Chapter 5
Linear Correlation and Regression

In Chap. 1, Pearson's correlation coefficient as a means to describe a linear association between two continuous measures was introduced. In this chapter, the inference of the correlation coefficient using sample data will be discussed first, and then the discussion will extend to a related method and its inference to examine a linear association of the continuous and binary outcomes with one or more variables using sample data.

5.1 Inference of a Single Pearson's Correlation Coefficient

A linear association measured by the Pearson's correlation coefficient between two continuous measures obtained from a sample, *r*, requires an inference. The two forms of inferences are hypothesis testing and interval estimation (i.e., construction of the confidence interval). Testing hypothesis is to state the null and alternative hypotheses, compute the test statistic, and determine if it is significant. Let us discuss the hypothesis testing first. The null hypothesis is that there is no linear association between two continuous outcomes (i.e., H_0: $\rho = 0$), and the alternative hypothesis is either a nondirectional alternative hypothesis (i.e., H_1: $\rho \neq 0$) or a directional alternative hypothesis (i.e., H_1: $\rho > 0$, or H_1: $\rho < 0$), depending on the researcher's objective. For such an inference we need a test statistic. A typical test statistic involves an arithmetic transformation of the sample correlation coefficient *r* because the sampling distribution of *r* is not approximately normal even when the sample size becomes large (i.e., the CLT is not applicable for the sample correlation coefficient). Nonetheless, it is noted that the sampling distribution of the transformation $z = ½[ln(1+r) - ln(1-r)]$ will follow $N(0, 1/\sqrt{(n-3)})$ under the null hypothesis as the number of observed data pairs, *n*, becomes sufficiently large. The idea of "*Observed Estimate ~ Null Value ~ SE* triplet (see Sect. 2.2.4.2)" is then applied to derive the test statistic. Instead of directly plugging in the observed sample correlation *r*, the above *z* transformation is substituted for the *observed estimate*,

H. Lee, *Foundations of Applied Statistical Methods*, DOI 10.1007/978-3-319-02402-8_5,

Table 5.1 Smallest absolute values of sample correlations that are significantly different from 0 by nondirectional *t*-test

	Level of significance of a one-sample nondirectional *t*-test (H_0: $\rho = 0$ versus H_1: $\rho \neq 0$)		
df = *n* of pairs–2	10 %	5 %	1 %
3	0.805	0.878	0.959
10	0.497	0.576	0.708
15	0.412	0.482	0.606
20	0.360	0.423	0.537
25	0.323	0.381	0.487
30	0.296	0.381	0.449

i.e., $z = \{0.5 \cdot [ln(1+r) - ln(1-r)] - 0\}/\sqrt{(n-3)}$, so that the sampling distribution of this resulting test statistic *z* can follow the standard normal distribution.

For small sample size, this test statistic can be resorted to *t*-distribution (i.e., one-sample *t*-test). The following table lists the minimum values of the sample correlation coefficients that would become statistically significant by a nondirectional *t*-test of which H_0: $\rho = 0$ and H_1: $\rho \neq 0$ for various sample sizes (Table 5.1).

The interval estimation can also be made by using this *z*-statistic. The lower and upper 95 % confidence limits of the population correlation coefficient can be obtained in two steps, of which the first step is to find the lower and upper 95 % confidence limits (i.e., 2.5th and 97.5th percentiles of the sampling distribution) of *z*, then equating these two limits to the expression $\{0.5\ [ln(1+\rho) - ln(1-\rho)] - 0\}/\sqrt{(n-3)}$, then finally solving them for ρ.

5.1.1 Q & A Discussion

Question: In correlation analyses, to what extent should we look at the *r*-value and the *p*-value? For instance, is $r = 0.7$ ($p < 0.05$), "stronger" than $r = 0.5$ ($p < 0.001$)? Is $r = 0.1$ a poor correlation even if $p < 0.001$? Is $r = 0.8$ a good correlation even if $p > 0.1$?

Answer: The magnitude of *r* and its *p*-value cannot be interpreted universally. The cross comparison of the magnitudes of *r*'s is only meaningful within one data set where all the *r*'s are obtained from the same sample size. Don't compare apples with oranges.

5.2 Linear Regression Model with One Independent Variable: Simple Regression Model

A statistical model usually appears as a mathematical description (often involves mathematical expression, i.e., equations, etc.) of how individual datum is determined with uncertainty (i.e., random sampling error). Linear regression model with one independent variable describes how a numeric (normally distributed) outcome (i.e., dependent) variable is determined by one independent nonrandom variable and a random error. More specifically, it appears as an equation where the left-hand side is the outcome variable and the right-hand side consists of two parts of which the first part articulates the nonrandom common rule and the second does the random error (i.e., individuals' deviations from the nonrandom common rule).

The following is a typical expression of the *i*th observed outcome y_i described by the linear regression model with one independent variable:

$$y_i = \underset{\substack{\uparrow \\ \text{Common rule}}}{\underline{\beta_0 + \beta_1 X_i +}} \qquad \underset{\substack{\uparrow \\ \text{Randomphenomenon}}}{\underline{\varepsilon_i}},$$

where ε_i, for individual *i*, is a random error term that follows a normal distribution with mean = 0 and variance = σ^2. The *i*th observed outcome y_i is expressed by the common value that is the same as the value for all other observations as long as the value of the independent variable is given to a certain value plus the random deviation from the common value. The regression refers to the rule, how this common value of the dependent variable is determined given a certain value of the independent variable. This model is called a *simple linear* regression model. It is called *simple* because there is only one independent variable and called *linear* because the common rule is expressed by a linear function of the independent variable. As discussed later in this chapter, a *multiple* (as opposed to simple) linear regression model is a linear model that includes more than one independent variable, e.g., $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon_i$.

The simple regression model can have its variants, and the following is such an example:

$$y_i = \beta_0 + \beta_1 X^2{}_i + \varepsilon_i,$$

where ε_i, for individual *i*, is a random error term that follows a normal distribution with mean = 0 and variance = σ^2. First, how many independent variables are there? Only one, so the *simple* part makes sense. Having x^2 in the model as the independent variable does not mean this is a nonlinear model. Let's note that the word *linear* means that the nonrandom common rule, $\beta_0 + \beta_1 x^2{}_i$, is linearly determined by a given value of the independent variable (i.e., the rate of linear change is β_1 for a unit change of x^2, and the amount $\beta_1 x^2{}_i$ determined by a particular value of x^2 is additive to β_0). To make it clearer, one can rename x^2 to a new name z, i.e., $y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$.

5.3 Simple Linear Regression Analysis

Regression analysis is to seek the best common rule equation that determines the mean value of the outcome variable given a certain value of the independent variable. The widely used computational procedure is the least squares method.

