

# Simple Linear Regression (Soc 504)

The linear regression model is one of the most important and widely-used models in statistics. Most of the statistical methods that are common in sociology (and other disciplines) today are extensions of this basic model. The model is used to summarize large quantities of data and to make inference about relationships between variables in a population. Before we discuss this model in depth, we should consider why we need statistics in the first place, in order to lay the foundation for understanding the importance of regression.

## 1 Why Statistics?

There are three goals of statistics:

1. Data summarization
2. Making inference about populations from which we have a sample
3. Making predictions about future observations

### 1.1 Data Summarization

You have already become familiar with this notion by creating univariate “summary statistics” about sample distributions, like measures of central tendency (mean, median, mode) and dispersion (variance, ranges). Linear regression is also primarily about summarizing data, but regression is a way to summarize information about relationships between 2 or more variables. Assume for instance, we have the following information:

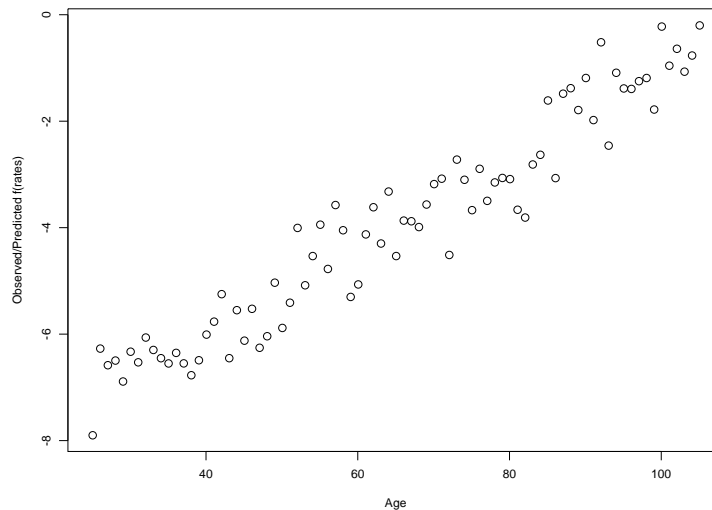


Figure 1. F(death rates) by Age

This plot is of a function of 1995 death rates for the total US population ages 25-105, by age (see <http://www.demog.berkeley.edu/wilmoth/mortality>). [Side Note: Death rates are approximately exponential across age, so I've taken the  $\ln()$  to linearize the relationship. Also, I've added a random normal variable to them  $[N(0, .5)]$ , because at the population level, there is no sampling error, something we will discuss in the next lecture].

There is clearly a linear pattern between age and these 81 death rates. One goal of a statistical model is to summarize a lot of information with a few parameters. In this example, since there is a linear pattern, we could summarize this data with 3 parameters in a linear model:

$$Y_i = \beta_0 + \beta_1 \text{Age}_i + \epsilon_i \quad (\textit{Population Equation})$$

Or

$$Y_i = b_0 + b_1 \text{Age}_i + e_i \quad (\textit{Sample Equation})$$

Or

$$\hat{Y}_i = b_0 + b_1 \text{Age}_i \quad (\textit{Sample Model})$$

We will discuss the meaning of each of these equations soon, but for now, note that, if this model 'fits' the data well, we will have reduced 81 pieces of information to 3 pieces of information (an intercept- $b_0$ , a slope- $b_1$ , and an error term-more specifically, an error variance,  $s_e^2$ ).

## 1.2 Inference

Typically we have a sample, but we would like to infer things about the population from which the sample was drawn. This is the same goal that you had when you conducted various statistical tests in previous courses. Inference in linear regression isn't much different. Our goal is to let the estimates  $b_0$  and  $b_1$  be our 'best guess' about the population parameters  $\beta_0$  and  $\beta_1$ , which represent the relationship between two (or more) variables at the population level. Formalizing inference in this model will be the topic of the next lecture.

