Multivariate Regression Analysis

9

Abstract

This final chapter provides an introduction into multivariate regression modeling. We will cover the logic behind multiple regression modeling and explain the interpretation of a multivariate regression model. We will further cover the assumptions this type of model is based upon. Finally, and using our data, we will provide concrete examples on how to interpret a multiple regression model.

9.1 The Logic Behind Multivariate Regression Analysis

Bivariate regression analysis is very rarely used in real applied research, because an outcome is hardly ever just dependent on one predictor. Rather, multiple factors normally explain the dependent variable (see Fig. 9.1). To highlight, if we want to explain a student's grade in an exam, several factors might come into play. A student's grade might depend on how much the respective student studied for the exam; it might depend on her health and even on her general mood. Multiple regression modeling allows us to absolutely and comparatively gauge the influence of all of these factors on the dependent variable.

Multiple regression analysis is an extension of bivariate regression analysis. It allows us to test the influence of multiple independent (predictor) variables on a dependent variable. Just like in the case of two variables, the goal of this method is to create an equation or a "model" that explains the impact of/relationship between these variables.

Let us assume that we want to explain the dependent variable "Y" and we have several independent variables $X_1 \ldots X_p$. Then, the multiple regression equation we need to calculate is:

D. Stockemer, *Quantitative Methods for the Social Sciences*,
https://doi.org/10.1007/978-3-319-99118-4_9

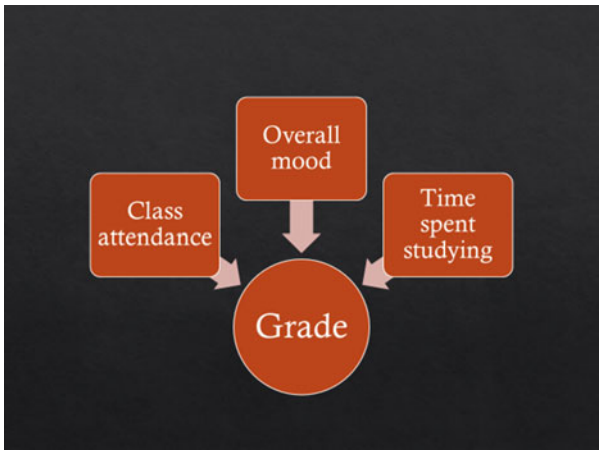


Fig. 9.1 Predictors of a student's grade

$$Y' = A + B_1X_1 + B_2X_2 + \ldots + B_pX_p$$

$\boldsymbol{Y'}$ is the predicted score of the dependent (criterion) variable (*dependent variable*).
$\boldsymbol{A}$ is a constant which gives the value of Y' when all Xs are zero (*Y-intercept or constant*).
Xs are all the independent variable values (*values of the independent variables*).
$\boldsymbol{B_1}$**–**$\boldsymbol{B_p}$ are regression weights. They are the contribution of each *independent* variable to the predicted value of the dependent variable (i.e., each represents the change in Y' resulting from a *unit* change in a specific predictor variable *when all other predictors are held at constant values*).

Example Suppose we want to study the predictors of a student's grade in a math exam (see Figure 8.21). We ask a random sample of students at a German high school about their grade in their last math exam, the time they spent studying for this exam, their general mood, and whether they were in good perceived health when taking the exam. Let us further assume that we run a multiple regression model and receive the following equation (for now we ignore the question of whether the variables are statistically significant or not). We receive the following equation:

$$Y' = 10.5 + 3.1X_1 + 1.5X_2 + 0.5X_3 + e$$

We would interpret the model as follows:

10.5 (on a scale from 0 to 100) is the hypothetical grade a student is expected to get, if she does not study at all, her general health is at its worst (she would rank her health by the value 0), and her general mood is also at the lowest value (she would also rank this at 0).

3.1 is the slope coefficient for the variable study time. This implies that for every hour a student studies, her grade is expected to increase by 3.1 points.

1.5. is the slope coefficient for somebody's general health, indicating that per each point somebody's perceived health increases, her math grade is predicted to improve by 1.5 points.

0.5 is the slope coefficient for somebody's general mood. In other words, for each point somebody's general mood increases, her test performance is expected to increase by 0.5 points.

As we can see from the example, the multivariate regression model is an extension of the bivariate model. It has the same parameters and the interpretation is analogous. To do a multiple regression analysis in SPSS (or Stata), follow the same steps as you would follow for a bivariate regression. Just add more variables as independent variables.

