14

Manipulating Independent Variables in Linear Regression

Learning Objectives

In this chapter, you will learn:

- What interaction variables are and why they are used
- How to interpret the results of linear regressions that use interaction variables
- What quadratic variables are and why they are used
- How to interpret the results of linear regressions that use quadratic variables
- Why researchers transform skewed variables
- What a logarithmic transformation is
- How to interpret the results of linear regressions that use log-transformed variables

Introduction

Chapters 14 and 15 introduce more advanced regression techniques that are regularly used by social science researchers. The goal of these chapters is to explain why researchers use these more advanced techniques and to illustrate how to interpret the regression coefficients that are produced. In this chapter, you'll learn why and how researchers manipulate independent variables in linear regressions. In particular, you'll learn how interaction variables allow researchers to estimate different slope coefficients for different subgroups within a sample, and how quadratic variables allow researchers to model curvilinear relationships. You'll also learn how skewed variables can be transformed so that they become more normally distributed. These techniques allow researchers to build regression models that fit the data well and that more accurately reflect real-world relationships and social processes.

The research focus of this chapter is literacy skills. *Literacy* refers to people's ability to understand, evaluate, use, and engage with written texts to participate in society, to achieve their goals, and to develop their knowledge and potential (PIAAC Literacy Expert Group 2009, 8). The United Nations identifies literacy as a fundamental human right that is essential to social and human development (UNESCO 2016). It asserts that "for individuals, families, and societies alike, [literacy] is an instrument

grinvalds/iStockphoto

Photo 14.1 Strong literacy skills, including digital literacy skills, are needed to participate fully in Canadian society.

of empowerment to improve one's health, one's income, and one's relationship with the world" (UNESCO 2016). Relatively high levels of literacy, including digital literacy, are essential for people to be able to participate fully in Canadian society. As federal, provincial, and municipal governments adopt e-government strategies, access to many basic services requires navigating and deciphering web pages and completing online forms. In addition, information available over the Internet varies in quality, so people need to be able to evaluate what they read, especially when that information relates to social and political issues or provides health and illness information. The emergence of "fake news"—that is, news stories that are fabricated for the purpose of drawing traffic to advertising—which is shared widely through social media, highlights the importance of developing these evaluative skills. Although only about 4 per cent of people in Canada are completely illiterate (Harwood 2012), people with low literacy skills experience substantial disadvantages. Compared to those with high literacy skills, people with low literacy skills are more likely to be unemployed, to work fewer weeks per year, and to earn lower wages (Murray and Shilington 2011). People with low literacy skills are also more likely than people with higher literacy skills to live in low-income households, even after controlling for level of education, immigration status, and family size and type (Heisz, Notten, and Situ 2016).

The statistical analyses in this chapter rely on Canadian data from a survey developed to measure adults' literacy, numeracy, and technological skill levels in OECD (Organisation for Economic Co-operation and Development) countries, called the Programme for the International Assessment of Adult Competencies

(PIAAC). (See the "Spotlight on Data" box for more information.) Literacy is measured by asking people to complete a series of tasks that require them to access and identify information, integrate and interpret information, and evaluate and reflect on information (OECD 2012). The results are combined to produce a literacy scale score ranging from 0 to 500. (I refer to this as a "literacy score" throughout this chapter.) Literacy scores are grouped into five general proficiency levels, as follows:

- *Below Level 1 (score of 0–175)*: Tasks at this level require people to read short texts on familiar topics to locate a single piece of specific information.
- *Level 1 (score of 176–225)*: Most tasks at this level require people to read relatively short texts to locate a single piece of information that is identical to the information given in the question.
- *Level 2 (score of 226–275)*: Tasks at this level require people to make matches between the text and information and may require paraphrasing or low-level inferences.
- *Level 3 (score of 276–325)*: Texts at this level are dense or lengthy and can include several pages. Understanding text and rhetorical structures becomes more central to successfully completing tasks, especially navigating complex digital texts.
- *Level 4 (score of 326–375)*: Tasks at this level require people to perform multiple-step operations to integrate, interpret, or synthesize information from complex or lengthy texts.
- *Level 5 (score of 376–500)*: Tasks at this level require people to search for and integrate information across multiple, dense texts; construct syntheses of similar and contrasting ideas or points of view; or evaluate evidence-based arguments.

The majority of jobs in Canada require at least Level 3 literacy skills, yet 43 per cent of all students leaving Canada's high schools have not achieved this level (Harwood 2012). There are many barriers to achieving high levels of literacy, including low personal motivation, lack of family support, increased family responsibilities, insufficient educational support, and the inability to obtain jobs that reinforce literacy skills (Tilleczek and Campbell 2013). In part, literacy skills are an indicator of the socio-economic status, or social class, of the family that people grow up in (Willms and Watson 2008). High-status and highly educated parents are more likely to foster high-level literacy skills in their children through their parenting practices and family activities.

In the Canadian PIAAC sample, the average literacy score is 273 (s.d. = 50). Slightly more than half of respondents (51 per cent) have a literacy score at Level 3 or higher, whereas 4 per cent are below Level 1, 13 per cent are at Level 1, and 32 per cent are at Level 2. Compared to other OECD nations, Canada has a relatively high percentage of adults at Level 3 or higher, ranking twelfth overall; the five nations with the largest proportion of adults at Level 3 or higher are Japan, Finland, Netherlands, Sweden, and Australia (OECD 2106). Notably, Canada has a higher proportion of people with a post-secondary educational credential than all of these top-five nations, suggesting that there may be a disconnect between educational achievement and applied literacy skills. In this chapter, statistical analysis is used to discover the following:

- How is age related to literacy scores?
- How is having a post-secondary education related to literacy scores?
- How does the relationship between age and literacy scores change when having a post-secondary education is accounted for?
- Is the amount of time that people spend engaged in non-formal educational activities related to literacy scores?

Spotlight on Data

The Programme for the International Assessment of Adult Competencies

The Programme for the International Assessment of Adult Competencies (PIAAC) is an initiative of the Organisation for Economic Co-operation and Development (OECD). It is designed to help governments assess, monitor, and analyze the level and distribution of skills among their adult populations, and to provide data that are comparable across many countries (OECD n.d.). Twenty-four countries have participated in the PIAAC adult skills survey; the analyses in this chapter use only the Canadian data.

The PIAAC survey collects information about people's participation in formal and informal education, their occupation and work history, their technology use, their language profile, and their demographic characteristics. In addition, respondents complete a series of tests designed to assess literacy, numeracy, and problem-solving and reading skills. In Canada, the survey questionnaire and tests are available in both English and French.

The survey population is people aged 16 to 65. People residing in institutions, on Aboriginal reserves, on military bases, and in some sparsely populated areas are not included. Multi-stage, stratified sampling is used to select respondents (Statistics Canada 2017). First, geographic clusters, stratified by rural or urban status, are selected. Then, households are selected from within each cluster, and, finally, one individual is randomly selected from within each household. The data used in this chapter were collected between November 2011 and June 2012, and the overall response rate was 58 per cent. Weights that account for the probability of selection, non-response bias, and population characteristics are included in the dataset.

Because of the complex sampling and measurement in the PIAAC survey, researchers must use a secondary software tool to analyze the data. The International Database (IDB) Analyzer produces SPSS syntax (code) that accounts for the complex sampling and assessment structures of several large international surveys, including PIAAC.

Using Interaction Variables in Linear Regression

interaction effect Occurs when two or more independent variables are jointly related to a dependent variable, and their combined influence is different than the summed influence of each variable alone.

In Chapter 12, you learned how multiple linear regression allows researchers to identify the unique influence of each independent variable on a dependent variable. In our complex social world, however, sometimes two or more things interact with each other to jointly influence something. An **interaction effect** occurs when two or more independent variables are jointly related to a dependent variable, and their combined influence is different than the sum of the influence of each variable alone. In other words, an interaction effect occurs when the relationship between two variables changes when a third variable is taken into account. You were first introduced to interaction effects when you learned about the elaboration model. For instance, in Chapter 9, the elaboration model showed how the relationship between contact with police and perceptions of police fairness is influenced by racialization. Similarly, in Chapter 10, the elaboration model showed how the relationship between years since immigration and annual employment income is also influenced by racialization.

To illustrate how interaction effects are captured in regression, let's investigate the relationship between age and literacy scores and consider how that relationship is influenced by education. Table 14.1 shows the average literacy scores for people in different age cohorts and with different levels of education. On average, people in younger age cohorts have higher literacy scores than people in older age cohorts. The 25- to 34-year-old age cohort has the highest average literacy score (mean = 285; s.d. = 48), and the 55- to 65-year-old age cohort has the lowest average literacy score (mean = 260; s.d. = 51). In part, these results reflect the educational trajectories of people in different age cohorts. People aged 65 in 2012 were born in 1947 and, thus, likely attended high school during the 1960s, when access to post-secondary education was more limited. As a result, a smaller proportion of people in the 55 to 65 age cohort have a post-secondary educational credential: only 58 per cent compared to 72 per cent in the 25 to 34 age cohort. In addition, the literacy skills of people in older age cohorts may have eroded during the time after leaving formal schooling; this is particularly likely for people whose jobs do not require higher literacy skills.

As expected, people with higher levels of education have higher literacy scores, on average, than those with lower levels of education. The average literacy score of people who do not have a high school diploma is 234 (s.d. = 53) whereas the average literacy score of people with a university bachelor's degree is 298 (s.d. = 44); the average literacy score of people with a university graduate degree is 305 (s.d. = 43). These results underscore the importance of ensuring widespread access to post-secondary education so that Canada remains competitive in a globalized, knowledge economy. Notably, more variation in literacy scores exists among people with lower levels of education, a result that may be related to people's different levels of engagement in everyday literacy practices and participation in informal learning.

It's possible that the relationship between age and literacy scores exists only because people in different age cohorts have different levels of education. To investigate this possibility, I use the elaboration model to show the relationship between age and

Table 14.1 Literacy Scale Scores, by Age Group and Highest Educational Credential (n = 26,653)

	Literacy Score		
Characteristic	Percentage	Mean	Std. Dev.
Age cohort			
16 to 24	17.2	275.7	45.2
25 to 34	20.1	285.0	48.2
35 to 44	19.5	279.5	50.4
45 to 54	22.6	267.8	52.7
55 to 65	20.6	260.3	50.6
Highest educational credential			
Less than high school	15.2	233.5	52.5
High school diploma or equivalent	24.7	267.1	46.3
Apprenticeship, upgrading, or trade certificate	14.0	271.4	42.8
College/CÉGEP or university diploma/ certificate	20.6	278.9	44.2
University bachelor's degree	19.5	298.3	44.4
University master's/research degree	6.0	304.8	43.1
Total/Overall	*100.0*	*273.3*	*50.4*

Source: Author generated; Calculated using data from Statistics Canada, 2017.

literacy scores for two separate groups: people with a post-secondary educational credential and people without a post-secondary educational credential. Table 14.2 shows the average literacy scores for people in each age cohort, after accounting for whether or not they have a post-secondary educational credential. Among people who do not have a post-secondary education, the highest average literacy score is 271 (s.d. = 45) in the age 16 to 24 cohort, and the lowest average literacy score is 240 (s.d. = 53), which occurs in the age 45 to 54 cohort. Among people who do not have a post-secondary education, there is a 31-point difference between the average literacy scores of people in the youngest age cohort and people in the two oldest age cohorts. As expected, in every age cohort, the average literacy scores of people who have a post-secondary education are higher than the average literacy scores of people who do not have a post-secondary education. Among people with a post-secondary education, the highest average literacy score is 295 (s.d. = 44), which occurs in the age 25 to 34 cohort (relatively few people aged 16 to 24 have completed post-secondary education); and the lowest average literacy score is 275 (s.d. = 46), which occurs in the age 55 to 65 cohort. In other words, among people with a post-secondary education, there is only a 20-point difference between the average literacy scores of people in the second-youngest age cohort and people in the oldest age cohort.