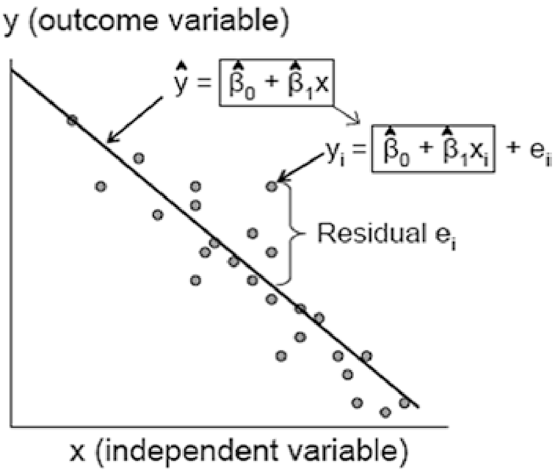
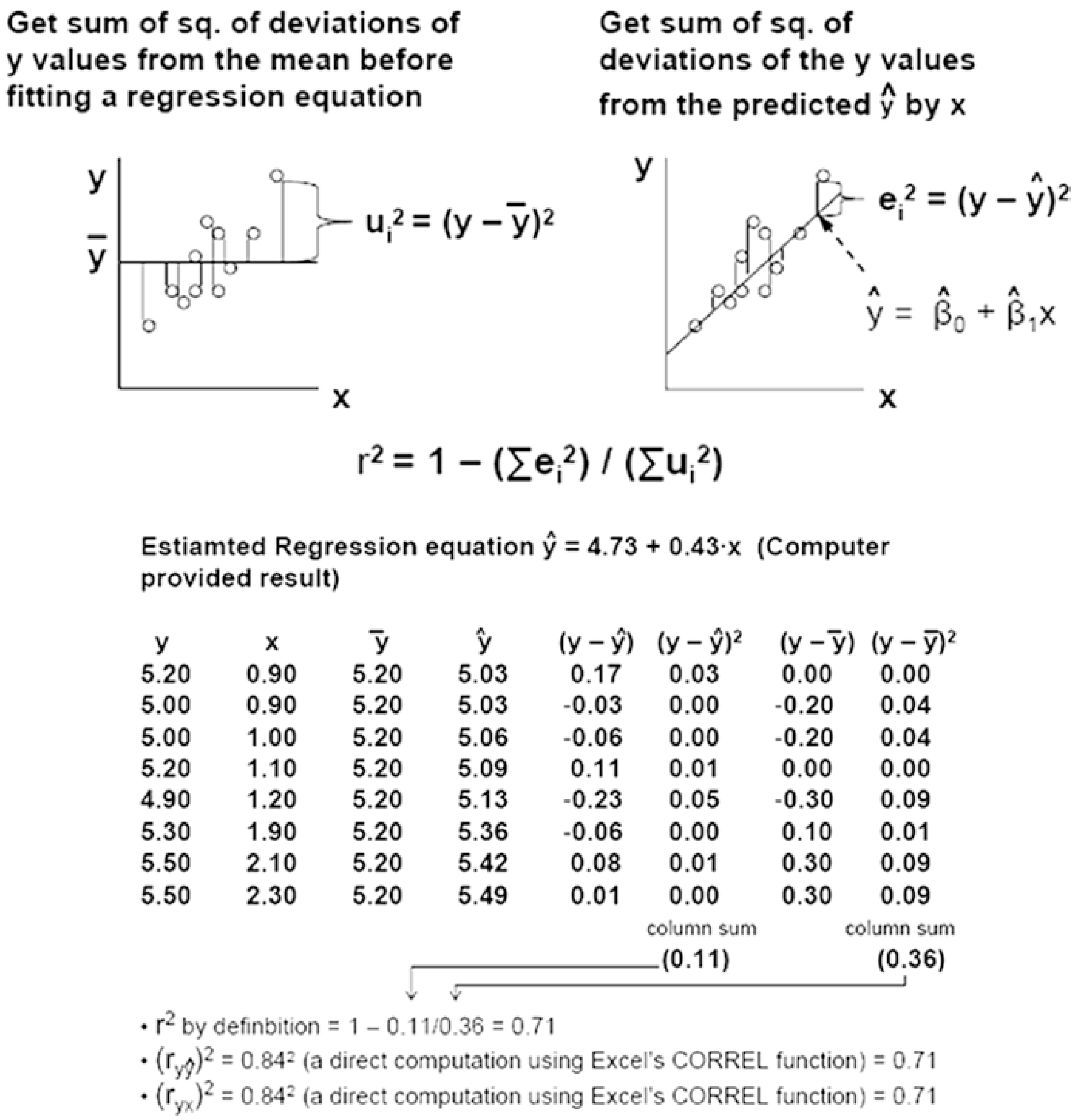

y (outcome variable)
$\hat{y} = \boxed{\hat{\beta}_0 + \hat{\beta}_1 x}$
$y_i = \boxed{\hat{\beta}_0 + \hat{\beta}_1 x_i} + e_i$
Residual e_i
x (independent variable)

Fig. 5.1 Illustration of the least squares method to estimate linear regression equation

In Fig. 5.1, the drawn line is the estimated regression line determined by the least squares method. Residual e_i is the difference between the observed y_i and the predicted value $\hat{y}_i$ via the estimated sample regression line. Note that the residual e_i is not the same random error term ε_i introduced in the model specification in that the term ε_i specified is the difference between the observed and the true population regression line. The least squares method is to estimate the intercept and slope of the regression line that minimize sum of squared residuals, e_i^2. The resulting estimated line is indeed the whole collection of the predicted means of the outcome variable y given the values of the independent variable x when the normality assumption of the ε_i error term's distribution is true. Computer programs (even Excel software has the feature) are widely available for estimating each regression equation parameter (i.e., intercept β_0, and slope β_1) and the standard error of each estimated regression parameter, and for providing the test statistic of the hypothesis testing whether or not each of the population coefficient is different from zero, as well as the 95 % confidence interval of each regression parameter.

A goodness of fit for the estimated simple linear regression equation is measured by r^2. This metric is the same as the squared value of the sample linear correlation coefficient computed from the observed y and x pairs. It is also the same as the proportion of the explained variation of the dependent variable by the estimated regression equation. The possible range is from 0 (0 % is explained) to 1 (100 % is explained). The r^2 is 1 – (sum of squares of the residuals/sum of squares deviations of the observed outcome values from the overall mean of the outcome values). Figure 5.2 illustrates the concept of r^2 and demonstrates the computational details. The first plot depicts the r^2 in the absence of a fitted regression equation for which the horizontal line represents the mean of y irrespective of the values of independent variable. The second plot depicts the r^2 of the fitted regression equation. It is also noted that r^2 is the squared value of the correlation coefficient between y and $\hat{y}$, and it is also the same as the squared value of the correlation coefficient between y and x. This can be shown algebraically and numerically.