9.2 The Functional Forms of Independent Variables to Include in a Multivariate Regression Model

In this introduction to survey research and quantitative methods, we only cover continuous dependent variables (noncontinuous dependent variables will be the subject of more advanced statistical courses). Yet, we still have to determine in what type of functional form we would include our independent variables. Provided that the relationship between a continuous independent and a continuous dependent variable is linear (the relationship follows a line), we will include the independent variables in its linear form. If a scatterplot would highlight that the relationship between independent and dependent variable is not linear (e.g., it follows a curve), we would need to transform the variable. However, this is also material for a more advanced class. We would also not change binary or dummy variables. The only variables we must be careful with, for the purpose of this introductory textbook, are categorical variables. If we have a categorical nominal variable (i.e., different religious affiliations), we create $N - 1$ dummy variables, with one of the categories serving as a reference category (see Table 4.14). If we have ordinal variables, we could also test whether the relationship is in fact ordered. For example, we could test via a multiple comparison test whether the relationship between times partying per week and money spent partying per week is in principle ordinal. By including the variable—times partying per week—in its linear ordinal form, we also assume that the relationship between partying less than one time per week and one time per week is the same as between four and five times per week. However, in the ANOVA or f-test, we find that this is not true (see Table 6.2). Rather, we find that students that party three times or less regardless of whether they party, on average, less than once, once, twice, or three times per week, spend approximately the same amount of money when they party; only students that party four or more times spend significantly more. Because of this dichotomy in the relationship, it would make sense to

create a dummy variable between the two categories. (In the dataset, we label this variable money spent partying 3.)

9.3 Interpretation Help for a Multivariate Regression Model

When you want to interpret a multivariate regression model, we can follow the same logic as for a bivariate regression model. The four guiding steps can help:

1. Look at what variables are significant.
2. Interpret the substantive value of significant variable.
3. Compare the relative strength of the significant variables.
4. Interpret the model fit.

9.4 Doing a Multiple Regression Model in SPSS

In our sample survey, we have included seven possible predictor variables, and we want to determine the relative and absolute influence of these seven predictor variables on the dependent variable, money spent partying per week. Because we know from the ANOVA analysis (see Sect. 7.2.) that the relationship between the ordinal variable times partying and money spent partying is not linear, but rather only matters for individuals who party four times per week or more, we create a binary variable, coded 0 for partying three times or less per week and 1 for partying four times or more. We add this recoded independent variable together with the remaining six independent variables into the model (see Sect. 9.7) and label it Times_Partying_3. The dependent variable is money spent partying.

9.5 Interpreting a Multiple Regression Model in SPSS

Following the four steps outlined under Sect. 9.3., we can proceed as follows (see Table 9.1):

1. If we look at the significance level, we find that two variables are statistically significant (i.e., quality of extra-curricular activities and times partying 3). For all other variables, the significance level is higher than 0.05. Hence, we would conclude that these indicators do not influence the amount of money students spent per week partying.
2. The first significant variable, the quality of extra-curricular activities, has the expected negative sign indicating that the more the students enjoy their extra-curricular activities at their institution, the less money they spent weekly partying. This observation also confirms our initial hypothesis. Holding everything else constant, the model predicts that per every point a student enjoys her extra-curricular activities more, she spends 62 cents less per week partying.

Table 9.1 Multiple regression output in SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.757[a]	.573	.480	21.74977

a. Predictors: (Constant), Times_Partying_3, Study_Time, Amount_Tuition_Student_Pays, Quality_Extra_Curricular_Activities, Gender, Year, Fun_Without_Alcohol

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20322.322	7	2903.189	6.137	.000[b]
	Residual	15137.678	32	473.052		
	Total	35460.000	39			

a. Dependent Variable: Money_Spent_Partying

b. Predictors: (Constant), Times_Partying_3, Study_Time, Amount_Tuition_Student_Pays, Quality_Extra_Curricular_Activities, Gender, Year, Fun_Without_Alcohol

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	75.467	34.621		2.180	.037
	Study_Time	1.371	2.185	.141	.627	.535
	Gender	-3.316	8.290	-.056	-.400	.692
	Year	1.771	4.152	.065	.427	.673
	Fun_Without_Alcohol	-.058	.284	-.047	-.206	.838
	Quality_Extra_Curricular_Activities	-.588	.180	-.421	-3.272	.003
	Amount_Tuition_Student_Pays	.179	.119	.212	1.498	.144
	Times_Partying_3	26.641	10.373	.387	2.568	.015

a. Dependent Variable: Money_Spent_Partying

For example, this implies that somebody who thinks that the extra-curricular activities are very bad at her university (i.e., she rates the quality of extra-curricular activities at 0) spends 62 dollars more per week studying than somebody who thinks that the extra-curricular activities are excellent (i.e., she rates the quality of extra-curricular activities at 100).