In the sample overall, the difference between the age cohort with the highest average literacy score and the age cohort with the lowest average literacy score—or the zero-order relationship—is 25 points. (See Table 14.1.) Thus, this elaboration

Table 14.2 Using the Elaboration Model to Understand How the Relationship between Age and Literacy Scores Is Influenced by Level of Education (n = 26,653)

	No Post-Secondary Educational Credential			Has Post-Secondary Educational Credential		
	Literacy Score			Literacy Score		
	Percentage	Mean	Std. Dev.	Percentage	Mean	Std. Dev.
Age cohort						
16 to 24	32.7	270.8	45.3	7.0	291.0	41.6
25 to 34	14.2	259.6	49.6	24.0	294.9	43.7
35 to 44	12.6	249.9	54.6	24.2	289.7	44.4
45 to 54	18.9	240.4	53.2	25.0	281.6	46.6
55 to 65	21.6	240.5	50.1	19.8	274.7	45.9
Total/Overall	*100.0*	*254.3*	*51.4*	*100.0*	*286.0*	*45.5*

Source: Author generated; Calculated using data from Statistics Canada, 2017.

model provides another example of specification: the relationship between age and literacy scores is stronger among people without a post-secondary education (a 31-point difference in average scores) and weaker among people with a post-secondary education (a 20-point difference in average scores). People with a post-secondary education are more likely to work in jobs that require high-level literacy skills; thus, they are likely to maintain and develop their literacy skills over their employment careers.

Now let's use multiple linear regression to model the general pattern of the relationship between age, level of education, and literacy scores. Age and level of education are both treated as independent variables, and literacy score is treated as the dependent variable. A pseudo-continuous age variable was created by recoding the variable capturing people's ages in five-year intervals, using the method described in the "Hands-on Data Analysis" box in Chapter 13. The pseudo-continuous age variable is centred on 40 to make the constant coefficient easier to interpret. Level of education is measured using a dummy variable indicating whether people have a post-secondary educational credential (anything above high school, which has a "1" value) or not (only high school or less, which has a "0" value). Table 14.3 shows the regression results. The constant coefficient indicates that the predicted literacy score of people who are 40 years old and who do not have a post-secondary education is 270, or at Level 2. Each one-year increase in age is associated with having a literacy score that is 0.71 points lower; thus, every ten-year increase in age is associated with having a literacy score that is 7.1 points lower. People with a post-secondary education are predicted to have a literacy score that is 17.49 points higher than people without a post-secondary education. The adjusted R^2 statistic indicates that 13 per cent of the variation in literacy scores can be explained by age and level of education.

Table 14.3 Results of a Multiple Linear Regression Predicting Literacy Scores, Using Age and Level of Education

Dependent variable: Literacy scale score (n = 25,653)

	Unstandardized Coefficient	Standardized Coefficient
Constant	270.23*	–
Age (in years, centred on 40)	−0.71*	−0.20
Has a post-secondary educational credential	17.49*	0.34
Adjusted R^2	*0.13*	

* Indicates that results are statistically significant at the $p < 0.05$ level.
Source: Author generated; Calculated using data from Statistics Canada, 2017.

These regression coefficients can be substituted into a regression prediction equation to find the predicted literacy scores of people at various ages, with and without a post-secondary education. The prediction equation for the regression shown in Table 14.3 is:

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

$$\begin{aligned}&\textit{Predicted literacy score}\\&\qquad = 270.23 + (-0.71)(\textit{age centred}) + 17.49(\textit{has post-secondary education})\end{aligned}$$

Since the age variable is centred on 40, this must be accounted for in the prediction equation; for instance, a 16-year-old has the value –24 on the centred age variable, since age 16 is 24 years less than age 40. The post-secondary education variable is a dummy variable, with the value “0” or “1”, indicating the presence or absence of a post-secondary educational credential. So the predicted literacy score of a 16-year-old without a post-secondary education is:

$$\begin{aligned}\textit{Predicted literacy score} &= 270.23 + (-0.71)(-24) + 17.49(0)\\ &= 270.23 + 17.04\\ &= 287.27\end{aligned}$$

The predicted literacy score of a (very bright!) 16-year-old with a post-secondary education is:

$$\begin{aligned}\textit{Predicted literacy score} &= 270.23 + (-0.71)(-24) + 17.49(1)\\ &= 270.23 + 17.04 + 17.49\\ &= 304.76\end{aligned}$$

Similarly, the predicted literacy scores of a 65-year-old, without and with a post-secondary education, respectively, are:

$$\begin{aligned}\textit{Predicted literacy score} &= 270.23 + (-0.71)(25) + 17.49(0)\\ &= 270.23 + (-17.75)\\ &= 252.48\end{aligned}$$

$$\begin{aligned} \textit{Predicted literacy score} &= 270.23 + (-0.71)(25) + 17.49(1) \\ &= 270.23 + (-17.75) + 17.49 \\ &= 269.97 \end{aligned}$$

As you learned in Chapter 12, these results can be used to plot the predicted literacy scores of people at various ages, with different levels of education. (See Figure 14.1.) The horizontal dotted line in Figure 14.1 is the cut-off point between Level 2 and Level 3 literacy skills; recall that most jobs in Canada require at least Level 3 literacy skills. There are two regression lines: one showing the predicted literacy scores among people with a post-secondary education and one showing the predicted literacy scores among people without a post-secondary education. The two lines are parallel, and the distance between them is determined by the regression coefficient of the "Has a post-secondary educational credential" dummy variable.

Although these predictions are a good start, the elaboration model results suggest that having a post-secondary education (or not) influences the magnitude of the relationship between age and literacy scores. Age seems to have a stronger relationship with literacy scores for people without a post-secondary education than for people with a post-secondary education. The implication of the elaboration model result is that the two regression lines depicted in Figure 14.1 should have different slopes: the angle of the line should be steeper for people without a post-secondary education, and flatter for people with a post-secondary education.

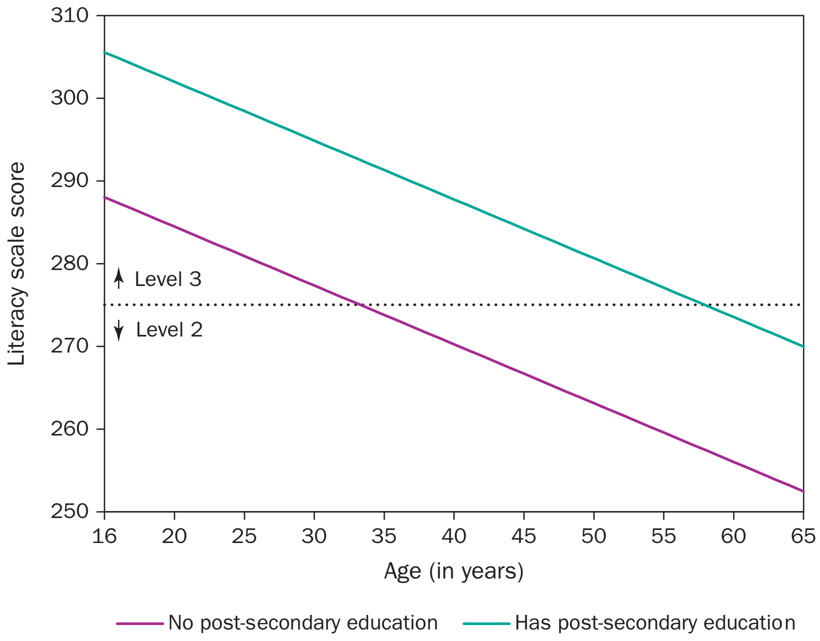

Figure 14.1 **Literacy Scores Predicted by the Regression in Table 14.3**

Source: Author generated; Calculated using data from Statistics Canada, 2017.

Interaction variables (also called interaction terms) allow researchers to use linear regression to predict relationships of different magnitudes for different groups of people. In other words, interaction variables make it possible to predict regression lines with different slopes for different subgroups, all in a single model. Let's add a variable to the regression in Table 14.3 that captures the interaction between age and level of education. Interaction variables are created by multiplying the values on two (or more) independent variables together for each case. Table 14.4 shows the values on an "Age" variable, a "Has a post-secondary educational credential" dummy variable, and a variable capturing their interaction for 10 hypothetical people. For the first five people, the "Has a post-secondary educational credential" dummy variable has a "0" value and, thus, the interaction variable has a "0" value, since 0 multiplied by any age equals 0. For the final five people, the "Has a post-secondary educational credential" dummy variable has a "1" value and, thus, the interaction variable has a value equal to people's age, since people's age multiplied by 1 is equal to their age. Interaction variables are usually created using statistical software.

interaction variable Created by multiplying the values on two or more variables together for every case.

Once an interaction variable has been created, it can be used as an independent variable in a regression, just like any other variable. Whenever an interaction variable is used in a regression, though, the variables used to create the interaction variable must also be used as independent variables in the regression; otherwise the results are very difficult to interpret. Table 14.5 shows the results of a multiple linear regression predicting literacy scores, using three independent variables: a (centred) "Age" variable, a "Has a post-secondary educational credential" dummy variable, and an interaction variable, which was created by multiplying the value on the (centred) "Age" variable by the value on the "Has a post-secondary educational credential" dummy variable for each case. As always, the constant coefficient

Table 14.4 Values on an Interaction Variable for 10 Hypothetical Cases

Person	Age (in Years)	Has a Post-Secondary Educational Credential (Dummy Variable)	Age x Post-Secondary Educational Credential (Interaction Variable)
Tamar	20	0	0
Liz	30	0	0
Danielle	40	0	0
Geza	50	0	0
David	60	0	0
Tai Lee	20	1	20
Sumi	30	1	30
Lucas	40	1	40
Mithi	50	1	50
William	60	1	60

Table 14.5 Results of a Multiple Linear Regression with Three Independent Variables (One Interaction Variable)

Dependent variable: Literacy scale score (n = 26,653)

	Unstandardized Coefficient	Standardized Coefficient
Constant	270.01*	–
Age (in years, centred on 40)	−0.80*	−0.22
Has a post-secondary educational credential	17.47*	0.34
Age (centred) x post-secondary educational credential	0.19*	0.04
Adjusted R^2	*0.13*	

* Indicates that results are statistically significant at the p < 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2017.

shows the predicted value on the dependent variable for people with a "0" value on all of the independent variables. Thus, 270 is the predicted literacy score for people aged 40 who do not have a post-secondary education. These people also have a "0" value on the interaction variable because a "0" value on the "Age" variable multiplied by a "0" value on the "Has a post-secondary educational credential" variable is equal to 0.

The interpretation of the slope coefficients is slightly different when a linear regression uses an interaction variable as an independent variable. The slope coefficients still show the change in the dependent variable that is associated with a one-unit increase in the independent variable. But, because of the interaction variable, the slope coefficient of the "Age" variable shows the predicted relationship between age and literacy scores, *when the "Has a post-secondary educational credential" variable equals 0.* In other words, it shows the slope of the regression line for people who do not have a post-secondary education. The slope coefficient of the "Has a post-secondary educational credential" variable shows the predicted relationship between having a post-secondary education and literacy scores, *when the "Age" variable equals 0.* In other words, it shows the distance between the two regression lines (for people with/without a post-secondary education) for people who are aged 40, since a "0" value on the age variable corresponds to age 40. Finally, the slope coefficient of the interaction variable shows how the predicted relationship between age and literacy scores *is different for people with a post-secondary education than for those without.* In practice, the slope coefficient of the interaction variable is added to the slope coefficient of the "Age" variable for people who have a post-secondary education when predicting literacy scores.

It is easiest to understand how interaction variables work by creating regression prediction equations for people of different ages and with different levels of education. The results can then be plotted to show how the slopes of the regression lines differ because of the coefficient of the interaction variable. The prediction equation for the regression shown in Table 14.5 is similar to that for the regression in Table 14.3, although the coefficients are slightly different, and the interaction variable is added to the end:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3(x_1x_2)$$

$$\begin{aligned} &\textit{Predicted literacy score} \\ &\quad = 270.01 + (-0.80)(\textit{age centred}) + 17.47(\textit{has post-secondary education}) \\ &\quad + 0.19(\textit{age centred} \times \textit{has post-secondary education}) \end{aligned}$$

For people without a post-secondary education, the value on the post-secondary education variable is "0", and the value on the interaction variable is also "0" (since any age multiplied by 0 equals 0). So for a 16-year-old without a post-secondary education:

$$\begin{aligned} \textit{Predicted literacy score} &= 270.01 + (-0.80)(-24) + 17.47(0) + 0.19(0) \\ &= 270.01 + 19.20 \\ &= 289.21 \end{aligned}$$

For a 40-year-old without a post-secondary education:

$$\begin{aligned} \textit{Predicted literacy score} &= 270.01 + (-0.80)(0) + 17.47(0) + 0.19(0) \\ &= 270.01 \end{aligned}$$

And for a 65-year-old without a post-secondary education:

$$\begin{aligned} \textit{Predicted literacy score} &= 270.01 + (-0.80)(25) + 17.47(0) + 0.19(0) \\ &= 270.01 + (-20.00) \\ &= 250.01 \end{aligned}$$

For people without a post-secondary education, the predicted literacy scores and, thus, the slope of the regression line are determined entirely by the coefficient of the "Age" variable.