Let's use an example of a simple linear regression equation estimated from an analysis, $y = 64.30 + 1.39 \cdot x$, where y denotes systolic blood pressure (SBP) and x



Get sum of sq. of deviations of y values from the mean before fitting a regression equation

Get sum of sq. of deviations of the y values from the predicted $\hat{y}$ by x

y
$\bar{y}$
x
$u_i^2 = (y - \bar{y})^2$
y
x
$e_i^2 = (y - \hat{y})^2$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$r^2 = 1 - (\sum e_i^2) / (\sum u_i^2)$$

Estiamted Regression equation $\hat{y} = 4.73 + 0.43 \cdot x$ (Computer provided result)

y	x	$\bar{y}$	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$	$(y - \bar{y})$	$(y - \bar{y})^2$
5.20	0.90	5.20	5.03	0.17	0.03	0.00	0.00
5.00	0.90	5.20	5.03	-0.03	0.00	-0.20	0.04
5.00	1.00	5.20	5.06	-0.06	0.00	-0.20	0.04
5.20	1.10	5.20	5.09	0.11	0.01	0.00	0.00
4.90	1.20	5.20	5.13	-0.23	0.05	-0.30	0.09
5.30	1.90	5.20	5.36	-0.06	0.00	0.10	0.01
5.50	2.10	5.20	5.42	0.08	0.01	0.30	0.09
5.50	2.30	5.20	5.49	0.01	0.00	0.30	0.09
					column sum (0.11)		column sum (0.36)

- r^2 by definbition = 1 – 0.11/0.36 = 0.71
- $(r_{y\hat{y}})^2 = 0.84^2$ (a direct computation using Excel's CORREL function) = 0.71
- $(r_{yx})^2 = 0.84^2$ (a direct computation using Excel's CORREL function) = 0.71

Fig. 5.2 Numerical illustration of r²

denotes age. The interpretation is that mean SBP increases linearly by 1.39 as a person's age increases by 1 year. For instance, a mean SBP of 30-year-old persons is predicted as 64.30 + 1.39·30 = 106. This value 106 is the common systematic rule to everyone whose age = 30. Note that this regression equation should be applied for a meaningful interval of the predictor variable *x* (e.g., age = 200 or age = −10 is nonsense). *y* = 64.30 when *x* = 0 is indeed the *y*-intercept and this may not be a value of interest (i.e., for age = 0).

It is important to know that what is being predicted by this linear regression equation is the mean value of the dependent variable given a particular value of the independent variable (*aka* conditional mean). In the above blood pressure prediction example, the predicted SBP value = 106 for a given age = 30 is indeed the estimated mean SBP of all subjects with age = 30. In Fig. 5.3, the estimated regression

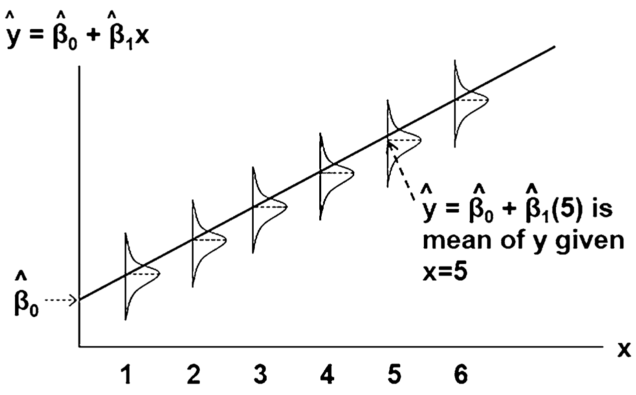

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(5)$ is mean of y given x=5
$\hat{\beta}_0$
x
1
2
3
4
5
6

Fig. 5.3 Illustration of regression mean

line represents the collection of predicted means (i.e., conditional means) of the dependent variable *y* given particular values of independent variable *x*.

All values on the predicted regression line are the means over the range of the given independent variable values, and a single point on that line is the estimated mean value given a particular value of the independent variable.

Many computer software programs offer to find the best (*unbiased minimum variance estimates)* regression coefficients of the specified model. Such programs also provide the estimated standard errors (SE) of the estimated regression coefficients for drawing inference. The hypothesis tests and interval estimations for the regression coefficients can be either directly available or easily completed by utilizing the computer-generated estimates.

The hypothesis test for the slope, β_1, is usually performed by a *z*- or *t*-test depending on the sample size. The practical choice of *z*-test is when the sample size is large enough (e.g., 30 or greater), otherwise a *t*-test is usually applied. The null hypothesis usually states that the regression slope is 0, i.e., H_0: $\beta_1 = 0$ (i.e., independent variable is not predictive of the outcome). The alternative hypothesis can either be nondirectional or directional depending on the research question, i.e., H_1: $\beta_1 \neq 0$ for a nondirectional test and H_1: $\beta_1 > 0$ for a directional test to claim a positive slope, etc. For both the *z*- and *t*-tests the test statistic is derived by the aforementioned "triplet," i.e., $\left[\hat{\beta}_1 - 0\right] / SE\left(\hat{\beta}_1\right)$ (see Sect. 2.2.4.5). The degrees of freedom for a *t*-test is $n-2$.

The interval estimation for each regression coefficient, i.e., the slope, can be constructed using *z*- or *t*-distribution depending on the sample size. For example, the 95 % confidence interval for the regression slope β_1 with a sample size of 20 is derived as $\left[\hat{\beta}_1 - 2.101 \times SE\left(\hat{\beta}_1\right), \hat{\beta}_1 + 2.101 \times SE\left(\hat{\beta}_1\right)\right]$, where 2.101 is the *t*-value, of which the tail area below −2.101 is 0.025 (i.e., 2.5th percentile) and the area above 2.101 is 0.025 (i.e., 97.5th percentile) with $df = 18$. The 95 % confidence interval using *z*-distribution when the sample size is large enough is derived as $\left[\hat{\beta}_1 - 1.96 \times SE\left(\hat{\beta}_1\right), \hat{\beta}_1 + 1.96 \times SE\left(\hat{\beta}_1\right)\right]$.