The second significant variable, times partying 3, also has the expected positive sign. The regression coefficient of 24.81 indicates that people that party four or more times are expected to spend nearly 25 dollars more on their partying habits than students that party three times or less.

3. If we compare the two statistically significant variables, we find that the standardized beta coefficient is higher for the variable quality of extra-curricular activities (i.e., the standardized beta coefficient is –0.421) than for the variable times partying 3 (0.387). This higher standard beta coefficient illustrates that the variable quality of extra-curricular activities has more explanatory power in the model than the variable times partying 3.
4. The model fits the data quite well; the seven independent variables explain 57% of the variance in the dependent variable, the amount of money students spent partying. (The *R*-squared is 0.568.)

9.6 Doing a Multiple Regression Model in Stata

In our survey, we have included seven possible predictor variables, and we want to determine the relative and absolute influence of these seven predictor variables on the dependent variable. Because we know from the ANOVA analysis (see Table 6.2) that the relationship between the ordinal variable times partying and money spent partying is not linear but rather only becomes different for individuals who party four times or more, we create a binary variable, coded 0 for partying three times or less per week and 1 for partying four times or more. We add this recoded independent variable together with the remaining six independent variables into the model (see Sect. 9.8). The dependent variable is money spent partying.

9.7 Interpreting a Multiple Regression Model in Stata

Following the four steps outlined under 10.3., we can proceed as follows (see Tables 9.2 and 9.3):

Table 9.2 Multiple regression output in Stata

Source	SS	df	MS	Number of obs	=	40
				F(7, 32)	=	6.14
Model	20322.3215	7	2903.18879	Prob > F	=	0.0001
Residual	15137.6785	32	473.052452	R-squared	=	0.5731
				Adj R-squared	=	0.4797
Total	35460	39	909.230769	Root MSE	=	21.75

Money_Spent_Partying	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Study_Time	1.370689	2.184869	0.63	0.535	-3.079744	5.821122
Gender	-3.316106	8.290119	-0.40	0.692	-20.20253	13.57031
Year	1.771348	4.152033	0.43	0.673	-6.686067	10.22876
Fun_Without_Alcohol	-.0583414	.2836462	-0.21	0.838	-.6361098	.519427
Quality_Extra_Curricular_Activ	-.5883588	.1798321	-3.27	0.003	-.9546649	-.2220527
Amount_Tuition_Student_Pays	.1788863	.1194251	1.50	0.144	-.0643748	.4221473
Times_Partying_3	26.6411	10.3729	2.57	0.015	5.512191	47.77
_cons	75.46679	34.62063	2.18	0.037	4.946871	145.9867

Table 9.3 Multiple regression output in Stata with standardized coefficients

Source	SS	df	MS	Number of obs	=	40
				F(7, 32)	=	6.14
Model	20322.3215	7	2903.18879	Prob > F	=	0.0001
Residual	15137.6785	32	473.052452	R-squared	=	0.5731
				Adj R-squared	=	0.4797
Total	35460	39	909.230769	Root MSE	=	21.75

Money_Spent_Partying	Coef.	Std. Err.	t	P>|t|	Beta
Study_Time	1.370689	2.184869	0.63	0.535	.1406508
Gender	-3.316106	8.290119	-0.40	0.692	-.055618
Year	1.771348	4.152033	0.43	0.673	.0654421
Fun_Without_Alcohol	-.0583414	.2836462	-0.21	0.838	-.04718
Quality_Extra_Curricular_Activ	-.5883588	.1798321	-3.27	0.003	-.4212501
Amount_Tuition_Student_Pays	.1788863	.1194251	1.50	0.144	.2120732
Times_Partying_3	26.6411	10.3729	2.57	0.015	.387448
_cons	75.46679	34.62063	2.18	0.037	.