Now let's find the predicted literacy scores for people at the same three ages (16, 40, and 65) who have a post-secondary education. This time, the value on the post-secondary education variable is "1", and the value on the interaction variable is the same as the value on the age variable (since any age multiplied by 1 is equal to that age). So for a 16-year-old with a post-secondary education:

$$\begin{aligned} \textit{Predicted literacy score} &= 270.01 + (-0.80)(-24) + 17.47(1) + 0.19(-24) \\ &= 270.01 + 19.20 + 17.47 + (-4.56) \\ &= 302.12 \end{aligned}$$

For a 40-year-old with a post-secondary education:

$$\begin{aligned} \textit{Predicted literacy score} &= 270.01 + (-0.80)(0) + 17.47(1) + 0.19(0) \\ &= 270.01 + 17.47 \\ &= 287.48 \end{aligned}$$

And for a 65-year-old with a post-secondary education:

$$\begin{aligned} \textit{Predicted literacy score} &= 270.01 + (-0.80)(25) + 17.47(1) + 0.19(25) \\ &= 270.01 + (-20.00) + 17.47 + 4.75 \\ &= 272.23 \end{aligned}$$

For people who are 40 years old (who have a "0" value on the age variable because of centring), the predicted literacy score for those with a post-secondary education is exactly 17.47 points higher than the predicted literacy score for those without a post-secondary education. That is, the slope coefficient of the post-secondary education variable shows the predicted relationship between having a post-secondary education and literacy scores, when the age variable equals 0.

For people at other ages, the 17.47-point difference associated with having a post-secondary education is still incorporated into the prediction, but the predicted literacy scores are also affected by the size of the slope coefficient of the interaction variable. So, the difference between the predicted literacy scores of 16-year-olds without a post-secondary education and 16-year-olds with a post-secondary education is determined by two elements: (1) the contribution of the slope coefficient of the post-secondary education variable, which adds 17.47 points for people at every age, *and* (2) the contribution of the slope coefficient of the interaction variable, which changes depending on people's age and subtracts 4.56 points for 16-year-olds (since 0.19 x –24 = –4.56). Similarly, the difference between the predicted literacy scores of 65-year-olds without a post-secondary education and 65-year-olds with a post-secondary education is determined by the coefficient of the post-secondary education variable (which adds 17.47 points for people at every age) *and* the contribution of the coefficient of the interaction variable, which adds 4.75 points for 65-year-olds (0.19 x 25 = 4.75).

Based on these calculations, among people without a post-secondary education, 16-year-olds are predicted to have literacy scores that are 39 points higher than 65-year-olds. In contrast, among people with a post-secondary education, 16-year-olds are predicted to have literacy scores that are only 30 points higher than 65-year-olds. In other words, the decrease in literacy scores associated with each older age cohort is smaller for people with a post-secondary education than for people without a post-secondary education.

Figure 14.2 shows the predicted relationship between age and literacy scores for people with and without a post-secondary education. Because of the interaction variable, the slopes of the two lines are different; that is, they are no longer parallel to one another (as in Figure 14.1). The different slopes show that the predicted influence of having a post-secondary education on literacy scores is smaller for people in younger age cohorts than for people in older cohorts. In other words, age cohort and level of education appear to jointly influence literacy scores.

This example illustrates how to interpret interactions between a ratio-level variable (age) and a categorical variable with only two attributes (has/does not have a post-secondary education). Interaction variables can also be used to predict more

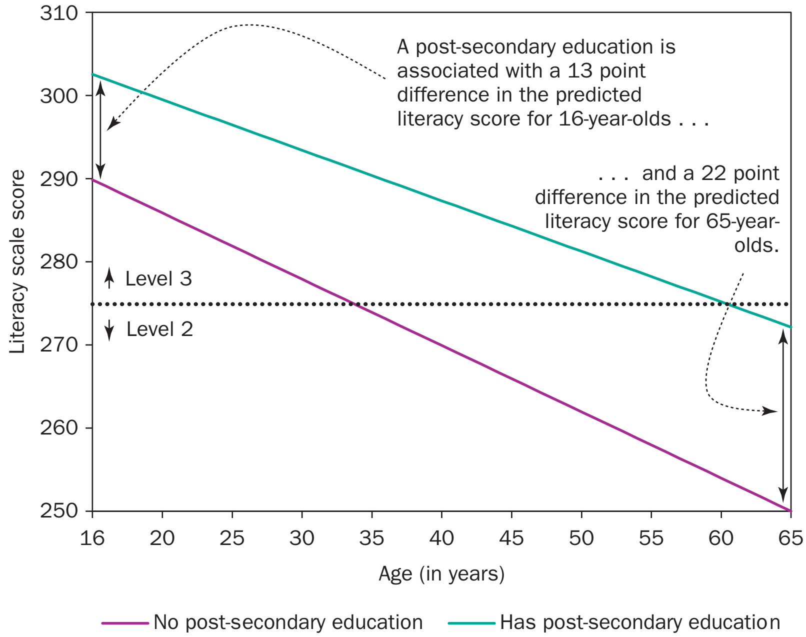

Figure 14.2 **Literacy Scores Predicted by the Regression in Table 14.5**

Source: Author generated; Calculated using data from Statistics Canada, 2017.

complex relationships. For instance, instead of dividing level of education into only two groups, I can retain five levels of education (collapsing the two university-level categories). To do so, I create five dummy variables to capture the education variable and use four of them as independent variables in a regression. (The omitted dummy variable becomes the reference group.) Four interaction variables are then needed to capture the interaction between age and education: one to correspond with each of the education dummy variables in the regression. Each interaction variable is created by multiplying the value on the age variable by the value on the corresponding education dummy variable.

Incorporating more detailed information about people's level of education results in a regression with nine independent variables: the age variable, the four education dummy variables, and the four corresponding interaction variables. Table 14.6 shows the results of this regression. But when regressions use more than one interaction variable, the predictions are much easier to interpret when they are graphed. In Figure 14.3, the slopes of the regression lines suggest that dividing people's highest educational credentials into three groups might be best when investigating the relationship between age cohort and literacy scores. People with less than a high school education are predicted to have a relatively steep decrease in literacy scores in older age cohorts. People with a university degree are predicted to have much less decrease in literacy scores in older age cohorts. And, finally, people with a high school diploma; people with an apprenticeship, upgrading, or trade certificate; and people with a college/CÉGEP or university diploma or certificate are predicted to have a similar decrease in literacy scores in

Table 14.6 **Results of a Multiple Linear Regression with Nine Independent Variables (Four Interaction Variables)**

Dependent variable: Literacy scale score (n = 26,653)

	Unstandardized Coefficient	Standardized Coefficient
Constant	276.94*	–
Age (in years, centred on 40)	−0.59*	−0.16
Highest educational credential (ref: high school diploma or equivalent)		
Less than high school	−17.38*	−0.25
Apprenticeship, upgrading, or trade certificate	3.48*	0.05
College/CÉGEP or university diploma/certificate	7.55*	0.12
University bachelor's degree or higher	17.45*	0.30
Age x education interaction variables		
Age (centred) x Less than high school	−0.49*	−0.07
Age (centred) x Apprenticeship, upgrading, or trade certificate	0.00	0.00
Age (centred) x College/CÉGEP or university diploma/certificate	−0.11	−0.01
Age (centred) x University bachelor's degree or higher	0.12	0.01
Adjusted R^2	*0.21*	

* Indicates that results are statistically significant at the $p < 0.05$ level.
Source: Author generated; Calculated using data from Statistics Canada, 2017.

older age cohorts. In other words, the slopes for these three groups are similar. Overall, these results suggest that the influence of age cohort on literacy scores is moderated by education, with three unique interactions or joint relationships: one for people with less than a high school education, one for people with a university degree, and one for everyone with a level of education between these two. This type of analysis helps to illustrate the complex interactions between age cohort, educational credentials, and literacy scores.

Interaction variables can also be used to capture the joint influence of two categorical variables. To do this, a researcher first creates dummy variables for each categorical variable and then uses those dummy variables to create a series of interaction variables that capture all of the possible combinations of attributes. Researchers also sometimes investigate the joint influence of two ratio-level variables. To do this, they simply multiply the values on the two ratio-level variables to create the interaction variable. Unfortunately, the slope coefficient of an interaction variable created using two ratio-level variables is harder to interpret and to display visually.

Overall, interaction variables help researchers to model more complex relationships between variables, by showing how two independent variables are jointly related to a dependent variable. Many social science researchers highlight the importance of intersectionality or intersectional identities for understanding people's experiences. Interaction variables allow quantitative social scientists to incorporate an understanding of intersectionality in regression models. But because the slope coefficients of interaction variables can be challenging to interpret and explain, most researchers use them sparingly in regression. Researchers typically present

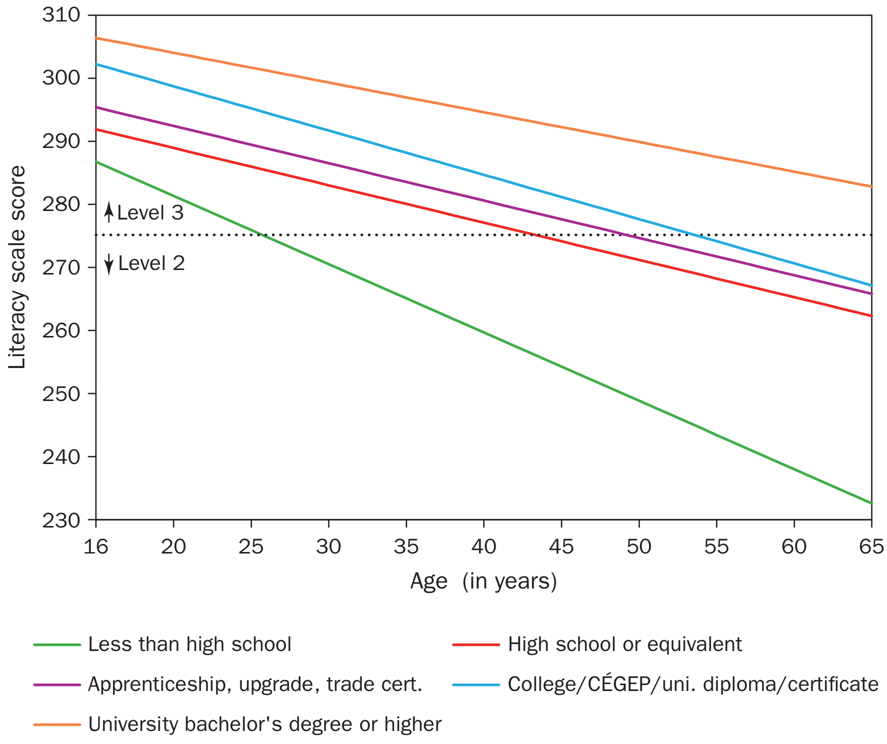

Figure 14.3 **Literacy Scores Predicted by the Regression in Table 14.6**

Source: Author generated; Calculated using data from Statistics Canada, 2017.

regression results related to interaction variables using graphs because the slope coefficients alone can be difficult to meaningfully describe.

Statistical Significance Tests When Interaction Variables Are Used

You just learned how the interpretation of slope coefficients changes when regressions use interaction variables as independent variables. Because an interaction variable changes what the slope coefficients show, it also changes what tests of statistical significance indicate.

Recall that in the regression shown in Table 14.5, the slope coefficient of the "Age" variable shows the predicted relationship between age and literacy scores but only when the post-secondary education variable equals 0. In other words, the slope coefficient of the "Age" variable shows the predicted relationship between age and literacy scores for people without a post-secondary education. Similarly, the test of statistical significance shows the likelihood of randomly selecting a sample with this observed relationship, or one of greater magnitude, if no relationship exists between age and literacy scores in the population of people without a post-secondary education. In other words, the test of statistical significance no longer shows the likelihood of selecting this sample if no relationship exists between age and literacy scores in the population overall. Instead, it shows the likelihood of selecting this sample if no relationship exists between these two variables in the population of people without a post-secondary education.

Recall also that the slope coefficient of the "Has a post-secondary educational credential" variable shows the predicted relationship between having a post-secondary education and literacy scores, when the age variable equals 0 (which represents people aged 40). As a result, the test of statistical significance shows the likelihood of randomly selecting a sample with the observed relationship, or one of greater magnitude, if no relationship exists between having a post-secondary education and literacy scores in the population of people aged 40.

When an interaction variable is used as an independent variable, the regression slope coefficients and the associated significance tests for the independent variables used to create that interaction variable no longer refer to the overall relationship between each independent variable and the dependent variable. Instead, the slope coefficients and the associated significance tests refer to the partial relationship between the independent variable and the dependent variable, for a group or condition that is defined by another variable. Statisticians sometimes refer to this as a conditional relationship.