## 1.3 Prediction

Sometimes, a goal of statistical modeling is to predict future observations from observed data. For example, given the data above, we might extrapolate our line out from age 105 to predict what the death rate should be for age 106, or 110. Or we might extrapolate our line back from age 25 to predict the death rate for persons at age 15. Alternatively, suppose I had a time series of death rates from 1950-2001, and I wanted to project death rates for 2002. Or, finally, suppose I had a model that predicted the death rates of smokers versus nonsmokers, and I wanted to predict whether a person would die within the next 10 years based on whether s/he were a smoker versus a nonsmoker.

## 1.4 Prediction and Problems with Causal Thinking

Inference and Prediction are not very different, but prediction tends to imply causal arguments. We often use causal arguments in interpreting regression coefficients, but we need to

realize that statistical models, no matter how complicated, may never be able to demonstrate causality. There are three rules of causality:

1. Temporality. Cause must precede effect.
2. Correlation. Two variables must be correlated (in some way) if one causes the other.
3. Nonspuriousness. The relationship between two variables can't be attributable to the influence of a third variable.

### 1.4.1 Temporality

Many, if not most, social science data sets are cross-sectional. This makes it impossible to determine whether  $A$  causes  $B$  or vice versa. Here is where theory (and some common sense) comes in. Theory may tell us that  $A$  is causally prior to  $B$ . For example, social psychological theory suggests that stress induces depression, and not that depression leads to stress. (note that two theories, however, may posit opposite causal directions). Common sense may also reveal the direction of the relationship. For example, in the mortality rate example, it is unreasonable to assume that death rates make people older.

### 1.4.2 Correlation and Nonspuriousness

Two variables must be related ( $B$ ) if there is a causal relationship between them ( $A$ ), but this does not imply the reverse statement that correlation demonstrates causation. Why? Because there could be any number of alternate explanations for the relationship between two variables. There are several types of alternate explanations. First, a variable  $C$  could be correlated with  $A$  but be the true cause of  $B$ . In that case, the relationship between  $A$  and  $B$  is “spurious.” A classic example of this is that ice cream consumption rates (in gallons per capita) are related to rape rates (in rapes per 100,000 persons). This is not a causal relationship, however, because both are ultimately driven by “season.” More rapes occur, and more ice cream is consumed, during the summer. Regression modeling can help us rule out such spurious relationships.

Second,  $A$  could affect  $C$ , with  $C$  affecting  $B$ . In that case,  $A$  *may* be considered a cause, but  $C$  is the more proximate cause (often called an “intervening variable”). For example, years of education is strongly linearly related to health. However, we seriously doubt that time spent in school is ultimately the cause of health; instead, education probably affects income, and income is more proximately related to health. Often, our goal is to find proximate causes of an outcome variable, and we will be discussing how the linear model is often used (albeit somewhat incorrectly) to find proximate causes.

Third, two variables may be correlated, but the relationship may not be a causal one. Generally, when we say that two variables are related, we are generally thinking about this at the within-individual level; that as a characteristic changes for an *individual* it will influence some other characteristic of the *individual*. Yet, our models generally capture covariance at the between-individual level. The fact that gender, for example, covaries with income DOES NOT imply that a sex change operation will automatically lead to an increase in pay. With fixed characteristics, like gender or race, we often use causal terminology, but because

gender cannot change, it technically cannot be a cause of anything. Instead, there may be more proximate and changeable factors that are associated with gender which are also associated with the outcome variable in which we are interested. Experimentalists realize this, and experiments involve observing average within-individual change, with a manipulable intervention.

As another example, life course research emphasizes three different types of effects of time: age, period, and cohort effects. Age effects refer to biological or social processes that occur at the within-individual level as the individual ages. Period effects refer to historical processes that occur at the macro level at some point in time and influence individuals at multiple ages. Cohort effects, at least one interpretation of them, refer to the interaction of period with age. A period event at time  $t$  may affect persons age  $x$  at time  $t$  differently than persons age  $x+5$  at  $t$ . Think, for example, about the difference between computer knowledge of persons currently age 20 versus those currently age 70. The fact that 70 year-olds know less about computers is not an artifact of some cognitive function decline across age-it's due to the differential effect of the period event of the invention of the home PC across birth cohorts.