The confidence interval (band) for the entire regression mean response line (i.e., whole collection of individual regression means given the individual values of independent variable) can also be constructed. The algebraic expression becomes more complex than that of the slope because the interval estimation for the

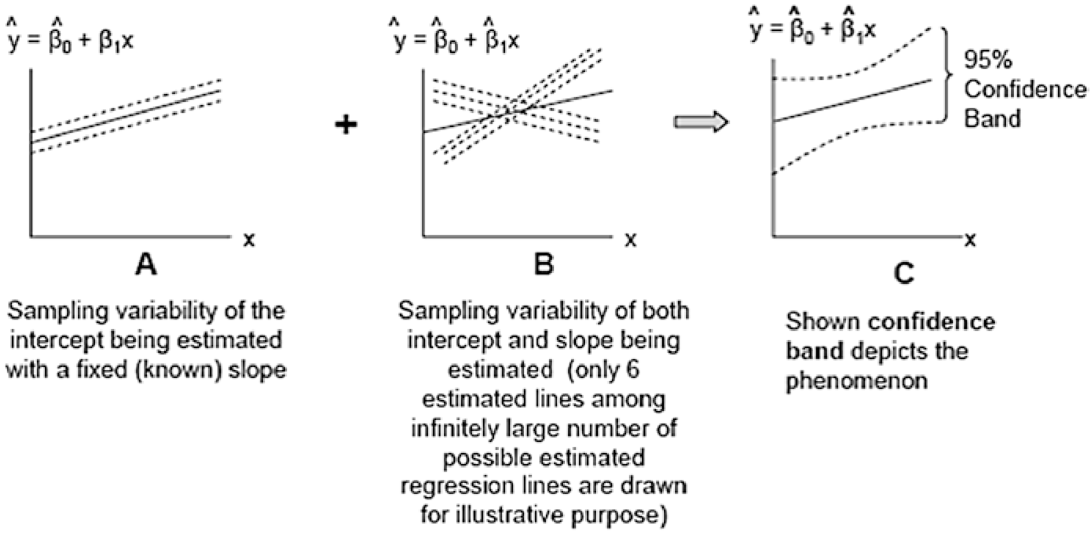

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
x
A
Sampling variability of the intercept being estimated with a fixed (known) slope
+
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
x
B
Sampling variability of both intercept and slope being estimated (only 6 estimated lines among infinitely large number of possible estimated regression lines are drawn for illustrative purpose)
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
95% Confidence Band
x
C
Shown confidence band depicts the phenomenon

Fig. 5.4 Illustration for aiding to understand confidence interval of the estimated linear regression equation

regression line involves issues of the underlying correlation between the intercept and slope estimates that are not independent with each other. The technical details may be beyond the level of knowledge of most of the readers. Letting alone the details, Fig. 5.4 demonstrates how the confidence band appears in that the bandwidth around the mean of independent variable is the narrowest and becomes wider as the values of the independent variable departs from its mean. Figure 5.4a demonstrates a special situation that only the intercept is estimated while the slope is not being estimated (assumed to be known and fixed during the estimation). Intuitively, the confidence band is parallel to the estimated regression line because the slope is always fixed to one value. Figure 5.4b demonstrates the variability of estimated regression line of which both the intercept and slope are being estimated (the shown lines are only several of infinitely large number of regression lines that are estimated and fluctuating due to the sampling variability of the raw data). Then, Fig. 5.4c illustrates the actual band of a regression line. Actual calculation of this is usually done by computer software.

Another interval estimation problem is to construct a confidence interval for predicted individual outcomes. When the regression equation is applied, the point estimate of an individual outcome value at a particular value of independent variable is indeed the estimated regression mean itself which is determined at that particular value of the independent variable. However, the confidence band of the predicted individual outcome values turn out to be a little bit wider than that of the regression line (i.e., the regression mean response line) because for a point on the regression line there are many individual values surrounded randomly above and below that single mean value on a particular point of the regression line. Such a band is called prediction band (e.g., 95 % prediction band), and its computational details take into account the additional random variability of these surrounded individual observations. Actual calculation of this is usually done by computer software.

Before we proceed to the next topic, a very important issue needs to be discussed. In many applications the data are observed at multiple time points within one subject and the observations are correlated within a subject (i.e., autocorrelation). Clinical studies may include a long-time series data of only a single subject (e.g., a long-time series of weekly incidence of an infectious disease in a particular place over many years, where the particular place can be viewed as a single study subject and the dependent variable is the number of new cases and the independent variable is the number of weeks since week 0) or multiple subjects with relatively short-time series data (monthly height growth pattern of a group infants over first 6 months after life, i.e., dependent variable is height and the independent variable is month after birth). The method of least squares estimation assumes that all data are uncorrelated (i.e., there is no autocorrelation). If this assumption is violated then the standard error of the regression coefficient estimate becomes inaccurate. Advanced techniques are available, but this material will not discuss. However, it is important to ensure that whether or not the study design (or data collection mechanism) would have induced such a problem and seek statistician's guidance to resolve the problem.

5.4 Linear Regression Models with Multiple Independent Variables

The outcome (dependent) variable of a regression models may need to be explained by more than one explanatory (independent) variable. For example, gray-haired people may show higher blood pressure than the rest, but the association between age and blood pressure is probably confounded with gray hair and age association and such a phenomenon needs to be taken into account. If multiple independent variables are additionally entered into the model, the model will decrease the residual variation of dependent variable that had not been explained solely by the primary independent variable of interest. Such a model with multiple independent variables is expressed as the following linear combination (i.e., a particular value of the dependent variable given a set of values of all independent variables in the model is expressed as a weighted sum of the independent variables where the regression coefficients β's being the weights).

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon_i,$$

where the assumption about ε_i is the same as what is specified in Sect. 5.2. The predicted value of the estimated regression equation for the *i*th individual, i.e., $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \ldots + \hat{\beta}_k \mathrm{x}_{ki}$, is the mean value of dependent variable *y* given the observed values of x_{1i}, x_{2i}, x_{3i}, …, and x_{ki}. The model fitting usually requires computer software. Below is a brief overview of how to perform such an analysis for model fitting (i.e., estimation of regression coefficients) and related inference.

The goodness of fit for a linear regression with multiple independent variables is measures by R^2 that is interpreted as the proportion of the explained variation of the

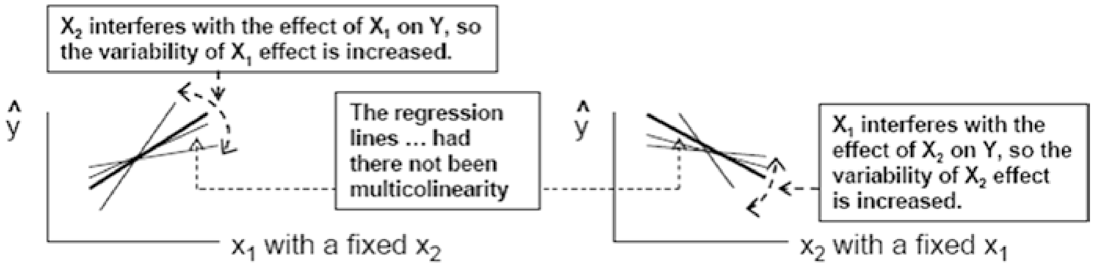

X₂ interferes with the effect of X₁ on Y, so the variability of X₁ effect is increased.
ŷ
The regression lines ... had there not been multicolinearity
x₁ with a fixed x₂
ŷ
X₁ interferes with the effect of X₂ on Y, so the variability of X₂ effect is increased.
x₂ with a fixed x₁

Fig. 5.5 Illustration of multi-colinearity in a multiple regression with two independent variables

dependent variable by the estimated regression equation. The least squares estimation seeks the regression coefficients that maximize and the least squares estimation seeks the regression coefficient estimates that maximize R^2. This R^2 is the squared value of the correlation coefficient between y and $\hat{y}$. In order to distinguish it from the case of simple linear regression's case (i.e., r^2), the notation uses capitalized R.