1. If we look at the significance level, we find that two variables are statistically significant (i.e., quality of extra-curricular activities and times partying 3). For all other variables, the significance level is higher than 0.05. Hence, we would conclude that these indicators do not influence the amount of money students spent per week partying.
2. The first significant variable, the quality of extra-curricular activities, has the expected negative sign indicating that the more students enjoy their extra-curricular activities at their institution, the less money they spent weekly partying. This observation also confirms our initial hypothesis. Holding everything else constant, the model predicts that per every point a student enjoys her extra-curricular activities more, she spends 62 cents less per week partying. For example, this implies that somebody who thinks that the extra-curricular activities are very bad at her university (i.e., she rates the quality of extra-curricular activities at 0) spends 62 dollars more per week studying than somebody who thinks that the extra-curricular activities are excellent (i.e., she rates the quality of extra-curricular activities at 100).

   The second significant variable, times partying 2, also has the expected positive sign. The regression coefficient of 24.81 indicates that people that party four or more times are expected to spend nearly 25 dollars more on their weekly partying habits than students that party three times or less.
3. If we compare the two statistically significant variables, we find that the standardized beta coefficient is higher for the variable quality of extra-curricular activities (i.e., the standardized beta coefficient is –0.421) than for the variable times partying 3 (0.387). This higher standard beta coefficient illustrates that the variable quality of extra-curricular activities has more explanatory power in the model than the variable times partying 3.

4. The model fits the data quite well; the seven independent variables explain nearly 57% of the variance in the dependent variable, the amount of money students spent partying. (The *R*-squared is 0.573.)

9.8 Reporting the Results of a Multiple Regression Analysis

In the multiple regression analysis (see Table 9.2), we evaluated the influence of seven independent variables (the quality of extra-curricular activities, students' study time per week, the year students are in, gender, whether they party two times or less or three times or more per week, the degree to which they think that they can have fun without alcohol, and the amount of tuition the students pay) on the dependent variable, the weekly amount of money students spent partying. We find that two of the seven variables are statistically significant and show the expected effect; that is, the more students think that the extra-curriculars at their university are good, the less money they spent partying per week. The same applies to students that party few times; they too spend less money going out. In substantive terms, the model predicts that per every point students increase their ranking of the extra-curricular activities at their school, they will spend 59 cents less partying per week. The coefficient for the dummy variable, partying two times or less or three times or more per week, indicates that students that party three or more times are predicted to spend 26 dollars more on their partying habits than students that party less. Using the 95% benchmark, none of the other variables is statistically significant. Consequently, we cannot interpret the other coefficients because they are not different from zero. In terms of model fit, the data fits the model fairly well: the seven independent variables explain 57% of the variance in the dependent variable.

9.9 Finding the Best Model

In real research the inclusion of variables into a regression model should be theoretically driven; that is, theory should tell us which independent variables we should include in a model to explain and predict a dependent variable. However, we might also be interested in finding the best model. There are two ways to proceed, and there is some disagreement among statisticians: One way is to only include statistically significant variables into the model. Another way is to use the adjusted *R*-squared as a benchmark. To recall, the adjusted *R*-squared is a measure of model fit that allows us to compare different models. For every additional predictor I include in the model, the adjusted *R*-squared increases only if the new term improves the model beyond pure chance. (Please note that a poor predictor can decrease the adjusted *R*-squared, but it can never decrease the *R*-squared.) Using the adjusted *R*-squared as a benchmark to find the best model, we should proceed as follows: (1) start with the complete model, which includes all the predictors, (2) remove the non-statistically significant predictor with the lowest standardized coefficient, and (3) continue this procedure until the

Table 9.4 Finding the best model

	Model 1	Model 2	Model 3	**Model 4**	Model 5
Quality of extra-curricular activities	–0.421	–0.415	–0.416	**0.421**	–0.442
Gender	0.056	0.062	0.051		
Study time per week	0.141	0.200	0.159	**0.140**	
Year of study	0.065	0.051			
Times partying (two times or less/ three times or more	0.387	0.401	0.421	**0.418**	0.418
Fun without alcohol	–0.047				
Amount of tuition student pays	0.179	0.218	0.201	**0.180**	0.150
Constant	75.47	70.22	76.36	**77.53**	92.66
R-squared	0.5731	0.5725	0.5707	**56.87**	0.5502
Adjusted *R*-squared	0.4797	0.4948	0.5075	**51.94**	0.5127