Recognizing these different types of time effects demonstrates why our models may fail at determining causality. Imagine some sort of life course process that is stable across an individual's life course but may vary across birth cohorts-suppose it is decreasing across birth cohorts. Assume that we take a cross section and look at the age pattern-we will observe a linear age pattern, even though this is not true for any individual:

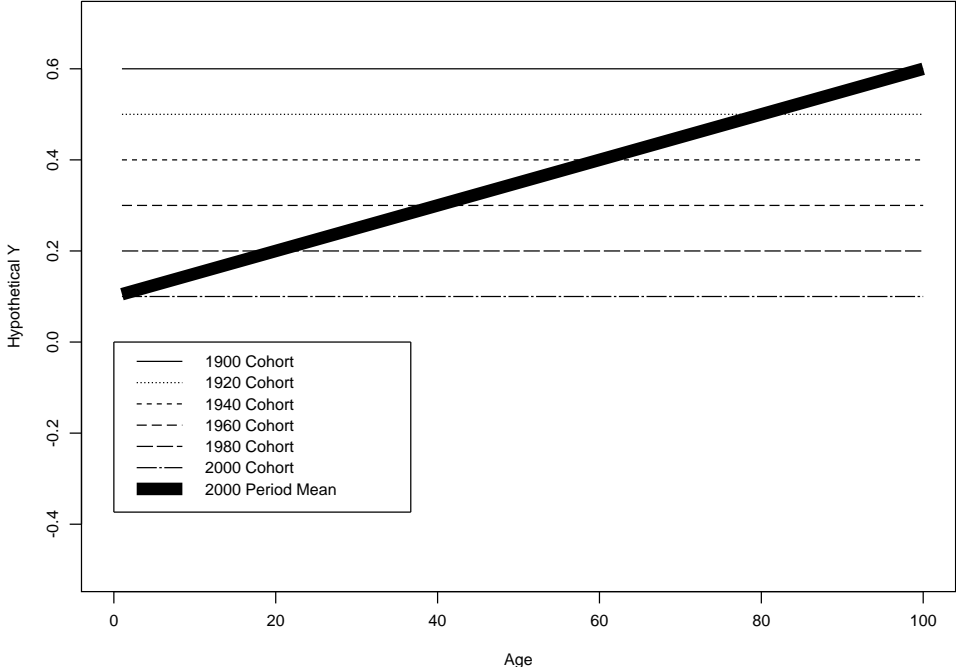


Figure 2. Hypothetical Y by Age

This example, as well as the ice cream and rape example, highlights additional fallacies that we need to be aware of when thinking causally: ecological fallacies and individualistic fallacies. If we assume that, because ice cream consumption rates and rape rates were related, that people who eat ice cream are rapists, this would be committing the ecological fallacy. Briefly stated, macro relationships aren't necessarily true at the micro level. To give a more reasonable example, we know that SES is related to heart disease mortality at the macro level (with richer countries having greater rates of heart disease mortality). This does not imply that having low SES at the individual level is an advantage—we know that the pattern is reversed at that level. The explanation for the former (macro) finding may be differences in competing risks or diet at the national level. The explanation for the latter (micro) finding may be differences in access to health care, diet, exercise, etc.

The individualistic fallacy is essentially the opposite fallacy—reverse the arguments above. We cannot infer macro patterns from micro patterns. Furthermore, exceptions to a pattern don't invalidate the pattern.

Because of the various problems with causality and its modeling, we will stay away from thinking causally, although to some extent, the semantics we use in discussing regression will be causal.

## 2 Back to Regression

Recalling the example above under 'data summarization,' we have data on death rates by age, and we would like to summarize the data by using a linear model. We saw 3 equations above. In these equations,  $\beta_0$  is the linear intercept term,  $\beta_1$  is the (linear) effect of age on death rates, and  $\epsilon_i$  is an error term that captures the difference (vertical distance) between the line implied by the coefficients and the actual data we observed—obviously every observation cannot fall exactly on a straight line through the data.

We assume that the first equation holds in the population. However, we don't have population data. So, we change our notation slightly to the second equation.

In the third equation, the error term has dropped out of the equation, because  $\hat{Y}$  ("y-hat") is the expected (mean) score for the rate applicable to a particular age. Note that simple algebra yields  $e_i = Y_i - \hat{Y}_i \equiv (b_0 + b_1 \text{Age}_i)$ .