Multi-colinearity is a phenomenon due to a set of correlated independent variables in a multiple regression setting. It affects the estimated regression equation adversely. For an estimated multiple regression, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_1 + \hat{\beta}_2 \mathrm{x}_2 + \hat{\beta}_3 \mathrm{x}_3 + \ldots \hat{\beta}_\mathrm{k} \mathrm{x}_\mathrm{k}$, if two independent variables (for instance x_1 and x_2) are highly correlated, then the uncertainty about $\hat{\beta}_1$ and $\hat{\beta}_2$ increases and the standard errors of these two estimated regression coefficients are inflated. A high overall R^2 value (i.e., the independent variables, as a whole set, predict the mean outcomes pretty well) but the test results for some individual coefficients may not be significant (due to the inflated standard error of the regression coefficient estimate) and the interpretation of such regression coefficients in conjunction with other regression coefficient(s) becomes dubious (Fig. 5.5).

Exclusion of the independent variables that are highly correlated (i.e., redundant to certain variables) will prevent such an adverse consequence. A formal diagnosis can be made by using Tolerance, which is the proportion of unexplained variance of the independent variable being diagnosed by all other remaining independent variables (i.e., 1- R^2 of the estimated regression of the independent variable being diagnosed on all other variables). The inverse of Tolerance is called Variance Inflation Factor (VIF). A common criterion is to exclude the independent variable if the tolerance is less than 0.1 (or VIF greater than 10).

5.5 Logistic Regression Model with One Independent Variable: Simple Logistic Regression Model

Modeling a binary outcome variable by a regression is different from that of continuous outcome that was introduced in the previous sections. Let's discuss the following example.

Figure 5.6 illustrates a set of raw data of a set of binary outcome y (e.g., certain disease; illness if $y = 1$ and $y = 0$ if illness free) versus a continuous measure of x

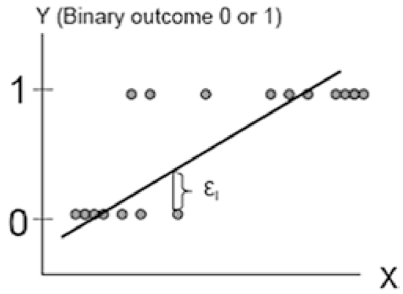


Fig. 5.6 illustration of inappropriate linear function to predict event probability of binary outcome given independent variable X

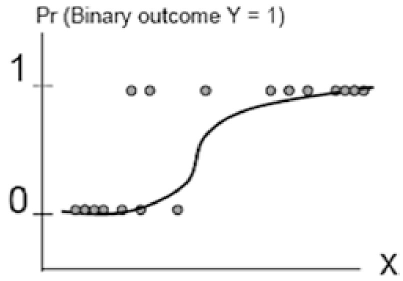


Fig. 5.7 Illustration of logistic function to predict event probability of binary outcome given independent variable X

(e.g., x = age in years) in the same way that was adopted to demonstrate the single independent variable linear regression model. The linear regression line stretches out above 1 and below 0, which is unrealistic. So, the idea of forcing the feasible range lies between 0 and 1, the logistic function is adopted and Fig. 5.7 illustrates this idea.

It is noted that the observations take values of either 0 or 1 but the regression curve does not exceed either 0 or 1, and it is also noted that the vertical axis is the probability of observing $y = 1$ given a particular value of x. This is called logistic regression model because the shape of the response curve is characterized by the cumulative distribution function of the logistic distribution (simply called logistic function). The mathematical expression of this function, where **e** is the base of natural logarithm, is

$$\text{Probability } \{y = 1 \text{ given } x\} = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}},$$

$$\text{and Probability } \{y = 0 \text{ given } x\} = 1 - \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}},$$

Unlike the linear regression model, the logistic regression model does not need the random error term because the transformed outcome variable of this logistic regression model specifies the probability of the event ($y = 1$) and this completely characterizes the probability distribution of the original outcomes of $y = 1$ and $y = 0$ (i.e., no other random error terns are necessary). A special emphasis is made here to the regression coefficient associated with the independent variable which measures the direction (positive or negative) and strength of association. Let's consider an example of which the outcome variable y is binary (1 = had an event, 0 = did not have an event) and independent variable x is a risk scale (1 = low risk, 2 = moderate risk, and 3 = high risk), the following probabilities of interests can be expressed via logistic equations:

Probability ($y = 1$ for moderate risk) = $[exp\ (\beta_0 + \beta_1 \cdot 2)]/[1 + exp\ (\beta_0 + \beta_1 \cdot 2)]$,
Probability ($y = 0$ for moderate risk) = $1 - [exp\ (\beta_0 + \beta_1 \cdot 2)]/[1 + exp\ (\beta_0 + \beta_1 \cdot 2)]$,
Probability ($y = 1$ for high risk) = $[exp\ (\beta_0 + \beta_1 \cdot 3)]/[1 + exp\ (\beta_0 + \beta_1 \cdot 3)]$, and
Probability ($y = 0$ for high risk) = $1 - [exp\ (\beta_0 + \beta_1 \cdot 3)]/[1 + exp\ (\beta_0 + \beta_1 \cdot 3)]$.

These probabilities are less of interest than the following odds ratio (OR see Sect. 1.4.3) in applied setting. If we are interested in the odds ratio of the event with high risk versus moderate risk then this odds ratio can be derived by a simple algebra as below.

$$
\begin{aligned}
OR &= \frac{\left[Probability\left(y = 1\, for\, high\, risk\right) / Probability\left(y = 0\, for\, high\, risk\right)\right]}{\left[Probability\left(y = 1\, for\, moderate\, risk\right) / Probability\left(y = 0\, for\, moderate\, risk\right)\right]} \\
&= \frac{\left[\exp\left(\beta_0 + \beta_1 \cdot 3\right)\right] / \left[1 + \exp\left(\beta_0 + \beta_1 \cdot 3\right)\right] / \left\{1 - \left[\exp\left(\beta_0 + \beta_1 \cdot 3\right)\right] / \left[1 + \exp\left(\beta_0 + \beta_1 \cdot 3\right)\right]\right\}}{\left[\exp\left(\beta_0 + \beta_1 \cdot 2\right)\right] / \left[1 + \exp\left(\beta_0 + \beta_1 \cdot 2\right)\right] / \left\{1 - \exp\left(\beta_0 + \beta_1 \cdot 2\right) / \left[1 + \exp\left(\beta_0 + \beta_1 \cdot 2\right)\right]\right\}}. \\
&= \exp\left(\beta_1 \cdot 3\right) - \exp(\beta_1 \cdot 2) = \exp(\beta_1).
\end{aligned}
$$

Likewise, the OR of moderate- versus low risk is $exp\ (\beta_1 \cdot 2)$ - exp $(\beta_1 \cdot 1)$, and the OR of high- versus low risk is exp $(\beta_1 \cdot 3) - exp\ (\beta_1 \cdot 1) = exp(\beta_1 \cdot 2) = 2 \cdot exp\ (\beta_1)$.

