

15: Regression

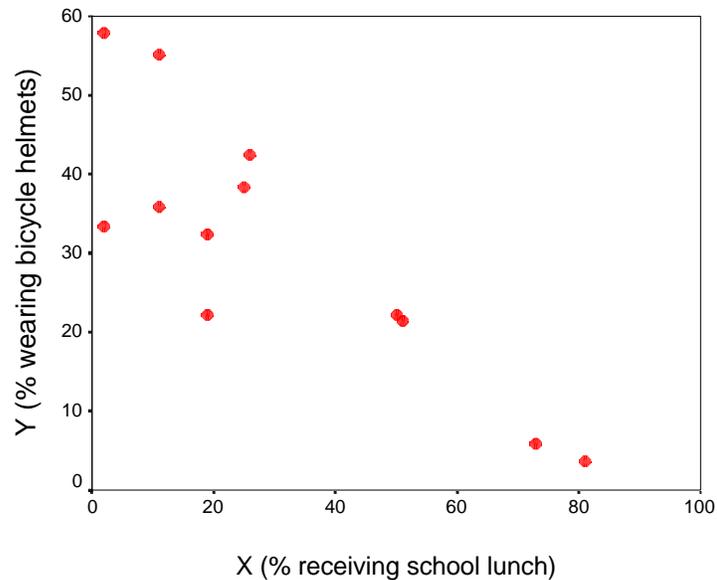
[Introduction](#) | [Regression Model](#) | [Inference About the Slope](#)

Introduction

As with correlation, regression is used to analyze the relation between two continuous (scale) variables. However, regression is better suited for studying **functional dependencies** between factors. The term *functional dependency* implies that X [partially] *determines* the level of Y. For example, there is a function dependency between age and blood pressure since as one ages, blood pressure increases. In contrast, there is *no* functional dependency between arm length and leg length since increasing the length of an arm will have no effect on leg length (or vice versa).

In addition, regression is better suited than correlation for studying samples in which the investigator **fixes the distribution of X**. For example, if I decide to select ten 30-year-olds, ten 40-year olds, and ten 50-year-olds to study the relation between age and blood pressure, I have fixed the distribution of the X variable in the sample. This would necessitate the use of regression and (in theory) prevent the use of correlation.

Illustrative data. We use the same data presented in the previous chapter (`bicycle.sav`) to illustrate regression techniques (Fig. below). Recall that the independent variable (X) in this data set represents the percent of children in the neighborhood receiving reduced-fee school lunches (a surrogate for neighborhood socioeconomic status). The dependent variables (Y) represents the percent of bicycle riders wearing helmets. (This study was done before bicycle helmet use laws were enacted.) Data are listed in the previous chapter. There is a strong negative correlation between X and Y ($r = -0.85$).



Regression Model

You might remember from algebra that a line is identified by its **slope** (the angle of the line describing the change in Y per unit X) and **intercept** (where the line crosses the Y axis). Regression describes the relation between X and Y with just such a line. When discussing our line, let

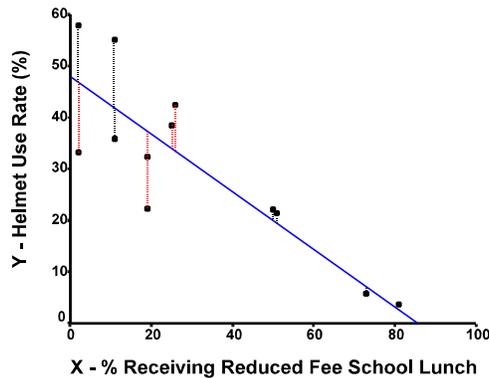
\hat{y} represent the **predicted value of Y**,
 a represent the **intercept** of the best fitting line, and
 b represent the **slope** of the line.

Thus, the **regression model** is denoted:

$$\hat{y} = a + bx \quad (1)$$

But how do we identify the best line for the data? If all data points were to fall on such a line, identifying the slope and intercept would be easy. However, because statistical data has random scatter, identifying a good line is not a trivial matter.

The random scatter around the line is identified as the distance of each point from the predicted line. These distances, called **residuals**, are shown as dotted lines in the figure below:



The goal is to determine a line that minimizes the *sum of the squared residuals*. This line is called the **least squares line**. The **slope**(b) of the least squares line is given by:

$$b = \frac{SS_{XY}}{SS_{XX}}$$

where SS_{XY} is the sum of the cross-products and SS_{XX} is the sum of the squares for variable X. (See previous chapter for SS formulas.) For the illustrative data, $SS_{XY} = -4231.1333$ and $SS_{XX} = 7855.67$. Therefore, $b = -4231.1333 / 7855.67 = -0.539$.