The interpretation of the statistical significance test associated with the slope coefficient of the interaction variable is closer to the typical interpretation of a significance test for a regression slope coefficient. The test of statistical significance associated with the slope coefficient of an interaction variable shows the probability of randomly selecting a sample with the observed relationship, or one of greater magnitude, if there is no joint relationship between the variables used to create the interaction variable and the dependent variable in the population. In other words, when an interaction variable is statistically significant, researchers are relatively confident that, in the population, the relationship between the dependent variable and one of the variables in the interaction variable differs depending on the value on the other variable(s) used to create the interaction variable.

Statistics in Use

How Does "Readiness to Learn" Affect Literacy Skills?

Original research: Smith, M. Cecil, Amy D. Rose, Jovita Ross-Gordon, and Thomas J. Smith. 2015. "Adults' Readiness to Learn as a Predictor of Literacy Skills." American Institutes for Research-PIAAC.

In this "Statistics in Use" box, I describe the results of a study that uses interaction variables as independent variables in a multiple linear regression. The researchers were interested in finding out whether adults' "readiness to learn" affected their literacy skills and the use of those skills. The data were taken from the 2013 PIAAC Survey of Adult Skills collected in the United States. A total of 5,010 adults between the ages of 15 and 65 completed the survey.

"Readiness to learn" (RtL) was defined as adults' "propensity to learn new things, relate these knowledge and skills to existing knowledge and life situations, and engage in problem solving and information-seeking behaviour" (p. 3). Each person was assigned a readiness-to-learn score, based on their answers to a series of questions about learning new things and linking information together. The researchers hypothesized that readiness to learn influences the relationship between people's demographic characteristics (gender, age, work experience, and education) and their literacy skills. To test this hypothesis, they developed a multiple linear regression model, with literacy scores as the dependent variable. Some of the regression results are shown in Table 14.7.

Table 14.7 Moderating Effects of Readiness to Learn on the Relationship between Selected Predictors and Literacy Scores

Outcome	Effect	*df*	*β*	*SE*	*t*
Literacy					
(R^2 = .35)	Readiness to Learn	1	0.03	0.02	1.52
	Gender (female)	1	0.01	0.02	0.54
	Age (5-year increments)	1	–0.35	0.03	–10.56***
	Work experience (in years)	1	0.21	0.04	5.73***
	Education (in years)	1	0.55	0.01	37.43***
	RtL x Gender	1	0.03	0.02	1.60
	RtL x Age	1	–0.07	0.03	–2.50*
	RtL x Work experience	1	0.03	0.03	1.00
	RtL x Education	1	–0.04	0.02	–2.31*

Note: *p < 0.05, **p <0.01, ***p < 0.001, df_{error} = 1,864

Source: Excerpt from Smith et al. 2015, 26.

Based on these regression results, the researchers conclude that people's readiness to learn (RtL) moderates the relationship between age and literacy and the relationship between education and literacy. They note that "specifically, as RtL increased, the effects of age and educational level on [literacy] skills outcomes decreased. Equivalently, at low educational levels (or younger ages), the effect of RtL on skill levels was more pronounced than at high educational levels (or older ages)." To better illustrate the results, they present the graph in Figure 14.4, which shows that for people with low education, readiness to learn is related to literacy skills, but for people with high education, readiness to learn has a weaker relationship with literacy skills.

In a regression that does not account for the interactions between readiness to learn and demographic characteristics, readiness to learn is a statistically significant predictor of literacy levels, although the relationship between readiness to learn and literacy is weaker than the relationship between years of education and literacy, and between age and literacy. But when the interactions between readiness to learn and demographic characteristics are considered,

Continued

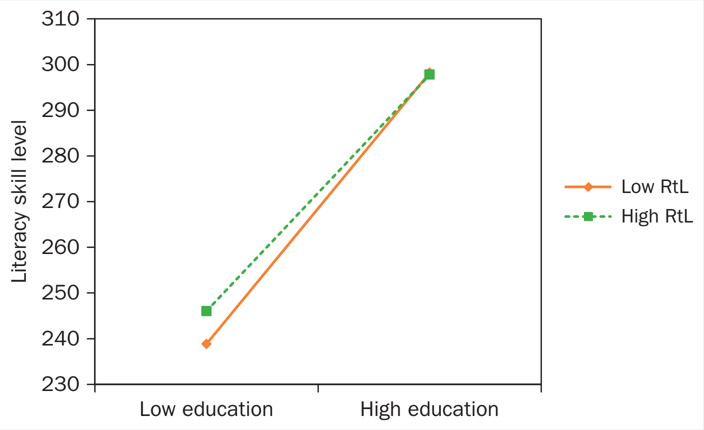

Figure 14.4 **The Moderating Effect of Readiness to Learn (RtL) on the Relationship between Educational Level and Literacy Skill Level**

Source: Smith et al. 2015, 31.

the researchers show that "although readiness to learn mediates some of the effects of education on skill level, this mediation is partial—that is, education still exerts considerable direct effect on skills. Therefore, while readiness to learn is a part of the overall picture of adults' literacy skills, schooling is much more important" (p. 11). Nonetheless, the researchers assert that increasing people's readiness to learn can potentially reduce some of the differences in literacy skills between people with lower education and people with higher education. As a result, they recommend implementing adult education policies and practices designed to enhance the readiness to learn of low-education workers in order to enhance their career prospects (p. 13).

Using Linear Regression to Predict Curvilinear Relationships

curvilinear relationship Occurs when the line of best fit between two variables is curved, not straight.

quadratic variable Created by squaring the value on a ratio-level variable for every case.

So far, you have learned how to use linear regression to predict straight-line relationships between one or more independent variables and a dependent variable. But linear regression can also be used to predict a **curvilinear** (or curved) **relationship** between an independent variable and a dependent variable. This is done by incorporating a **quadratic variable**, or a quadratic term (which is the same thing as a squared variable or a squared term), into a linear regression. Whenever you see a linear regression that uses a quadratic or squared independent variable, it indicates that the researcher thinks that the relationship between that variable and the dependent variable is curvilinear instead of linear. For instance, age is often related to things in a curvilinear way: there are many characteristics that generally improve as we grow older, and then reach a point where they begin to decline. This applies to many biological characteristics (such as motor skills), psychological characteristics (such as memory), and social characteristics (such as income). It is only sensible to predict curvilinear relationships between ratio-level independent variables and the ratio-level dependent variable in a regression; because categorical independent variables are incorporated into regressions using dummy variables, it doesn't make sense to square them.

To predict a curvilinear relationship between a ratio-level independent variable and the dependent variable in a linear regression, two versions of the independent variable are used in the regression: the original version of the variable (called the linear variable) and a quadratic version of the variable. The slope coefficient of the linear version of the independent variable indicates the angle of the predicted straight-line relationship between the independent variable and the dependent variable. The slope coefficient of the quadratic version of the variable indicates the direction and shape of the predicted curvilinear relationship between the independent variable and the dependent variable.

Let's start by using a hypothetical example to illustrate how quadratic variables predict curvilinear relationships. To create a quadratic variable, researchers simply square the value on the original variable for each case (multiply the value by itself). The values on the quadratic variable grow exponentially as a result of the squaring. Table 14.8 shows the values on a hypothetical original variable and the quadratic variable that corresponds to it.

A linear regression that uses both the linear and the quadratic version of this independent variable produces a constant coefficient and two slope coefficients: one for the linear version of the variable and one for the quadratic version of the

Table 14.8 The Values on Linear and Quadratic Versions of a Single Independent Variable (Hypothetical Data)

Person	Original Linear Variable (x)	Quadratic Variable (x^2)
Mandeep	0	0
Mikaela	10	100
Ishaan	20	400
Noam	30	900
Althea	40	1,600
Christos	50	2,500
Luka	60	3,600
Katerina	70	4,900
Fatima	80	6,400
Joaquin	90	8,100
Zeynep	100	10,000

variable. If they are the only independent variables used in the regression, the prediction equation is:

$$\hat{y} = a + b_1 x + b_2 \left(x^2\right)$$

$$\begin{aligned}&\textit{Predicted value on the DV}\\ &\quad = a + \left(b_1\right)\left(\textit{linear version of IV}\right) + \left(b_2\right)\left(\textit{quadratic version of IV}\right)\end{aligned}$$

Notice that the linear regression prediction equation has not changed—the only difference is that the second independent variable is just the square of the first independent variable, instead of being a completely different independent variable. To illustrate how a curvilinear relationship is predicted, let's imagine that the regression constant coefficient is 0, the slope coefficient of the original (linear) independent variable is 5, and the slope coefficient of the corresponding quadratic variable is 10. In this hypothetical situation, the regression prediction equation becomes:

$$\hat{y} = 0 + (5)(x) + (10)\left(x^2\right)$$

$$\begin{aligned}&\textit{Predicted value on the DV}\\ &\quad = 0 + (5)\left(\textit{linear version of IV}\right) + (10)\left(\textit{quadratic version of IV}\right)\end{aligned}$$

Table 14.9 illustrates what happens when the values on the two versions of the independent variable—the original linear version and the quadratic version—shown in Table 14.8 are substituted into this regression prediction equation.

Because of the exponential growth in the values on the quadratic variable, the predicted values on the dependent variable no longer correspond to a straight-line

Table 14.9 Using Linear and Quadratic Versions of an Independent Variable in a Hypothetical Regression Prediction Equation (Hypothetical Data)

Original Linear Variable (x)	Quadratic Variable (x^2)	Prediction Equation $0 + (5)(x) + (10)(x^2) =$	Predicted Value on the Dependent Variable ($\hat{y}$)
0	0	0 + (5)(0) + (10)(0) =	0
10	100	0 + (5)(10) + (10)(100) =	1,050
20	400	0 + (5)(20) + (10)(400) =	4,100
30	900	0 + (5)(30) + (10)(900) =	9,150
40	1,600	0 + (5)(40) + (10)(1,600) =	16,200
50	2,500	0 + (5)(50) + (10)(2,500) =	25,250
60	3,600	0 + (5)(60) + (10)(3,600) =	36,300
70	4,900	0 + (5)(70) + (10)(4,900) =	49,350
80	6,400	0 + (5)(80) + (10)(6,400) =	64,400
90	8,100	0 + (5)(90) + (10)(8,100) =	81,450
100	10,000	0 + (5)(100) + (10)(10,000) =	100,500

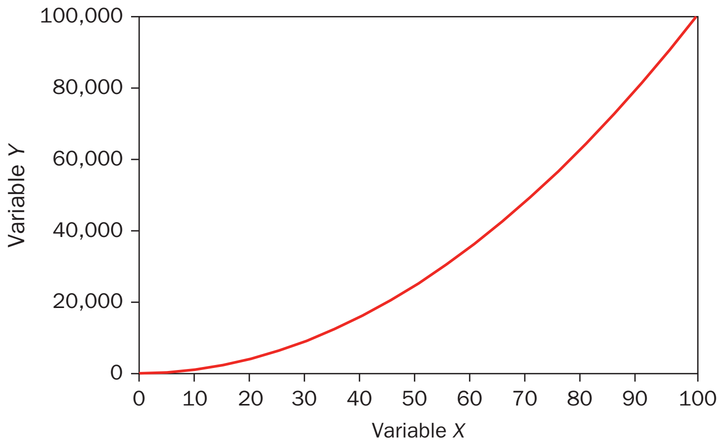

Figure 14.5 **The Predicted Relationship between the Independent Variable and the Dependent Variable in Table 14.9 (Hypothetical Data)**

relationship. When the predicted relationship between the original independent variable and the dependent variable is graphed, the curved shape becomes apparent. (See Figure 14.5.) This is how quadratic variables enable researchers to use linear regression to predict curvilinear relationships.

As with interaction variables, interpreting the slope coefficients of quadratic variables is difficult without using graphs. But it's possible to get a general idea about the shape of a relationship by looking at the sign of the slope coefficient of a quadratic variable, along with the sign of the slope coefficient of the corresponding linear variable. The top row of Figure 14.6 shows how to interpret the slope coefficient of a quadratic variable when the slope coefficient of the corresponding linear variable is positive. The bottom row of Figure 14.6 shows how to interpret the slope coefficient of a quadratic variable when the slope coefficient of the corresponding linear variable is negative. When the slope coefficient of a quadratic variable is positive, the predicted relationship has an upward curve, with the tails of the line moving up and away from the straight line. When the slope coefficient of a quadratic variable is negative, the predicted relationship has a downward curve, with the tails dropping further below the straight line. The size of the slope coefficient of a quadratic variable indicates how quickly or slowly the regression line curves away from the straight line. Although in Figure 14.6 each curved line overlaps the straight line at 0, this will not always occur. The best way to assess the shape and magnitude of a curvilinear relationship is to use the regression coefficients to calculate the predicted value on the dependent variable for several plausible values on the independent variable and to graph the relationship, as in Figure 14.5.