While the OR is the measure of association of our ultimate interest, its inference is made on the regression coefficient, β_1, because the OR is merely the transformed value of the regression coefficient (i.e., OR = $e^{\beta}{}_{1}$). The standard method for estimating the regression coefficients (i.e., fitting the logistic regression function) is the maximum likelihood (ML) method. This is a calculus approach to find the solution for the following likelihood function which is constructed by β_0 and β_1 and the observed data. The likelihood function, denoted by L, will be proportional to the joint probability of all observed events, i.e., the product of all probabilities of $y = 1$ given x for all observations with the outcome value 1 and all probabilities of $y = 0$ given x for all observations with the outcome value 0. The following is the spelled out expression of the illustrative observation set listed below.

Observation No.	Outcome y (0 or 1)	Predictor x (0 or 1)
1	1	1
2	1	0
3	1	1
.	.	.
.	.	.
.	.	.
.	.	.
n-1	0	1
n	0	o

$$L \approx \begin{array}{ccccccc} \left(\dfrac{e^{\beta_0+\beta_1(x=1)}}{1+e^{\beta_0+\beta_1(x=1)}}\right) & \times & \left(\dfrac{e^{\beta_0+\beta_1(x=0)}}{1+e^{\beta_0+\beta_1(x=0)}}\right) & \times\ \cdots\ \times & \left(1-\dfrac{e^{\beta_0+\beta_1(x=1)}}{1+e^{\beta_0+\beta_1(x=1)}}\right) & & \left(1-\dfrac{e^{\beta_0+\beta_1(x=0)}}{1+e^{\beta_0+\beta_1(x=0)}}\right) \\ \uparrow & & \uparrow & \cdots & \uparrow & & \uparrow \\ Observation1 & & Observation2 & & Observationn-1 & & Observationn \\ y=1, x=1 & & y=1, x=0 & & y=0, x=1 & & y=0, x=0 \end{array}$$

Each term in the above product is the logistic model-based probability of either $y=1$ or 0 given x. The maximum likelihood estimation procedure is a calculus problem to find the solutions for β_0 and β_1 that maximize this function. The actual computation uses its natural logarithm, *ln(L)*, instead of *L* by which the computation becomes much less burdensome. Letting alone the further detail of mathematical statistics aspect not addressed here, it is important and practically useful to note the property of the regression coefficients that are obtained from the ML method (*aka*, ML estimators). The property is that the sampling distribution of such an estimator follows Gaussian (i.e., normal) distribution as long as the sample size is sufficiently large. Relying on this property, similar to the simple linear regression case (see Sect. 5.2), a one-sample *z*-test (*aka* Wald's *z*-test) or *t*-test, if sample size is not large, is a common method for a regression coefficient β_1 to be tested for H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$. For interval estimation, the lower and upper limits of 95 % confidence interval for the regression coefficient (see Sect. 5.2) are obtained first, then these limits are transformed to OR limits, i.e., the limits are $e^{\textit{Lower limit of the regression coefficient}}$ and $e^{\textit{Upper limit of the regression coefficient}}$.

Because the maximum likelihood method does not resort to the least squares method, there is no goodness of fit such as the r^2 (for one independent variable) or R^2 (for multiple independent variables). Goodness of fit for an estimated logistic regression equation can be examined by several options. The most common option is to use Hosmer–Lameshow statistic, which measures the disagreement between observed versus expected events of interest in partitioned deciles (or three to nine if fewer than ten observed patterns of the independent variable(s) existed) of the predicted probabilities, and transform it to a Chi-square statistic with *g*-2 degrees of freedom where *g* is number of ordered partitions of the predicted probabilities (see Sect. 6.1).

5.6 Consolidation of Regression Models

5.6.1 General and Generalized Linear Models

Linear regression models that have more than one independent variable are called general linear models. If the regression models with more than one independent variable with its model equation is not linear (e.g., logistic) but is transformed into a linear form, then such transformed models are called generalized linear models. The meaning of "linear" is that the predicted mean value given the independent variables is expressed as a linear combination of the regression coefficients (i.e., simple addition of more than one term of which each individual term is the product of a regression coefficient and the corresponding independent variable) (see Sect. 5.2). For example, $a + bx$ is a linear combination of the two terms a and bx, and $c + dx^2$ is also a linear combination of c and dx^2. In the case of $c + dx^2$ the linearity is held between c and d. What is often confusing is that the resulting value of $c + dx^2$ turns out as a quadratic function with respect to x. However, by definition, such a regression equation is a linear model rather than a nonlinear model because the linearity between c and d is held as long as the observed x^2 value is viewed as the weight of the linear combination.

Unlike the linear models, nonlinear models are the ones that the model equation cannot be expressed by linear sum of the products created by the regression coefficients and their corresponding independent variables. For example, the logistic regression equation is a nonlinear function called logistic function (see Sect. 5.5). Nevertheless, the nonlinear function often can be converted to a linear function via algebra (i.e., linearization), and such transformed models are called generalized linear models. In the case of logistic regression, the logistic function to predict the probability of event can be transformed into a linear function to predict the log of the odds.

For the logistic regression equation Probability $\{y = 1$ given $x\}$ = $\dfrac{\mathrm{e}^{\beta_0+\beta_1 \mathrm{x}}}{1+\mathrm{e}^{\beta_0+\beta_1 \mathrm{x}}}$,

By letting *logit [p]* denote the transformation log_e [odds] = log_e [$p/(1\text{-}p)$] = log_e [Probability of $y = 1$ given x / (1- Probability of $y = 1$ given x)], the resulting equation becomes *Logit [p]* = $\beta_0 + \beta_1 x$, which is now a linear function to predict the *logit* (i.e., natural logarithm of the odds) while preserving the interpretation of both β_0 and β_1, the same as that were made in the original form, i.e., OR ($x = 1$ versus 0) = $e^{\beta}{}_1$. Such a linearization makes the computation of the estimation less burdensome. The computational detail is beyond the objective of this monograph and is not described.

5.6.2 Multivariate Analyses and Multivariate Model

The terminologies *Multivariate Analyses* and *Multivariate Model* are very often misused by the applied researchers, and such errors appear frequently even in published articles.