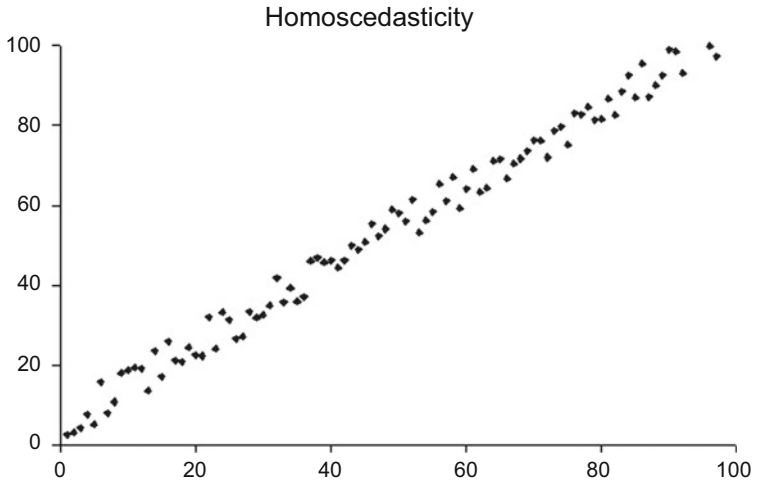
adjusted *R*-squared does no longer increases. Table 9.4 highlights this procedure. We start with the full model. The full model has an adjusted *R*-squared of 0.4797. We take out the variable with the lowed standardized beta coefficient (fun without alcohol). After taking out this variable, we see that the adjusted *R*-squared increases to 0.4948 (see Model 2). This indicates that the variable fun without alcohol does not add anything substantial to the model and should be removed. In a next step, we remove the variable, year of study. Removing this variable leads to another increase in the adjusted *R*-squared (i.e., the new adjusted *R*-squared is 0.5075), indicating again that this variable does not add anything substantively to the model and should be removed (see Model 3). Next, we remove the variable gender and see another increase in the adjusted *R*-squared to 0.5194. If we now remove the variable with the lowest adjusted *R*-squared, the study time per week, we find that the adjusted *R*-squared decreases to 0.5127 (see Model 5), which is lower than the adjusted *R*-squared from Model 4, which is 0.5114. Based on these calculations, we can conclude that Model 4 has the best model fit.

9.10 Assumptions of the Classical Linear Regression Model or Ordinary Least Square Regression Model (OLS)

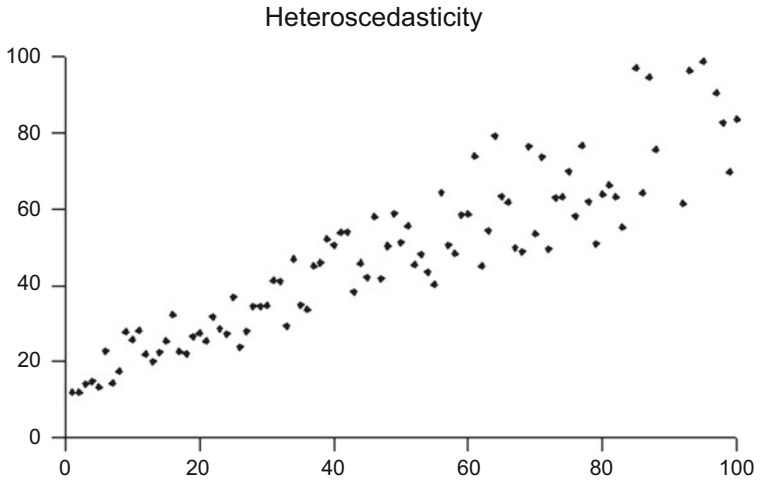
The classical linear regression model (OLS) is the simplest type of regression model. OLS only works with a continuous dependent variable. It has ten underlying assumptions:

1. **Linearity in the parameters**: Linearity in the parameters implies that the relationship between a continuous independent variable and a dependent variable must roughly follow a line. Relationships that do not follow a line (e.g., they might follow a quadratic function or a logarithmic function) must be included into the model using the correct functional forms (more advanced textbooks in regression analysis will capture these cases).