Let's return to the Canadian PIAAC data to investigate whether the relationship between age cohort and literacy scores is linear or curvilinear. Table 14.10 shows the results of a linear regression that predicts literacy scores using people's age. To assess the possibility of a curvilinear relationship, two versions of the "Age" variable are used as independent variables: a linear version and a quadratic version.

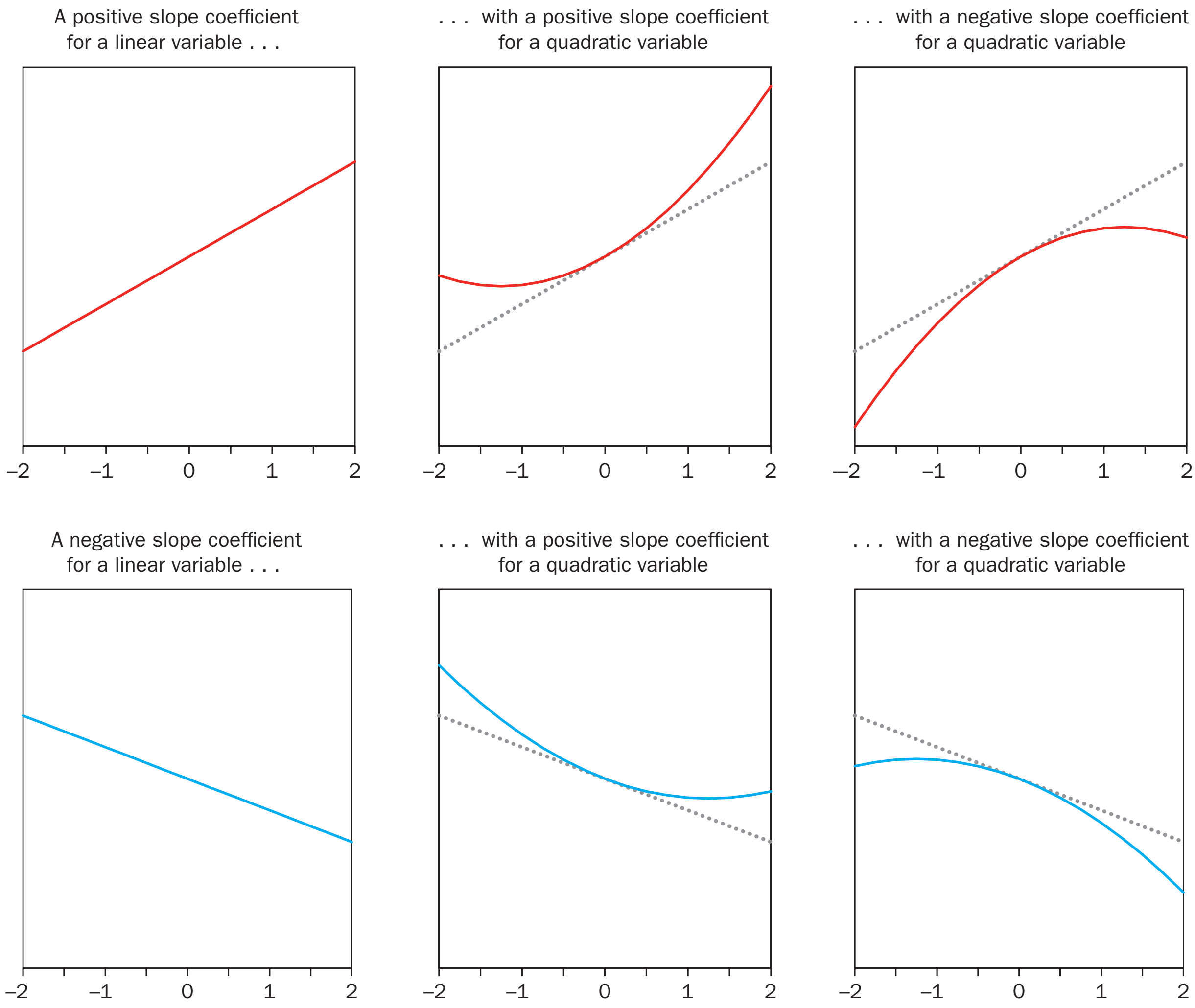

Figure 14.6 **A Guide to Interpreting Slope Coefficients When a Linear Regression Uses a Quadratic Variable (Hypothetical Data)**

As in the other examples in this chapter, the linear version of the "Age" variable was centred on 40 before the variable was squared. Centring variables is especially useful when researchers are modelling curvilinear relationships because doing so helps to avoid collinearity problems.

The constant coefficient of the regression shown in Table 14.10 indicates that the predicted literacy score of people who are aged 40 is 279. Both the linear version of the "Age" variable and the quadratic version of the "Age" variable have negative slope coefficients, so the ends of the curve will be pointing downwards, as in the bottom right panel of Figure 14.6. In other words, literacy scores are predicted to increase with each older age cohort, up to a certain point, and then they are predicted to decrease.

Figure 14.7 shows the predicted relationship between age cohort and literacy scores from two different linear regressions. The straight blue line shows

Table 14.10 Results of a Multiple Linear Regression with Two Independent Variables (One Linear and One Quadratic)

Dependent variable: Literacy scale score (n = 26,653)

	Unstandardized Coefficient	Standardized Coefficient
Constant	279.23*	–
Age (in years, centred on 40)	−0.51*	−0.14
Age (in years, centred on 40), squared	−0.03*	−0.10
Adjusted R^2	*0.03*	

* Indicates that results are statistically significant at the $p < 0.05$ level.
Source: Author generated; Calculated using data from Statistics Canada, 2017.

the predicted relationship between age and literacy scores when only the linear "Age" variable is used as an independent variable in the regression. (These regression coefficients are not shown.) The linear relationship predicted by the blue line suggests that the youngest age cohort has the highest literacy scores and that literacy scores decrease steadily in each older age cohort. This prediction doesn't quite match with how learning is organized in our society: it's unlikely that teens have higher literacy scores than people who are in their early twenties, some of whom are engaged in post-secondary education. The curved red line shows the predicted relationship between age and literacy scores when both a linear "Age" variable and a quadratic "Age" variable are used as independent variables in the regression (the regression shown in Table 14.10). This predicted relationship clearly fits better with our understanding of the relationship between age and

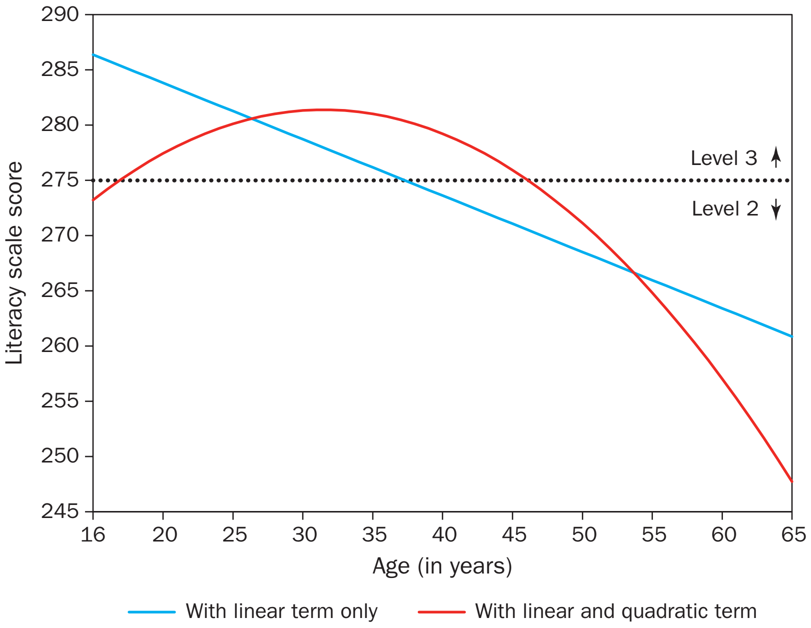

Figure 14.7 Predicted Linear and Curvilinear Relationships between Age and Literacy Scores

Source: Author generated; Calculated using data from Statistics Canada, 2017.

learning, including literacy. The curved red line suggests that there is a positive relationship between literacy scores and age, up until the age cohort who are in their early thirties. The apex of the curve in the early thirties corresponds to the years when many people are most attentive to formal education and specialized job training, which might be the reason for the high literacy scores in these age cohorts. For the age cohorts from the mid-thirties onward, literacy scores are predicted to be lower in each older cohort. As described earlier, this may reflect the different educational expectations and experiences of people in older age cohorts. It may also indicate that literacy skills have deteriorated in older age cohorts, as a result of the length of time away from formal schooling or other educational activities.

Quadratic variables allow researchers to use regression to model relationships that more accurately reflect real-world processes because they are no longer restricted to predicting straight-line relationships, even in the context of linear regression. Curvilinear relationships capture situations where the relationship between two variables is positive at some values of the independent variable and negative at other values of the independent variable.

Using quadratic variables to predict curvilinear relationships can be combined with other regression modelling techniques. For instance, given the analyses earlier in this chapter, it's possible that the curvilinear relationship between age and literacy scores is also related to education. To investigate this possibility, I can use both quadratic and interaction variables as independent variables in a regression. For example, the regression shown in Table 14.10 can be refined by adding a variable (or variables) for people's highest educational credential and variables for the interaction between people's highest educational credential and their age. For simplicity, educational credentials are once again divided into only two groups: people with a post-secondary education and those without, captured in a dummy variable. Then, two interaction variables are created: one that multiplies the linear "Age" variable by the "Has a post-secondary educational credential" dummy variable, and one that multiplies the quadratic "Age" variable by the "Has a post-secondary educational credential" dummy variable. Thus, the multiple linear regression predicting literacy scores has five independent variables: the dummy variable for "Has a post-secondary educational credential," two "Age" variables (one linear and one quadratic), and two interaction variables (the "Has a post-secondary educational credential" dummy variable multiplied by the linear "Age" variable and the "Has a post-secondary educational credential" multiplied by the quadratic "Age" variable). Since the regression coefficients are difficult to interpret on their own, the results are graphed in Figure 14.8.

Figure 14.8 shows the predicted curvilinear relationship between age and literacy scores for people who have a post-secondary education and people who do not. For people who do not have a post-secondary education, each older age cohort (beyond 16) is predicted to have a lower literacy score. The relationship is monotonic—that is, it always moves in the same direction, even though it is not linear. In contrast, for people with a post-secondary education, literacy scores are predicted to be positively related to age for the cohorts between the mid-teens and

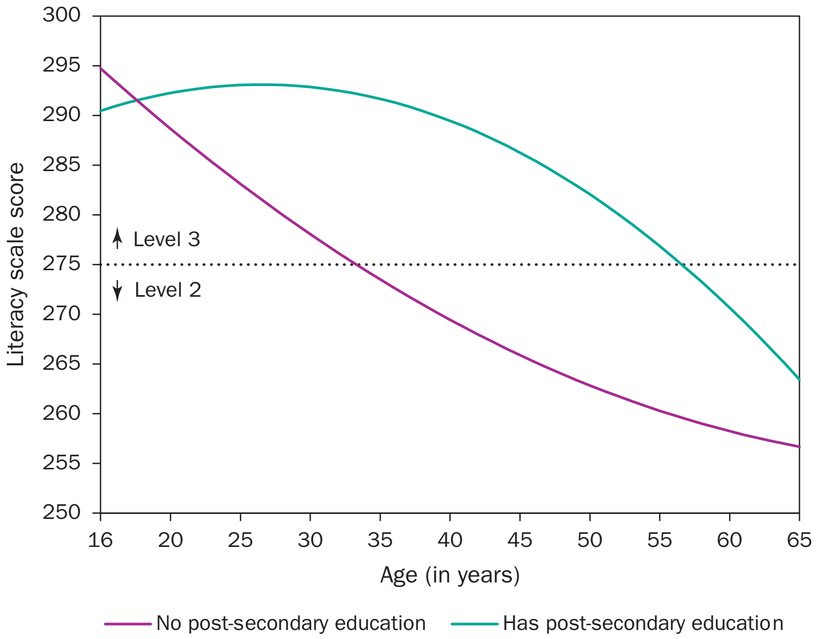

Figure 14.8 **Literacy Scores Predicted by a Multiple Linear Regression That Uses Both a Quadratic Variable and Interaction Variables**

Source: Author generated; Calculated using data from Statistics Canada, 2017.

the mid-twenties, and then level off. Among people with a post-secondary education, after age 26 each older age cohort is then predicted to have a lower literacy score. In other words, this relationship is not monotonic—it changes direction, with the curve peaking in the mid-twenties.