A *Multivariate Analysis* is the simultaneous analysis of two or more related numeric outcome variables (i.e., dependent variables). Such methods are commonly applied in the social science research, and some popular methods are T^2-test for simultaneous comparison of two or more related means between two groups (e.g., comparison of mean weight and mean height between men and women), Multivariate Analysis of Variance (MANOVA) for simultaneous comparison of two or more related means among three or more groups (e.g., comparison of mean weight and mean height among three ethnic groups), Multivariate Regression Analysis to fit more than one correlated dependent variables by means of more than one related regression equations, Factor Analysis and Principal Component analysis to reduce a large dimension of linearly correlated variables into a small dimension, Canonical Correlation Analysis to examine a set of correlated variables with another set of correlated variables, and Linear Discriminant Analysis to build a linear equation by a set of linearly correlated random variables to differentiate the individuals into two or more groups, etc.

A *Multivariate Model* refers exclusively to a regression model of a single outcome variable with two or more independent variables (e.g., multiple linear regression models, ANCOVA models, etc.), and the analysis method is univariate because there is only one dependent variable. Note that the multiplicity of the independent variables in a model does not mean that the method is multivariate.

5.7 Application of Linear Models with Multiple Independent Variables

Figures 5.8 and 5.9 demonstrate a particular type of applications of general linear models to predict the mean of dependent variable using multiple independent variables. In the first case, the predicted mean given independent variables

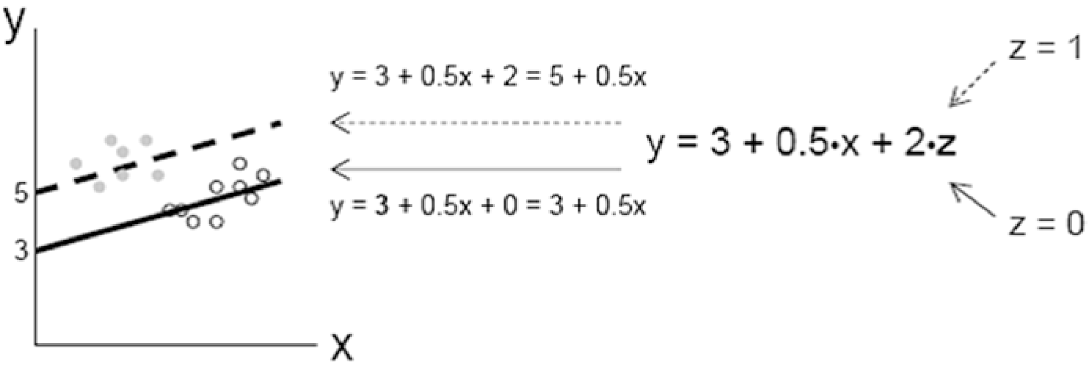


Fig. 5.8 illustration of dummy variable technique without modeling an effect of interaction

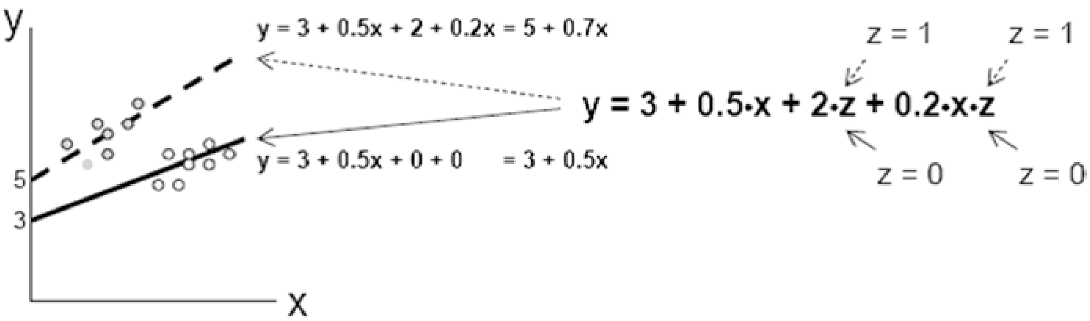

y
5
3
x
y = 3 + 0.5x + 2 + 0.2x = 5 + 0.7x
y = 3 + 0.5x + 0 + 0 = 3 + 0.5x
y = 3 + 0.5·x + 2·z + 0.2·x·z
z = 1
z = 1
z = 0
z = 0

Fig. 5.9 Illustration of dummy variable technique applied to model a main effect and an effect of interaction

(i.e., regression equation) is determined by two independent variables, of which the first is a continuous variable x and the second, z, is to take either 1 or 0. Such a dichotomized independent variable to take either 0 or 1 is called dummy variable.

In Fig. 5.8 the dummy variable was used to fit the two regression lines with the same slopes but different intercepts.

In Fig. 5.9, the dummy variable was used to fit the two regression lines with two different slopes and intercepts. The last term of the regression equation is $0.2 \cdot x \cdot z$, of which the variable that takes data values is the product of x and z. Such a term is called interaction term. The corresponding regression coefficient is the size of the difference in slopes between the two subgroups of having $z = 1$ and having $z = 0$. Note that the product term variable, $x \cdot z$, is considered as a single variable (e.g., it can be renamed as any one letter variable name such as "w," etc.).

5.8 Worked Examples of General and Generalized Linear Modes

5.8.1 Worked Example of a General Linear Model

Four hundred ($n = 400$) over-weighted adults with age between 35 and 45 years participated in a 1:1 randomized 1-year study of a weight loss intervention program (i.e., 200 on the invention arm and 200 on control arm). The study collected the baseline weight and the weight change after the completion of the study.

The baseline mean (± standard deviation) weight (in lb) among all participants was 201.6 (±32.9) and their mean values of the 1-year weight changes were −3.74 (±9.34) and 4.82 (±7.12) in the intervention and control group, respectively. A general linear model analysis was applied to determine the intervention effects on the mean weight change without and with adjusting for the individual participant's age (Table 5.2 and Fig. 5.10). The dependent variable was the 1-year weight change (WC: post 1 year weight – baseline weight), and the independent variables were intervention (I: 1 = yes, 0 = no) and age (AGE: continuous. Note that intervention (I) is a dummy variable.

