	Parent Education	Model Predicted GPA
6	8	76.12
6	9	76.99
6	10	77.86

Now a one-year difference in parent education level is associated with a 0.87-point difference in predicted GPA, *holding time spent on homework constant*. We could also have calculated this directly from the earlier result. Since a four-year difference in parent education is associated with a 3.48-point difference in predicted GPA, a one-year difference in parent education is associated with a $^{3.48}/_{4} = 0.87$-point difference in predicted GPA. This algebra works since the relationship is constant (i.e., linear).

Triptych Plots: Displaying the Results from a Multiple Regression Model

Remember, to create a plot that helps us interpret the results of a multiple regression analysis, we pick fixed values for all but one of the predictors and substitute those into the fitted equation. We can then rewrite the equation and use `geom_abline()` to draw the fitted lines. We illustrated this earlier by choosing three fixed values for parent level of education (namely 8, 12, and 16) and rewriting the three equations:

$$\begin{aligned}
\textbf{Parent Education} = \textbf{8}: &\quad \hat{\text{GPA}}_i = 70.18 + 0.99(\text{Homework}_i) \\
\textbf{Parent Education} = \textbf{12}: &\quad \hat{\text{GPA}}_i = 73.66 + 0.99(\text{Homework}_i) \\
\textbf{Parent Education} = \textbf{16}: &\quad \hat{\text{GPA}}_i = 77.14 + 0.99(\text{Homework}_i)
\end{aligned}$$

The plot we created earlier put all three fitted lines on the same plot. An alternative plot is to show each line in a different plot, and to place this plots side-by-side in a "triptych". (Note: I borrowed this terminology from Richard McElreath.) To do this we save each plot into an object and then use the `grid.arrange()` function from the **gridExtra** package to put the plots side-by-side.

```
# Load package
library(gridExtra)

# Create plot 1
p1 = ggplot(data = keith, aes(x = homework, y = gpa)) +
      geom_point(alpha = 0) +
      geom_abline(intercept = 70.18, slope = 0.99) +
      theme_bw() +
      xlab("Time spent on homework") +
      ylab("Model predicted GPA") +
      ggtitle("Parent Education = 8 Years")

# Create plot 2
p2 = ggplot(data = keith, aes(x = homework, y = gpa)) +
      geom_point(alpha = 0) +
```

```
      geom_abline(intercept = 73.66, slope = 0.99) +
      theme_bw() +
      xlab("Time spent on homework") +
      ylab("Model predicted GPA") +
      ggtitle("Parent Education = 12 Years")

# Create plot 3
p3 = ggplot(data = keith, aes(x = homework, y = gpa)) +
      geom_point(alpha = 0) +
      geom_abline(intercept = 77.14, slope = 0.99) +
      theme_bw() +
      xlab("Time spent on homework") +
      ylab("Model predicted GPA") +
      ggtitle("Parent Education = 16 Years")

# Put plots side-by-side
grid.arrange(p1, p2, p3, nrow = 1)
```

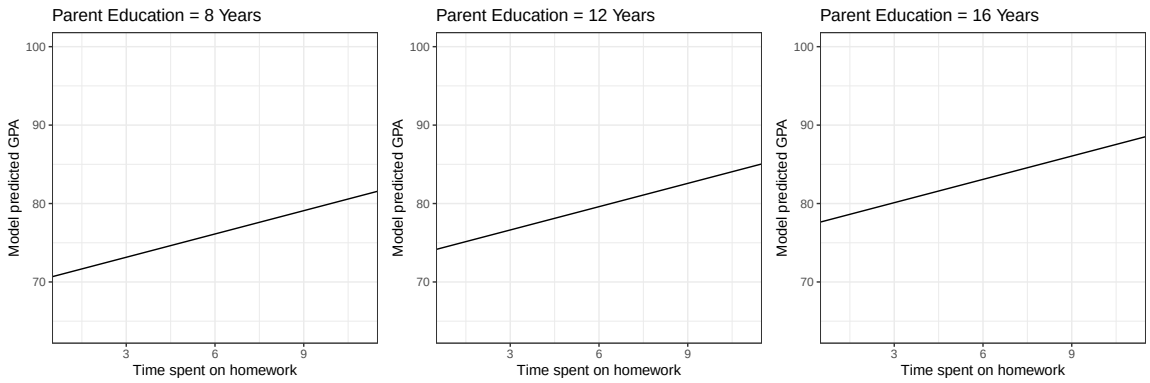

Figure 4. Model predicted GPA as a function of time spent on homework for students with a parent education level of 8, 12, and 16 years.

Emphasis on the Effect of Parent Level of Education

With two (or more) effects in the model we have more than one potential way to display the fitted results. In general, we will display one predictor through the slope of the line plotted (same as we did with only one predictor), and EVERY OTHER predictor will be shown through one or more lines. In the previous plots, below, we have dsiplayed the effect of time spent on homework (on the x-axis) through the slope of the lines, and the effect of parent level of education through the vertical distance between the three different lines.

In general, effect seen via the slope of the line is more cognitively apparent than the vertical distance between different lines. Thus whichever effect you want to emphasize should be placed on the x-axis; or left as a variable when you algebraically simplify the fitted regression equation.