These regression results make it clear that the relationship between age and literacy scores is both curvilinear and influenced by people's level of education. The lower literacy scores among older age cohorts can partly be attributed to their lower educational credentials. But, even when having a post-secondary educational credential is accounted for, people in age cohorts beyond 35 have lower literacy skills than their younger counterparts. Recall that most jobs in Canada require Level 3 literacy skills. Among those without a post-secondary education, people aged 34 or older are predicted to have literacy scores that are below this cut-off. In contrast, among those with a post-secondary education, only people aged 57 or older are predicted to have literacy scores that are below this cut-off. Overall, these findings reinforce the importance of post-secondary education in developing and maintaining a highly literate workforce.

Transforming Skewed Variables

Researchers also sometimes transform independent variables in order to create better-fitting regression models—that is, models that explain more of the variation in the dependent variable or models with smaller residuals (or smaller errors). A **transformation** refers to replacing the values on a variable with values that are a

transformation Replacing the values on a variable with values that are a mathematical function of the original value.

mathematical function of the original value. You are probably already familiar with simple transformations—you do this whenever you convert your weight in pounds to your weight in kilograms or your height in feet and inches to your height in centimetres. The actual thing that you are measuring (weight or height) doesn't change when you transform it; only the units of measurement change. Ratio-level variables can be transformed; however, because nominal- and ordinal-level variables have arbitrary values, it doesn't make sense to mathematically manipulate them.

linear transformation A transformation where the relative sequence of cases and the relative distance between the cases remains the same in the original variable and the transformed variable.

non-linear transformation A transformation where the relative sequence of cases remains the same in the original variable and the transformed variable, but the relative distance between the cases changes.

When a variable undergoes **a linear transformation**, both the relative sequence of cases (from smallest to largest) and the relative distance between the cases remain the same in the new variable. But in the context of regression, **non-linear transformations** are often more useful for improving the fit of a model. In a non-linear transformation the relative sequence of cases (from smallest to largest) remains the same, but the relative distance between the cases changes. Non-linear transformations are typically used when a researcher wants to use a highly skewed variable as an independent variable in a regression. When an independent variable (or a dependent variable) is highly skewed, the regression residuals are often also skewed, which violates the regression assumption that the errors are normally distributed. Independent variables that are highly skewed also typically result in some cases having very large regression residuals or errors. In these circumstances, transforming the skewed variable using a non-linear transformation can help to make the distribution of regression residuals more normal and improve the regression predictions overall.

logarithmic transformation A non-linear transformation where each value on the original variable is expressed as a common base number raised to an exponent, and that exponent is assigned as the value on the transformed variable.

For variables that are right-skewed—that is, variables with a distribution that has a long tail trailing off to the right—the most common non-linear transformation that researchers use is called a **logarithmic transformation**, or a log transformation. The first step in a logarithmic transformation is to express each value on the original variable as a common base number raised to an exponent. Any positive number can be used as the common base number, but social scientists regularly use either base 2 or base 10. A logarithmic transformation using base 2 is denoted as $\log_2$ and a logarithmic transformation using base 10 is denoted as $\log_{10}$. Then, the transformed variable is created by assigning each case the value of the *exponent* that is produced when the original value on the variable is represented as the common base number raised to an exponent.

Let's illustrate this process using a base 2 transformation. A case with the value "2" on the original variable is assigned the value "1" on the transformed variable because 2 is equal to 2^1. Similarly, a case with the value "4" on the original variable is assigned the value "2" on the transformed variable, because 4 is equal to 2^2, and a case with the value "8" on the original variable is assigned the value "3" on the transformed variable, because 8 is equal to 2^3. (See Table 14.11.) But exponents don't need to be whole numbers. For instance, a case with the value "3" on the original variable is assigned the value "1.58" on the transformed variable because 3 is equal to $2^{1.58}$.

When a variable is log-transformed using base 2, each one-unit increase in the transformed variable is equivalent to doubling the original value. So, in the transformed variable shown in the final column of Table 14.11, Amira's value is 1 unit higher than Alan's value. Looking at the values on the original variable, you can see

Table 14.11 Log-Transforming a Variable Using Base 2 (Hypothetical Data)

Person	Value on Original Variable	Value in Base 2 Exponent Form	Value on the Transformed Variable ($\log_2$)
Asmita	1	2^0	0
Parv	2	2^1	1
Terry	3	$2^{1.58}$	1.58
Alan	4	2^2	2
Nuvdeep	5	$2^{2.32}$	2.32
Josh	6	$2^{2.58}$	2.58
Chloe	7	$2^{2.81}$	2.81
Amira	8	2^3	3
Becky	9	$2^{3.17}$	3.17
Liam	10	$2^{3.32}$	3.32

that Amira has double the value that Alan does (8 compared to 4). Understanding what a one-unit increase in the transformed variable represents will help you to interpret slope coefficients when transformed variables are used as independent variables in a linear regression.

Figure 14.9 shows how a non-linear log transformation compares to a linear transformation. Both graphs in Figure 14.9 plot the values on a hypothetical original variable on the horizontal axis (x-axis), and the values on a corresponding transformed variable on the vertical axis (y-axis). The left panel shows a simple linear transformation of the original variable: each original value is multiplied by two. In the linear-transformed variable, shown on the vertical axis, the cases stay in the same relative order (smallest to largest) as in the original variable. The relative distance (height) between the cases also stays the same. A one-unit increase in the original variable is equivalent to a two-unit increase in the linear-transformed variable for every case. The right panel of Figure 14.9 shows a $\log_2$ transformation of the original variable that corresponds to Table 14.11. In the $\log_2$ transformed variable, the cases stay in the same relative order (smallest to largest) as in the original variable. But the relative distance (height) between the cases changes in the $\log_2$ transformed variable. The higher the value on the original variable, the smaller the height difference between the cases in the transformed variable. The curved pattern of the dots shows that the transformation is non-linear—that is, it is not a straight line.

When a variable is log-transformed, cases with low values on the original variable are moved farther apart in the new variable, and cases with high values on the original variable are moved closer together in the new variable. This is why log transformations are useful for variables that are right-skewed—they move the cases with high values in the long tail on the right closer together, and move the cases with the low values that are clustered together on the left farther apart.

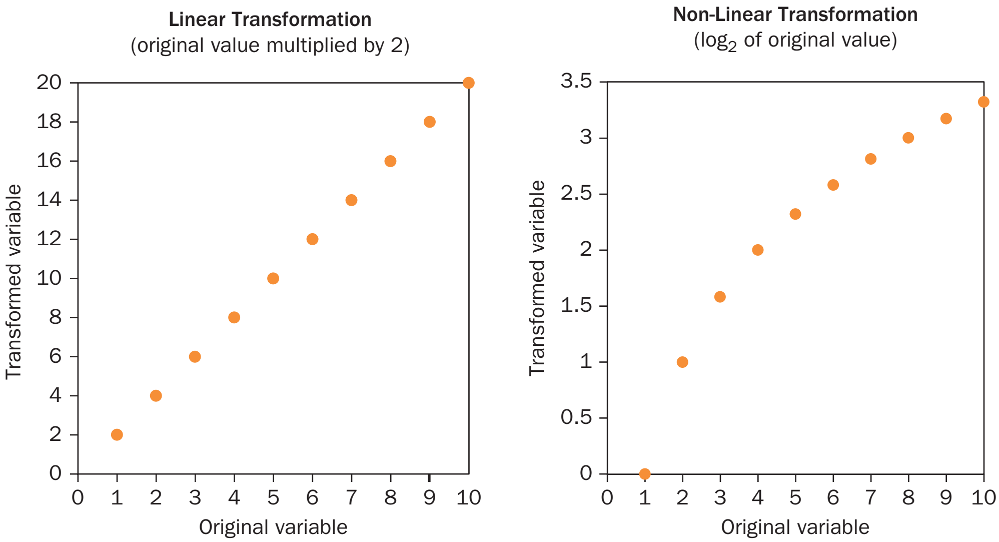

Figure 14.9 **Comparing a Linear Transformation and a Non-Linear Transformation for the Same Variable (Hypothetical Data)**

An example from the PIAAC data is useful for illustrating how log transformations work. The dataset includes a variable that captures the number of hours people spend participating in non-formal education each year (people who did not participate in any non-formal education in the past 12 months are excluded from the analysis). Non-formal educational activities are defined as organized and sustained educational activities that are outside the formal "ladder" system of schooling. They can include educational programs to support adult literacy, life skills, work skills, cultural interests, or basic education for people who are out of school. Types of non-formal education include open and distance education, on-the-job training sessions, as well as other types of seminars, workshops, courses, or private lessons (PIAAC 2011). The distribution of time spent participating in non-formal education is right-skewed. (See the left panel of Figure 14.10.) About a third of people spend less than 20 hours participating in non-formal education each year. But the long tail on the right of the distribution shows that there are a few people who participate in hundreds of hours of non-formal education each year. The taller bar at 800 hours reflects the fact that this variable is top-coded. Everyone who would have been spread out over the values larger than 800 hours is clustered together at this value. The original variable capturing the amount of time spent participating in non-formal education was log-transformed using base 2. The distribution of the $\log_2$-transformed variable is shown in the right panel of Figure 14.10. The same cases are shown, in the same relative order, but the distance between the cases is changed by the log transformation. As you can see, the $\log_2$-transformed variable is no longer highly skewed and is closer to being normally distributed. The centre of the distribution is at 5, which represents 2^5 or 32 hours participating in non-formal education.

Log-transformed variables can be used as independent variables in a regression in the same way as any other ratio-level variable. They replace the original

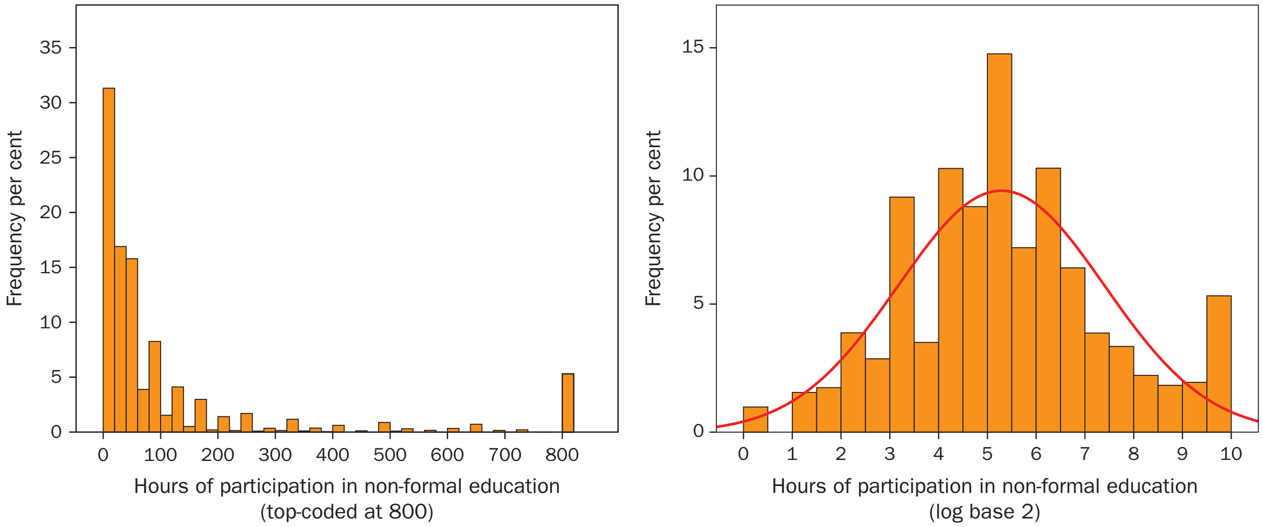

Figure 14.10 **The Distribution of Time Spent Participating in Non-Formal Education, Original Variable and $\log_2$-Transformed Variable**

Source: Author generated; Calculated using data from Statistics Canada, 2017.

variable in the regression: when a log-transformed variable is used as an independent variable, the original variable is omitted from the regression. The slope coefficients of log-transformed variables are interpreted in relation to the base of the log. For instance, when a variable is log-transformed to base 2, a one-unit increase in the transformed variable is equivalent to doubling the original value; therefore, the slope coefficient of the transformed variable shows the predicted change in the dependent variable that is associated with doubling the original value on the independent variable.