Table 5.2 Summary of general linear model analysis: weight loss intervention study

Model	Independent variables	$\hat{\beta}$	SE ($\hat{\beta}$)	p-Value
Model 1	Intercept	4.82	0.59	<0.0001
	Intervention (I)	− 8.56	0.83	<0.0001
Model 2, $R^2 = 0.76$	Intercept	−78.55	4.77	<0.0001
	Intervention (I)	−49.32	6.92	<0.0001
	Age (AGE)	2.09	0.12	<0.0001
	Interaction of intervention and age (I × AGE)	0.98	0.17	<0.0001

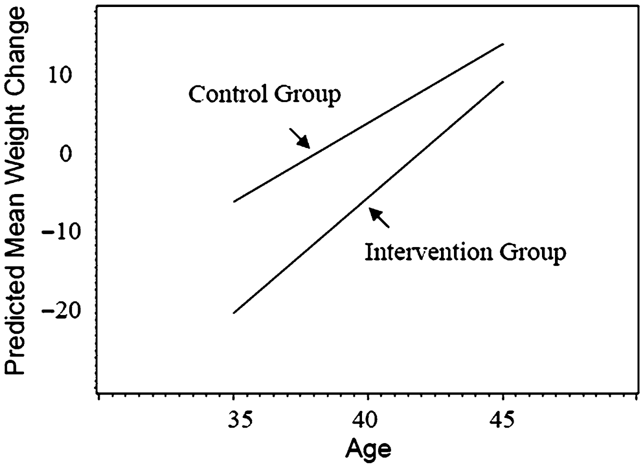


Fig. 5.10 Illustration of effect of interaction between intervention and age

In Model 1, the estimated intercept value of 4.82 lb is the mean weight change among the control participants, and the estimated parameter value of the intervention variable (I), -8.56 ($p<0.0001$), directly offers the significant estimated difference (i.e., effect) in the mean weight changes between the two groups. With these two regression coefficients, the mean change in the intervention group can be estimated by 4.82 – 8.56 = −3.74, which is the same as the group-specific descriptive summary statistics presented above (before performing the general linear model analysis). Model 2 was constructed in order to predict the mean weight change not only by the given intervention status but also by the age. The main effect of age as well as its interaction with the intervention (i.e., whether or not the age effects were different between the intervention and control subjects) were added to this model. Note that the estimated parameter value of the intervention (I) does not directly offer the difference in the mean weight changes between the intervention and control groups because the additional variables are included now and those effects must be taken into account simultaneously. The estimated parameter value of −49.32 ($p<0.0001$) is the group difference of the mean weight changes only for the persons with age 0. The age of 0 is unrealistic. So, if we chose a particular age of 40 for a meaningful interpretation, then the intervention group's mean weight change is predicted by $-78.55 - 49.32\times1+2.09\times40+0.98\times1\times40=-5.07$, and that of the control group is $-78.55 - 49.32\times0+2.09\times40+0.98\times0\times40=5.05$, thus the estimated effect (i.e., the mean difference) at age 40 is $-5.07-(-5.05)=-10.12$, which is the conditional effect of the intervention for 40-year-old participants. As shown in Fig. 5.10, the conditional effect decreased as the age increased.

5.8.2 Worked Example of a Generalized Linear Model (Logistic Model) Where All Multiple Independent Variables Are Dummy Variables

A large survey study investigated if the college students in California are less involved in binge drinking (Wechsler et al. 1997). The survey sample comprised

Table 5.3 Summary of generalized linear model analysis: California college students binge drinking study

Model	Independent variables	$\hat{\beta}$	SE ($\hat{\beta}$)	$\widehat{OR} = \exp(\hat{\beta})$	p-Value
Model 1	California	−0.66	0.053	0.52	<0.0001
Model 2	California	0.18	0.19	1.20	0.353
	Age < 24	0.81	0.05	2.24	<0.0001
	Male	0.44	0.03	1.56	<0.0001
	Never married	1.27	0.06	3.58	<0.0001
	White	1.08	0.05	2.95	<0.0001
	Non-commuter	0.68	0.04	1.97	<0.0001
	Smoker	1.54	0.04	4.38	<0.0001

1864 college students from California and 17,592 from elsewhere in the USA. The logistic regression analysis was performed as below.

Dependent variable – Binge drinking (1 vs. 0).

Independent variables – California student (1 vs. 0); Age < 24 (1 vs. 0); Male gender (1 vs. 0); Never married (1 vs. 0); White ethnicity (1 vs. 0); Non-commuter (1 vs. 0); Smoker (1 vs. 0) (Table 5.3).

Unlike the result summary of the general linear model (Table 5.2), the result summary of this generalized linear model analysis did not show the estimated parameter values of the intercepts (Table 5.3) because the intercept is the nuisance parameter for the odds ratio (see Sect. 5.5). The simple logistic regression model of binge drinking solely on the California residency indicator variable (1 = live in California, 0 = elsewhere) showed that there was significant decrease in binge drinking among the California college students ($\widehat{OR}$ = 0.52, p < 0.0001). However, after simultaneously adjusting for other demographic variables and other risk factors (every variable was dichotomized as 1 = yes and 0 = no), this effect was no longer significant (Adjusted $\widehat{OR}$ = 1.20, not significantly different from 1 at a 5 % significance level) while all the other covariates were significantly associated with the binge drinking in that students under 24 years old (Adjusted $\widehat{OR}$ =2.24, p < 0.0001), male students (Adjusted $\widehat{OR}$ =1.56, p < 0.0001), never married students (Adjusted $\widehat{OR}$ =3.58, p < 0.0001), students with white ethnic background (Adjusted $\widehat{OR}$ =2.95, p < 0.0001), non-commuter students (Adjusted $\widehat{OR}$ =1.97, p < 0.0001), and smoker students (Adjusted $\widehat{OR}$ =4.38, p < 0.0001) were involved more in binge drinking.

5.9 Study Questions

1. The estimated least square linear regression equation (simple or multiple regression) does not predict an individual's specific outcome value given the subject's value(s) of the independent variable(s)? What value does the regression equation predict?
2. What is the quantitative interpretation of the regression coefficient (i.e., the slope) of a least square linear regression equation?

3. What value does a logistic regression equation predict given an individual's value(s) of the independent variable(s)?
4. What is the quantitative interpretation of the regression coefficient of a logistic regression equation?
5. What are the definitions of the following?

   Odds ratio
   General linear model
   Generalized linear model

6. Explain why the multiple linear regression and multiple logistic regression are not multivariate analyses.

Bibliography

Breslow NE, Day NE (1980) Statistical methods in cancer research, vol 1, the analysis of case–control studies (IARC scientific publications no. 32). IARC, Lyon

Cox DR (1970) Analysis of binary data. Chapman and Hall, New York

Draper NR, Smith H (1998) Applied regression analysis, 3rd edn. Wiley, NJ

Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression, 3rd edn. Wiley, Hoboken, NJ

Johnston J (1998) Econometric methods, 4th edn. McGraw-Hill, Boston, MA

MacCullagh P, Nelder JA (1989) Generalized linear models. Chapman & Hall, Boca Raton, FL

Pagano M, Gauvreau K (1993) Principles of biostatistics. Duxbury, Belmont, CA

Rosner B (2010) Fundamentals of biostatistics, 7th edn. Cengage Learnings, Inc. Boston, MA

Wechsler H, Fulop M, Padilla A, Lee H, Patrick K (1997) Binge drinking among college students: a comparison of California with other states. J Am Coll Health 45(6):273–277