For example, what if we wantd to emphasize parent level of education? In that case, we would choose fixed values for time spent on homework, substitute these into the fitted equation, and simplify. Here we choose fixed values for time spent on homework of 2, 5, and 10 hours. Rewriting the three equations:

$$\textbf{Homework} = \textbf{2}: \quad \hat{\text{GPA}}_i = 65.18 + 0.87(\text{Parent Education}_i)$$
$$\textbf{Homework} = \textbf{5}: \quad \hat{\text{GPA}}_i = 68.14 + 0.87(\text{Parent Education}_i)$$
$$\textbf{Homework} = \textbf{10}: \quad \hat{\text{GPA}}_i = 73.08 + 0.87(\text{Parent Education}_i)$$

Then we create the triptych plot showing the resulting equations:

```
# Create plot 1
p4 = ggplot(data = keith, aes(x = parent_ed, y = gpa)) +
      geom_point(alpha = 0) +
      geom_abline(intercept = 65.18, slope = 0.87) +
      theme_bw() +
      xlab("Parent education (in years)") +
      ylab("Model predicted GPA") +
      ggtitle("Time Spent on Homework = 2 Hours")

# Create plot 2
p5 = ggplot(data = keith, aes(x = parent_ed, y = gpa)) +
      geom_point(alpha = 0) +
      geom_abline(intercept = 68.14, slope = 0.87) +
      theme_bw() +
      xlab("Parent education (in years)") +
      ylab("Model predicted GPA") +
      ggtitle("Time Spent on Homework = 5 Hours")

# Create plot 3
p6 = ggplot(data = keith, aes(x = parent_ed, y = gpa)) +
      geom_point(alpha = 0) +
      geom_abline(intercept = 73.08, slope = 0.87) +
      theme_bw() +
      xlab("Parent education (in years)") +
      ylab("Model predicted GPA") +
      ggtitle("Time Spent on Homework = 10 Hours")

# Put plots side-by-side
grid.arrange(p4, p5, p6, nrow = 1)
```

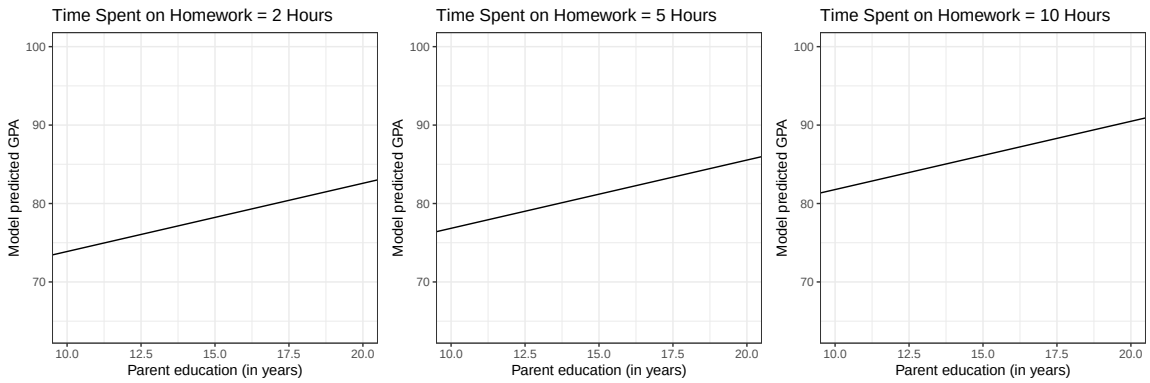

Figure 5. Model predicted GPA as a function of parent education level for students who spend 2 hours, 5 hours, and 10 hours a week on homework.

Only Displaying a Single Effect

Sometimes you only want to show the effect of a single predictor from the model. For example, in educational studies we often control for SES and mother's level of education when we fit the model, but we don't want to display those effects in our plot. There is no rule that just because you included an effect in the fitted model that you are obligated to display it.

Any effect that you do not want to display graphically can be fixed to a single value, typically the mean value. Fixing the effect to a single value will produce only one line. For example, here we set the parent education value to the its mean value of 14.03 years. After substituting this value into the fitted equation, this results in,

$$\hat{\text{GPA}}_i = 75.43 + 0.99(\text{Homework}_i)$$

Plotting this we get a single line which displays the effect of time spent on homework on GPA. Even though the effect of parent education is not displayed, it is still included as the intercept value of the plotted line is based on fixing this effect to its mean.

```
ggplot(data = keith, aes(x = homework, y = gpa)) +
 geom_point(alpha = 0) +
 geom_abline(intercept = 75.43, slope = 0.99) +
 theme_bw() +
 xlab("Time spent on homework") +
 ylab("Model predicted GPA")
```

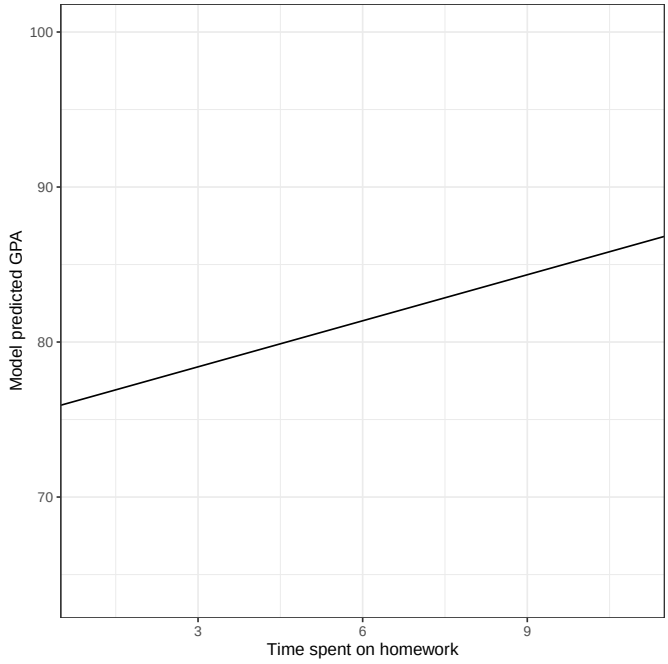

Figure 6. Model predicted GPA as a function of time spent on homework for students with an average parent education level (14.03 years).