Table 14.12 shows the results of a linear regression predicting literacy scores using the $\log_2$-transformed variable capturing hours spent participating in non-formal education. The constant coefficient indicates that people who spend one hour engaged in non-formal education each year are predicted to have a literacy score of 280. Recall that the constant coefficient shows the predicted value on the dependent variable when the independent variable equals 0. A “0” value on the $\log_2$-transformed independent variable represents people who spend 2^0 hours participating in non-formal education, and 2^0 is equal to 1. (Any base to the power of 0 is equal to 1.) This is why the constant coefficient shows the predicted value on the dependent variable for people who spend one hour participating in non-formal education each year. The slope coefficient shows that every doubling of hours spent participating in non-formal education is associated with an increase of 1.39 in people’s literacy scores. Notice that it is the *percentage* change in hours that matters, not the absolute change in hours. So, people who spend two hours participating in non-formal education each year are predicted to have a literacy score that is 1.39 points higher than people who spend one hour participating in non-formal education: the one-hour difference represents a 100 per cent increase in time, or a doubling of time. But people who spend eight hours participating in non-formal

Table 14.12 Results of a Simple Linear Regression with a Log-Transformed Independent Variable

Dependent variable: Literacy scale score (n = 12,969)

	Unstandardized Coefficient	Standardized Coefficient
Constant	280.47*	—
Hours spent participating in non-formal education in past 12 months (log base 2)	1.39*	0.07
Adjusted R^2	*0.00*	

* Indicates that results are statistically significant at the p< 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2017.

education each year are predicted to have a literacy score that is 1.39 points higher than people who spend four hours participating in non-formal education: the four-hour difference also represents a 100 per cent increase in time, or a doubling of time.

The very low adjusted R^2 of this regression indicates that the number of hours spent participating in non-formal education does not explain much of the variation in literacy scores. In part, this is because this regression only includes people who reported spending at least some time participating in non-formal education. Regardless of the base that is used, before any variable is log-transformed, all values must be above 0. It is not possible to represent "0" values or negative values in a log-transformed variable. People who spend zero hours participating in non-formal education do not have a value on the log-transformed variable and, thus, are excluded from this regression.

When a variable is log-transformed to base 10, a one-unit increase in the transformed variable is equivalent to a tenfold increase in the original value. When a $\log_{10}$ transformed variable is used as an independent variable in a regression, the slope coefficient of the transformed variable shows the predicted change in the dependent variable that is associated with a tenfold increase in the original value on the independent variable. An example of a base 10 log transformation is shown in Table 14.13. Notice that values on the original variable between "0" and "1" become negative exponents when they are log-transformed. (This occurs for all bases.)

Log-transformed variables can also be used as the dependent variable in a regression, although the interpretation of the slope coefficients becomes more complex. A discussion of how to interpret regression coefficients when a dependent variable is log-transformed is beyond the scope of this book. But in Chapter 15 you'll learn how log transformations are used in logistic regression, which is a type of regression that is used when the dependent variable is a dichotomous variable.

So far, I have only described how to transform variables that are right-skewed. In social science data, right-skewed variables are far more common than left-skewed variables. But left-skewed variables can also be transformed so that they are more normally distributed. Left-skewed variables are typically transformed by taking the square root (or the cube root) of the original variable. Interpreting the slope coefficients of variables that are transformed this way is more complex than

Table 14.13 Log-Transforming a Variable Using Base 10 (Hypothetical Data)

Person	Value on the Original Variable	Value in Base 10 Exponent Form	Value on the Transformed Variable ($\log_{10}$)
Aaron	0.1	10^{-1}	−1
Emma	0.5	$10^{-0.3}$	−0.3
Sohailia	1	10^{0}	0
Amanda	5	$10^{0.7}$	0.7
Millicent	10	10^{1}	1
Mike	50	$10^{1.7}$	1.7
Jeanette	100	10^{2}	2
Jacob	500	$10^{2.7}$	2.7
Sergio	1,000	10^{3}	3
Louisa	5,000	$10^{3.7}$	3.7
Ricardo	10,000	10^{4}	4

for log-transformed variables. Given how rarely this technique is used, I do not discuss it further here.

Since highly skewed variables can result in poor-fitting regression models, it's important for researchers to examine the distribution of each of the variables in a regression as they build their model. For variables that are highly skewed, transforming the variable so that the distribution becomes more normal can substantially improve the regression predictions.

What You Have Learned

In this chapter, you learned several strategies for manipulating independent variables in order to investigate more complex relationships and to improve the fit of a regression model. You learned how to use interaction variables to investigate how two independent variables jointly influence a dependent variable. You also learned how to use quadratic independent variables in a linear regression in order to predict curvilinear relationships. Finally, you learned how to transform skewed variables so that they can be used as independent variables in a linear regression without violating any distributional assumptions. Particular emphasis was placed on log transformations of right-skewed variables since these occur more often than left-skewed variables in social science data.

The research focus of this chapter was adults' literacy scores and how they are related to age cohort and level of education. An analysis of Canada's PIAAC data shows that literacy scores are predicted to be the highest for adults between the ages of 25 and 40, and lower for people in both younger and older age cohorts. Most notably, education is predicted to be strongly related to literacy skills, both directly and indirectly. Directly, people with higher educational credentials are predicted to have higher literacy scores. Indirectly, the negative relationship between age and literacy scores is moderated by people's level of education; among people with a post-secondary education, older age cohorts have less of a decrease in literacy scores than among people without a post-secondary education. This finding, along with the established link between literacy skills and income (Heisz, Notten, and Situ 2016), reinforces the long-term benefits of investing in a post-secondary education.

Check Your Understanding

Check to see if you understand the key concepts in this chapter by answering the following questions:

1. Why do researchers use interaction variables as independent variables in linear regression?
2. How does the interpretation of the slope coefficients change when an interaction variable is used as an independent variable in a linear regression?
3. Why do researchers use quadratic variables as independent variables in linear regression?
4. Why do researchers log-transform variables?
5. How does the interpretation of the constant coefficient and the slope coefficients change when a log-transformed variable is used as an independent variable in a linear regression?

Practice What You Have Learned

Check to see if you can apply the key concepts in this chapter by answering the following questions. Keep two decimal places in any calculations.

1. In "Practice What You Have Learned" in Chapter 13, you interpreted a multiple linear regression model developed by a community agency to predict people's charitable giving. It showed that two indicators of community engagement—whether people volunteered their time in the past 12 months and whether they participate in religious activities/services at least once a month—were both strongly related to the amount of money that people donated to charitable organizations. The agency's research team decides to investigate whether the relationship between participating in religious activities/services at least once a month and the total amount of money that people donated to charitable organizations is different for people who volunteer and for people who do not volunteer. Table 14.14 shows a multiple linear regression that uses people's volunteer status, their participation in religious activities/services, and a variable capturing the interaction between them to predict the total amount of money that people donated to charitable organizations in the past 12 months.
   a. Explain what the unstandardized slope coefficient of the "Volunteered in the past 12 months" dummy variable shows.
   b. Explain what the unstandardized slope coefficient of the "Participates in religious activities/services once a month or more often" dummy variable shows.
   c. Explain what the unstandardized slope coefficient of the interaction variable shows.

Table 14.14 Results of a Multiple Linear Regression with Three Independent Variables (One Interaction Variable)

Dependent variable: Total financial donations to charitable organizations ($) in the past 12 months (n = 13,556)

	Unstandardized Coefficient	Standardized Coefficient
Constant	163.98*	—
Volunteered in the past 12 months	91.19*	0.069
Participates in religious activities/services once a month or more often	280.71*	0.186
Volunteered x religious participation	321.00*	0.171
Adjusted R^2	*0.13*	

*Indicates that results are statistically significant at the p < 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2017.

2. Using the information in Table 14.14:
   a. Write out the regression prediction equation for people who volunteer but who do not participate in religious activities/services at least once a month. Then, calculate the result.
   b. Write out the regression prediction equation for people who participate in religious activities/services at least once a month but who do not volunteer. Then, calculate the result.
   c. Write out the regression prediction equation for people who both volunteer and participate in religious activities/services at least once a month. Then, calculate the result.
3. The community agency's research team is also interested in understanding how the relationship between age and the total amount of money that people donate to charitable organizations is different for people who volunteer and for people who do not volunteer. Table 14.15 shows the results of a multiple linear regression that uses people's age, their volunteer status, and a variable capturing the interaction between them to predict the amount donated to charitable organizations in the past 12 months.
   a. Explain what the unstandardized slope coefficient of the "Age" variable shows.
   b. Explain what the unstandardized slope coefficient of the "Volunteered in the past 12 months" dummy variable shows.
   c. Explain what the unstandardized slope coefficient of the interaction variable shows.
4. Using the information in Table 14.15:
   a. Calculate how much money a 20-year-old who volunteers is predicted to donate to charitable organizations in a 12-month period. Then, repeat the calculation for an 80-year-old who volunteers. Be sure to account for the centring of the age variable in your calculations.
   b. Calculate how much money a 20-year-old who does not volunteer is predicted to donate to charitable organizations in a 12-month period. Then, repeat the calculation for an 80-year-old who does not volunteer.
   c. Compare the influence of volunteering on people's predicted donations to charitable organizations for people who are age 20 and for people who are age 80. Is the relationship between volunteering and the amount of money donated to charitable organizations stronger for young people or older people?

Table 14.15 Results of a Multiple Linear Regression with Three Independent Variables (One Interaction Variable)

Dependent variable: Total financial donations to charitable organizations ($) in the past 12 months (n = 14,538)

	Unstandardized Coefficient	Standardized Coefficient
Constant	208.93*	—
Age (in years, centred on 45)	4.24*	0.119
Volunteered in the past 12 months	257.65*	0.195
Age (in years, centred on 45) x volunteered	6.04*	0.107
Adjusted R^2	*0.07*	

*Indicates that results are statistically significant at the $p < 0.05$ level.
Source: Author generated; Calculated using data from Statistics Canada, 2017.

5. Using the information in Table 14.15:
   a. Calculate how much money people at different ages, ranging from 20 to 80 (i.e., age 20, 30, 40, 50, 60, 70, and 80), are predicted to donate to charitable organizations in a 12-month period, for both volunteers and non-volunteers.
   b. Either by hand or using a spreadsheet program, create a graph showing these predictions.
6. Based on their analyses, the agency's research team is confident that there is a relationship between people's age and the amount of money they donate to charitable organizations. But the researchers want to assess whether this relationship is linear or curvilinear. Table 14.16 shows the results of a multiple linear regression that uses a linear version of the "Age" variable and a quadratic version of the "Age" variable to predict the total amount of money that people donated to charitable organizations in the past 12 months.

Table 14.16 **Results of a Multiple Linear Regression with Two Independent Variables (One Linear and One Quadratic)**

Dependent variable: Total financial donations to charitable organizations ($) in the past 12 months (n = 14,538)

	Unstandardized Coefficient	Standardized Coefficient
Constant	334.35*	—
Age (in years, centred on 45)	6.18*	0.173
Age (in years, centred on 45), squared	−0.06*	−0.031
Adjusted R^2	*0.03*	

*Indicates that results are statistically significant at the p < 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2017.

a. Explain what the unstandardized slope coefficient of the linear "Age" variable shows, in terms of the direction of the relationship between age and charitable giving.

b. Explain what the unstandardized slope coefficient of the quadratic "Age" variable (age squared) shows, in terms of whether and how the relationship between age and charitable giving is curvilinear.

7. Using the information in Table 14.16:

   a. Write out the regression prediction equation for a person who is 20 years old. Be sure to account for the centring of the age variable. Then, calculate the result.

   b. Write out the regression prediction equation for a person who is 50 years old, and calculate the result.

   c. Write out the regression prediction equation for a person who is 80 years old, and calculate the result.

8. Using the information in Table 14.16 and the approach you used in question 7, calculate how much money people at different ages ranging from 20 to 80 (i.e., age 20, 30, 40, 50, 60, 70, and 80), are predicted to donate to charitable organizations in a 12-month period. Either by hand or using a spreadsheet program, create a graph showing these predictions.

9. The first step in a logarithmic transformation is to express each value on the original variable as a common base number raised to an exponent. Express the following numbers in base 2 exponent form. (You will need to use a calculator or a spreadsheet program to do these calculations.)

   a. 20

   b. 50

   c. 80

10. Table 14.17 shows the results of a simple linear regression that uses a variable capturing the *number* of hours that people volunteered in the past 12 months, log-transformed to base 2, to predict the total amount of money that people donated to charitable organizations in the past 12 months. Only people who volunteered for at least one hour in the past 12 months are included in the analysis.

    a. Explain what the unstandardized slope coefficient of the log-transformed "Hours spent volunteering" variable shows.

    b. Explain what the constant coefficient shows.