The **intercept** of the least squares line is given by the equation:

$$a = \bar{y} - b\bar{x} \quad (3)$$

where \bar{y} is the average value of Y, b is the slope, and \bar{x} is the average value of X. For the illustrative data, $\bar{y} = 30.8833$, $b = -0.54$, and $\bar{x} = 30.8333$. Therefore $a = (30.8833) + (-0.539)(30.8333) = 47.49$ and the regression model is: $\hat{y} = 47.49 + (-0.54)x$.

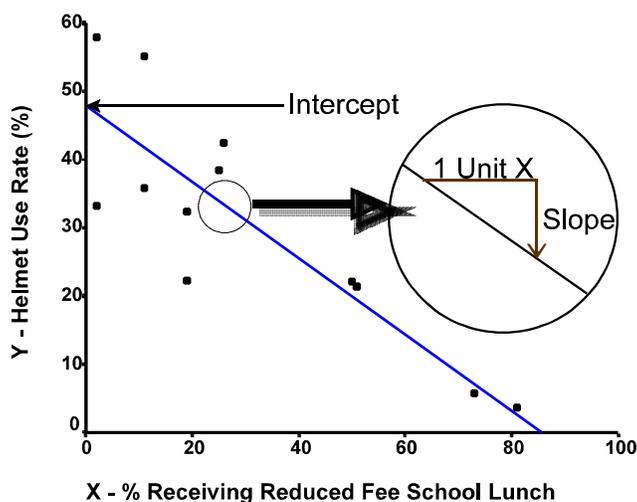
SPSS. Regression coefficients are requested in SPSS by clicking ANALYZE > REGRESSION > LINEAR. Output for the illustrative data includes the following table:

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	47.490	4.242		11.194	.000
	X (% children receiving school lunch)	-.539	.106	-.849	-5.087	.000

a. Dependent Variable: Y (% bicycle riders wearing helmets)

The column labeled *Unstandardized Coefficients* contains the coefficients we seek. The intercept (a) is reported as the unstandardized coefficient for the (Constant). The slope (b) is reported as the coefficient for the X variable.

Interpretation of the slope estimate. The slope a regression model represents *the average change in Y per unit X*:



The slope of -0.54 predicts 0.54 fewer helmet users (per 100 bicycle riders) for each additional percentage of children receiving reduced-fee meals. Notice that in order to interpret the regression coefficient, you must keep track of the units of measurement for each variable. For example, if helmet use was expressed per 1000 riders (instead of per 100), the regression coefficient would be increased by a corresponding factor of ten up to 5.4 fewer helmet uses per 1000 riders for each percentage increase in reduced-fee school lunches. This conspicuous point is worth keeping in mind.

Using Regression Coefficients for Prediction. The regression model can be used to predict the value of Y at a given level of X. For example, a neighborhood in which half the children receive reduced-fee lunch ($X = 50$) has an expected helmet use rate (per 100 riders) that is equal to $47.49 + (-0.54)(50) = 20.5$.

Inference About the Slope

The slope in the sample is not the same as the slope in the population. Thus, different symbols are needed to refer to each. Let b represent the calculated slope in the sample and let β represent the slope in the population.

It is possible to find a positive slope in the sample (i.e., $b > 0$) when in fact there is a negative slope in the population ($\beta < 0$), and vice versa. We use the standard techniques of estimation and hypothesis testing to infer the value of β .

Confidence Interval for the Slope

The **standard error of the regression**, denoted $se_{Y|x}$, is:

$$se_{Y|x} = \sqrt{\frac{SS_{YY} - b(SS_{XY})}{n - 2}} \quad (4)$$

where SS_{YY} is the sum of squares of Y, b is the slope estimate, SS_{XY} is the cross-product of X and Y, and n is the sample size. This statistic quantifies the standard deviation of Y after taking into account its dependency on X and is a rough measure of the predicted error.

$$\text{For the illustrative data, } se_{Y|x} = \sqrt{\frac{3159.68 - (-0.54)(-4231.133)}{12 - 2}} = 9.38.$$

The above statistic allows us to calculate the **standard error of the slope**, which is:

$$se_b = \frac{se_{Y|x}}{\sqrt{SS_{XX}}} \quad (5)$$

$$\text{For the illustrative data, } se_b = \frac{9.38}{\sqrt{7855.67}} = 0.1058.$$

The random distribution of the slope is assumed to be normal with a mean of 0 and standard deviation of se_b . Thus, a **95% confidence interval for b** is:

$$b \pm (t_{n-2, .975})(se_b) \quad (6)$$

where $t_{n-2, .975}$ is the 97.5th percentile on a t distribution with $n-2$ degrees of freedom. For the illustrative data, the 95% confidence interval for $\beta = -0.54 \pm (t_{10, .975})(0.1058) = -0.54 \pm (2.23)(0.1058) = -0.54 \pm 0.24 = (-0.30, -0.78)$. We can say with 95% confidence that population slope β lies between -0.30 and -0.78 .