2. ***X* is fixed**: This rule implies that one observation can only have one x and one y value.
3. **Mean of disturbance is zero**: This follows the rule to draw the ordinary least square line. We draw the best fitting line, which implies that the summed up distance of the points below the line is the same as the summed up distance above the line.
4. **Homoscedasticity:** The homoscedasticity assumption implies that the variance around the regression line is similar for all the predictor variables around the regression line (X) (see Fig. 9.2). To highlight, in the first graph, the points are distributed rather equally around a hypothetical line. In the second graph, the points are closer to the hypothetical line at the bottom of the graph in comparison to the top of the graph. In our example, the first graph would be an example of homoscedasticity and the second graph an example of data suffering from heteroscedasticity. At this stage in your learning, it is important that you have heard about heteroscedasticity, but details of the problem will be covered in more advanced textbooks and classes.
5. **No autocorrelation:** There are basically two forms of autocorrelation: (1) contemporaneous correlation, where the dependent variable from one observations affects the dependent variable of another observation in the same dataset (e.g., Mexican growth rates might not be independent because growth rates in the United States might affect growth rates in Mexico), and (2) autocorrelation in pooled time series datasets. That is, past values of the dependent variable influence future values of the dependent (e.g., the US growth rate in 2017 might affect the US growth rate in 2018). This second type of autocorrelation is not really pertinent for cross-sectional analysis but becomes relevant for panel analysis.
6. **No endogeneity:** Endogeneity is one of the fundamental problems in regression analysis. Regression analysis is based on the assumption that the independent variable impacts the dependent variable but not vice versa. In many real-world political science scenarios, this assumption is problematic. For example, there is debate in the literature whether high women's representation in instances of power influences/decreases corruption or whether low levels of corruption foster the election of women (see Esarey and Schwindt-Bayer 2017). There are statistical remedies such as instrumental regression techniques, which can model a feedback loop, that is, more advanced techniques can measure whether two variables influence themselves mutually. These techniques will also be covered in more advanced books and classes.
7. **No omitted variables:** We have an omitted variable problem if we do not include a variable in our regression model that theory tells us that we should include. Omitting a relevant or important variable from a model can have four negative consequences: (1) If the omitted variable is correlated with the included variables, then the parameters estimated in the model are biased, meaning that their expected values do not match their true values. (2) The error variance of the estimated parameters is biased. (3) The confidence intervals of included variables and more general the hypothesis testing procedures are unreliable, and (4) the R-squared of the estimated model is unreliable.



(Graph image published under the CC-BY-SA-3.0 license (http://creativecommons.org/licenses/by-sa/3.0/), via Wikimedia Commons)



(Graph image published under the CC-BY-SA-3.0 license (http://creativecommons.org/licenses/by-sa/3.0/), via Wikimedia Commons)

Fig. 9.2 Homoscedasticity and heteroscedasticity

8. **More cases than parameters ($N > k$):** Technically, a regression analysis only runs if we have more cases than parameters. In more general terms, the regression estimates become more reliable the more cases we have.
9. **No constant "variables":** For an independent variable to explain variation in a dependent variable, there must be variation in the independent variable. If there is no variation, then there is no reason to include the independent variable in a

regression model. The same applies to the dependent variable. If the dependent variable is constant or near constant, and does not vary with independent variables, then there is no reason to conduct any analysis in the first place.

10. **No perfect collinearity among regressors**: This rule means that the independent variables included in a regression should represent different concepts. To highlight, the more two variables are correlated, the more they will take explanatory power from each other (if they are perfectly collinear, a regression program such as Stata or SPSS cannot distinguish these variables from one another). This becomes problematic because relevant variables might become nonsignificant in a regression model, if they are too highly correlated with other relevant variables. More advanced books and classes will also tackle the problem of perfect collinearity and multicollinearity. For the purposes of an introductory course, it is enough if you have heard about multicollinearity.

Reference

Esarey, J., & Schwindt-Bayer, L. A. (2017). Women's representation, accountability and corruption in democracies. *British Journal of Political Science*, 1–32.

Further Reading

Since basically all books listed under bivariate correlation and regression analysis also cover multiple regression analysis, the books I present here go beyond the scope of this textbook here. These books could be interesting further reads, in particular to students, who want to learn more what is covered here.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis*. Boca Raton: Chapman and Hall/CRC. An overview of different approaches to analyze complex sample survey data. In addition to multiple linear regression analysis the topics covered include different types of maximum likelihood estimations such as logit, probit, and ordinal regression analysis, as well as survival or event history analysis.

Lewis-Beck, C., & Lewis-Beck, M. (2015). *Applied regression: An introduction* (Vol. 22). Thousand Oaks: Sage A comprehensive introduction into different types of regression techniques.

Pesaran, M. H. (2015). *Time series and panel data econometrics*. Oxford: Oxford University Press. Comprehensive introduction into different forms of time series models and panel data estimations.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Mason, OH: Nelson Education. Comprehensive book about various regression techniques; it is, however, mathematically relatively advanced.