11. Using the information in Table 14.17 and your answers to question 9:

    a. Write out the regression prediction equation for people who volunteered for 20 hours in the past 12 months, and calculate the result.

Table 14.17 **Results of a Simple Linear Regression with a Log-Transformed Independent Variable**

Dependent variable: Total financial donations to charitable organizations ($) in the past 12 months (n = 6,278)

	Unstandardized Coefficient	Standardized Coefficient
Constant	170.11*	—
Hours spent volunteering in past 12 months (log base 2)	49.47*	0.143
Adjusted R^2	*0.02*	

*Indicates that results are statistically significant at the p < 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2017.

Table 14.18 Proportion of Canadian-Born University Graduates Aged 25 to 65 at Level 2 or Below in Skill Proficiency Levels across Various Socio-Economic Characteristics

	Level 2 or Below, Literacy	Level 2 or Below, Numeracy
	Percentage	
Gender		
Men (ref.)	13.6	16.8
Women	16.2	26.8*
Age group		
25 to 29	12.3	19.5
30 to 34	11.3	17.4
35 to 39 (ref.)	8.6	17.4
40 to 44	13.7	21.2
45 to 49	14.2	22.0
50 to 54	17.9*	26.8
55 to 59	23.8*	29.0*
60 to 65	23.9*	29.3*
First language spoken (mother tongue)		
English (ref.)	12.7	20.6
French	18.5*	24.0
Other	23.8*	30.3
Parental education		
Less than a high school diploma (ref.)	26.6	34.7
At least one has a high school diploma	18.8	26.8
At least one has a PSE below bachelor	11.7*	18.7*
At least one has a university degree	11.5*	18.0*
Number of books in home at age 16		
10 or less (ref.)	30.9	39.1
11 to 25	20.5	30.5
26 to 100	17.0	24.0*
101 to 200	14.4*	20.8*
More than 200	8.6*	15.4*
Field of study		
STEM (ref.)	9.4	11.5
Humanities, languages, and arts	17.9*	29.4*
Social science, business, and law	13.1	20.9
Teacher training and education science	22.2*	29.1*
Health and welfare	15.5	26.5*
Educational attainment		
Bachelor’s degree (ref.)	16.0	23.4
1st professional degree (medical, veterinary medical, dental, optometry, law, divinity)	13.1	19.3
Master’s or doctorate	13.2	19.9

*Significantly different from the reference category (ref.) ($p < 0.05$).

Note: PSE = Post-secondary education. STEM includes science, technology, engineering, mathematics, and computer science.

Source: Programme for the International Assessment of Adult Competencies (PIAAC) 2012.

Source: Excerpt from Hango 2014, 5.

b. Write out the regression prediction equation for people who volunteered for 50 hours in the past 12 months, and calculate the result.

c. Write out the regression prediction equation for a people who volunteered for 80 hours in the past 12 months, and calculate the result.

12. Table 14.18, excerpted from a Statistics Canada report, shows the relationship between low levels of literacy or numeracy and a series of socio-economic characteristics (Hango 2014). In particular, it shows the percentage of Canadian-born university graduates in each category with Level 2 literacy or less, and the percentage with Level 2 numeracy or less. (People with Level 2 numeracy or less can typically only perform simple mathematical operations and cannot understand complex mathematical information or engage in problem-solving strategies.)

    a. Describe how gender is related to literacy and numeracy levels.

    b. Describe how age is related to literacy and numeracy levels.

    c. Describe how people's first language spoken (mother tongue) is related to literacy and numeracy levels.

13. Social scientists know that the socio-economic status of a person's family is related to that person's literacy and numeracy skills. Using the information in Table 14.18:

    a. Describe how parental education is related to literacy and numeracy levels.

    b. Describe how the number of books in people's homes is related to literacy and numeracy levels.

14. Using the information in Table 14.18:

    a. Describe how people's field of study is related to their literacy and numeracy levels.

    b. Describe how people's level of educational attainment is related to their literacy and numeracy levels.

Practice Using Statistical Software (IBM SPSS)

Answer these questions using IBM SPSS and the GSS27.sav or the GSS27_student.sav dataset available from the Student Resources area of the companion website for this book. Weight the data using the "Standardized person weight" [STD_WGHT] variable you created following the instructions in Chapter 5. Report two decimal places in your answers, unless fewer are printed by IBM SPSS. It is imperative that you save the dataset to keep any new variables that you create.

1. Use the Compute Variable tool to create an interaction variable by multiplying the "Women" [WOMEN] dummy variable by the "Visible minority" [IS_VISMIN] dummy variable. (You created these variables in "Practice Using Statistical Software" in Chapters 12 and 13, respectively.) To ensure that cases with missing values are treated appropriately, use the If Cases option to assign a value on the new variable only if the value of IS_VISMIN is greater than or equal to "0" (IS_VISMIN >= 0). Call the new variable "Visible minority women" [VISMIN_WOMEN]. Produce a cross-tabulation of the "Women" [WOMEN] and "Visible minority" [IS_VISMIN] variables, and compare it to a frequency distribution of "Visible minority women" [VISMIN_WOMEN] to ensure that the variable was created correctly.

2. Use the Linear Regression procedure to produce a regression of the independent variables "Women" [WOMEN], "Visible minority" [IS_VISMIN], and "Visible minority women" [VISMIN_WOMEN] on the dependent variable "Number of paid hours worked per week - All jobs" [WKWEHRC].

    a. Explain what the constant coefficient shows.

    b. Explain what each of the unstandardized slope coefficients show.

    c. Use the regression results to calculate the predicted number of paid hours of work each week for the following:

- Women who are visible minorities
- Men who are visible minorities
- Women who are not visible minorities
- Men who are not visible minorities

3. Create a variable that captures the interaction between age and having a post-secondary educational credential.
   a. Use the Recode into Different Variables tool to recode the "Education - Highest degree (4 categories)" [DH1GED] variable into a "Has a postsecondary education" [POSTSECONDARY] dummy variable. In the new variable, assign the value "1" to people who have a post-secondary diploma or a university degree, and assign the value "0" to people who only have a high school education or have less than a high school education. The remaining value can be designated as system-missing in the new variable. Produce frequency distributions of the original variable "Education - Highest degree (4 categories)" [DH1GED] and the new variable "Has a postsecondary education" [POSTSECONDARY], and compare them to be sure that the recoding is correct.
   b. Use the Compute Variable tool to create an interaction variable by multiplying the "Has a postsecondary education" [POSTSECONDARY] dummy variable by the "Age (centred)" [AGE_CENTRED] variable (which you created in question 1[b] of "Practice Using Statistical Software" in Chapter 13). Call the new variable "Age x Postsecondary education" [AGE_POSTSEC]. Find the mean of the new interaction variable.

4. Use the Linear Regression procedure to produce a regression of the independent variables "Age (centred)" [AGE_CENTRED], "Has a postsecondary education" [POSTSECONDARY], and "Age x Postsecondary education" [AGE_POSTSEC] on the dependent variable "Number of paid hours worked per week - All jobs" [WKWEHRC].
   a. Explain what the constant coefficient shows.
   b. Explain what each of the unstandardized slope coefficients show.
   c. Use the regression results to calculate the predicted number of paid hours worked each week for the following:
      - A 20-year-old with a post-secondary education
      - A 20-year-old without a post-secondary education
      - A 65-year-old with a post-secondary education
      - A 65-year-old without a post-secondary education

   Be sure to account for centring of the age variable in your calculations.

5. Use the Compute Variable tool to create a quadratic variable by multiplying the "Age (centred)" [AGE_CENTRED] variable by itself. Call the new variable "Age (squared)" [AGE_SQUARED]. Find the mean of the new quadratic variable.

6. Use the Linear Regression procedure to produce a regression of the independent variables "Age (centred)" [AGE_CENTRED] and "Age squared" [AGE_SQUARED] on the dependent variable "Number of paid hours worked per week - All jobs" [WKWEHRC].
   a. Explain what the unstandardized slope coefficient of the linear "Age" variable shows, in terms of the direction of the relationship between age and the number of hours that people work each week.
   b. Explain what the unstandardized slope coefficient of the quadratic "Age" variable shows, in terms of whether and how the relationship between age and the number of hours that people work each week is curvilinear.

7. The distribution of the variable "Number of new people met - Past month" [SCP_110] is right-skewed.
   a. In the *Variable View*, designate the value "0" as missing for this variable.
   b. Use the Compute Variable tool to create a version of this variable that is log-transformed to base 2, by programming this command into the

"Numeric expression" box: LN(SCP_110)/LN(2). Call the new variable "Number of new people met - past month (log base 2)" [SCP_110_LOG2]. Find the mean of the new log-transformed variable.

8. Use the Linear Regression procedure to produce a regression of the independent variable "Number of new people met - past month - Log base 2" [SCP_110_LOG2] on the dependent variable "Number of close friends" [SCF_100C].

   a. Explain what the constant coefficient shows.
   b. Explain what the unstandardized slope coefficient of the log-transformed variable shows.

References

Hango, Darcy. 2014. "University Graduates with Lower Levels of Literacy and Numeracy Skills." Catalogue no. 75–006–X. Insights on Canadian Society. Ottawa: Statistics Canada. http://www.statcan.gc.ca/pub/75-006-x/2014001/article/14094-eng.pdf.

Harwood, Chris. 2012. "State of the Literacy and Essential Skills Field." Ottawa: Canadian Literacy and Learning Network. http://www.literacy.ca/content/uploads/2012/12/State-of-the-Field-2012.pdf.

Heisz, Andrew, Geranda Notten, and Jerry Situ. 2016. "The Association between Skills and Low Income." Catalogue no. 75-006-X. Ottawa: Statistics Canada. http://www.statcan.gc.ca/pub/75-006-x/2016001/article/14322-eng.pdf.

Murray, Scott, and Richard Shilington. 2011. "From Poverty to Prosperity: Literacy's Impact on Canada's Economic Success." Ottawa: Canadian Literacy and Learning Network. http://www.literacy.ca/content/uploads/2012/09/Poverty-to-Prosperity-full-report.pdf.

Organisation for Economic Co-operation and (OECD). 2012. *Literacy, Numeracy and Problem Solving in Technology-Rich Environments*. Paris: OECD Publishing. http://www.oecd-ilibrary.org/education/literacy-numeracy-and-problem-solving-in-technology-rich-environments_9789264128859-en.

———. 2016. *Skills Matter: Further Results from the Survey of Adult Skills*. Paris: OECD Publishing. http://www.oecd-ilibrary.org/education/skills-matter_9789264258051-en.

———. n.d. "Key Facts about the Survey of Adult Skills (PIAAC)." http://www.oecd.org/skills/piaac/Key%20facts%20about%20the%20Survey%20of%20Adult%20Skills.pdf.

Programme for the International Assessment of Adult Competencies (PIAAC). 2011. "PIAAC Conceptual Framework of the Background Questionnaire Main Survey." http://www.oecd.org/skills/piaac/PIAAC(2011_11)MS_BQ_ConceptualFramework_1%20Dec%202011.pdf.

Programme for the International Assessment of Adult Competencies (PIAAC), Literacy Expert Group. 2009. "PIAAC Literacy: A Conceptual Framework." OECD Education Working Papers 34. http://www.oecd-ilibrary.org/education/piaac-literacy-a-conceptual-framework_220348414075.

Smith, M. Cecil, Amy D. Rose, Jovita Ross-Gordon, and Thomas J. Smith. 2015. "Adults' Readiness to Learn as a Predictor of Literacy Skills." Washington, D.C.: American Institutes for Research-PIAAC. https://static1.squarespace.com/static/51bb74b8e4b0139570ddf020/t/54da7802e4b08c6b90107b4f/1423603714198/Smith_Rose_Ross-Gordon_Smith_PIAAC.pdf.

Statistics Canada. 2015. General Social Survey, 2013: Cycle 27, Giving, Volunteering and Participating. *Public Use Microdata File*. Ottawa, ON: Statistics Canada.

———. 2017. "Programme for the International Assessment of Adult Competencies (PIAAC) 2012." http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4406.

Tilleczek, Kate, and Valerie Campbell. 2013. "Barriers to Youth Literacy: Sociological and Canadian Insights." *Language and Literacy* 15 (2): 77–95.

United Nations Educational Scientific and Cultural Organization (UNESCO). 2016. "Literacy." *Education*. http://www.unesco.org/new/en/education/themes/education-building-blocks/literacy/browse/3/.

Willms, J. Douglas, and Barry Watson. 2008. *Literacy, Numeracy and Problem-Solving Skills of Canadian Youth*. Gatineau: Human Resources and Social Development Canada.