Hypothesis Test

The test of $H_0: \beta = 0$ can be performed with a t statistic:

$$t_{\text{stat}} = \frac{b - 0}{se_b} \quad (7)$$

This statistic $n - 2$ degrees of freedom.

Illustrative example. The illustrative data shows $t_{\text{stat}} = (-0.54 - 0) / 0.1058 = -5.10$ with $df = 12 - 2 = 10$, $p = .00046$, providing strong evidence against H_0 .

Assumptions. The confidence interval and hypothesis test assume:

- Linearity(between X and Y)
- Independence (of bivariate observations)
- Normality (of the residuals)
- Equal variance (of residuals at each point on the line)

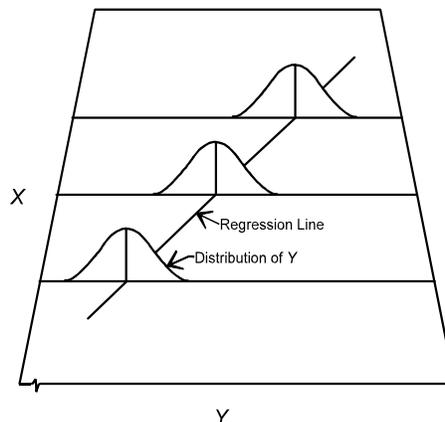
The first letters of these assumptions form the handy mnemonic “LINE”.

Briefly, **linearity** implies the relation between X and Y can be described by a straight line. The most direct way to assess linearity is with a scatter plot. When the relation between X and Y is not linear, regression should be avoided. Alternatively, data may be algebraically transformed to straightened-out the relation or, if linearity exists in part of the data but not in all, we can limit descriptions to that portion which is linear. (Illustrations of data transformations and range restrictions are provided in lab.)

The **independence** assumption provides for random sampling of bivariate observations.

The **normality** and **equal variance** assumptions address distribution of residuals around the regression model’s line. Random scatter should be normal with a mean of zero and consistent variance. This is shown graphically in the figure below.

With this said, regression models are robust allowing for departure from model assumptions while still providing meaningful results.



Comments on Interpretation

It is clear from the outset that statistical associations are not always causal. The distinction between causal and noncausal associations in health and disease has several explanations, each of which must be examined individually.

Working with aggregate-level (“ecological”) data can be especially problematic. A common type of aggregate-level analysis compares morbidity or mortality rates according to geographic region. In this type of analysis, the independent variable (“exposure”) is a summary characteristic of the region and the dependent variable (“disease”) is the rate or prevalence of disease in that region. This applies to the primary illustrative data set used in this and the prior chapter (`bicycle.sav`), as well as exercise 14.3/15.3 concerning the correlation between per capita cigarette consumption and lung cancer mortality.

One problem in interpreting ecological data is referred to as the **ecological fallacy**. This problem consists in thinking the association seen in the aggregate holds for individuals when in fact it does not (Thorndike, 1939; Selvin, 1958). This problem can often be viewed as one of confounding due to unmeasured co-factors.

Illustrative example: `farr1852.sav`. (Farr’s *Studies of Elevation Above Sea Level and Cholera Mortality*; Gerstman, 2003, p. 196). Farr’s 1852 ecological study of altitude above sea level and cholera mortality provides a vivid illustration of the ecological fallacy. At the time of the study, Farr was a confirmed *miasmaist*, placing emphasis on general social and environmental conditions as contributors to epidemics while giving little credibility to what we would now call contagion (Eyler, 1980). Farr supported his theory of miasma with ecological data on cholera mortality in which districts were classified according to the “Mean Elevation of the Ground above the Highwater Mark” (`elevati`) and observed average mortality from cholera per 10,000 (`obs`). Additional information is available in *Epi Kept Simple* pp. 196-197. Data are:

<code>elevati</code>	<code>obs</code>
0	177
10	102
30	65
50	34
70	27
90	22
100	17
350	7

A plot of the data is to the right. If we exclude the outlier in the lower right quadrant, $r = -.88$ and the regression model is $\hat{y} = 129.9 - 1.33X$ ($p = .009$).

We now recognize that the elevation variable was merely a surrogate for the likelihood of exposure to raw sewage and was thus confounded by the unmeasured contaminated water supply variable.