## 3 Least Squares Estimation

How do we find estimates of the regression coefficients? We should develop some criteria that lead us to prefer one set of coefficients over another set.

One reasonable strategy is to find the line that gives us the least error. This must be done for the entire sample so, we would like the  $b_0$  and  $b_1$  that yield  $(\min \sum_{i=1}^n e_i)$ . However, the sum of the raw errors will always be 0, so long as the regression line passes through the point  $(\bar{x}, \bar{y})$ . An alternate strategy is to find the line that gives us the least absolute value error:  $(\min \sum_{i=1}^n |e_i|)$ . We will discuss this strategy later in the semester.

It is easier to work with squares (which are also positive), so we may consider:  $(\min \sum_{i=1}^n e_i^2)$ . In fact, this is the criteria generally used, which is why it is called Ordinary Least Squares regression.

We need to minimize this term, so, recalling from calculus that a maximum or minimum is reached wherever a curve inflects, we simply need to take the derivative of the least squares function, set it equal to 0, and solve for  $b_0$  and  $b_1$ .

There is one catch-in this model we have two parameters that we need to solve for. So, we need to take two partial derivatives: one with respect to  $b_0$  and the other with respect to  $b_1$ . Then we will need to solve the set of two equations:

$$\frac{\partial F}{\partial b_0} = \frac{\partial}{\partial b_0} \left[ \sum_{i=1}^n (Y_i - (b_0 + b_1 X_i))^2 \right] = 2nb_0 - 2 \sum Y_i + 2b_1 \sum X_i$$

and:

$$\frac{\partial F}{\partial b_1} = \frac{\partial}{\partial b_1} \left[ \sum_{i=1}^n (Y_i - (b_0 + b_1 X_i))^2 \right] = -2 \sum X_i Y_i + 2b_0 \sum X_i + 2b_1 \sum X_i^2.$$

Setting these equations to 0 and solving for  $b_0$  and  $b_1$  yields:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

and

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Notice that the denominator of  $b_1$  is  $(n-1) \times s_x^2$ , and the numerator is  $(n-1) \times cov(X, Y)$ . So,  $b_1 = \frac{Cov(X, Y)}{Var(X)}$ .

## 4 Maximum Likelihood Estimation

An alternative approach to estimating the coefficients is to use maximum likelihood estimation. Recall that for ML estimation, we first establish a likelihood function. If we assume the errors ( $e_i$ ) are  $\sim N(0, s_e)$ , then we can establish a likelihood function based on the error. Once again, assuming observations are independent, the joint pdf for the data given the parameters (the likelihood function) is:

$$p(Y | b_0, b_1, X) \equiv L(b_0, b_1 | X, Y) = \prod_{i=1}^n \frac{1}{s_e \sqrt{2\pi}} \exp \left\{ -\frac{(Y_i - (b_0 + b_1 X_i))^2}{2s_e^2} \right\}$$

This likelihood reduces to:

$$p(Y | b_0, b_1, X) \equiv L(b_0, b_1 | X, Y) = (2\pi s_e^2)^{-\frac{n}{2}} \exp \left\{ -\frac{\sum_{i=1}^n (Y_i - (b_0 + b_1 X_i))^2}{2s_e^2} \right\}$$

Taking the log of the likelihood, we get:

$$LL(b_0, b_1, X) \propto -n \log(s_e) - \frac{1}{2s_e^2} \left( \sum_{i=1}^n (Y_i - (b_0 + b_1 X_i))^2 \right)$$

Taking the derivative of this function with respect to each parameter yields the following 3 equations:

$$\frac{\partial LL}{\partial b_0} = -\frac{1}{2s_e^2} \left( 2nb_0 - 2 \sum Y_i + 2b_1 \sum X_i \right)$$

and

$$\frac{\partial LL}{\partial b_1} = -\frac{1}{2s_e^2} \left( -2 \sum X_i Y_i + 2b_0 \sum X_i + 2b_1 \sum X_i^2 \right)$$

and

$$\frac{\partial LL}{\partial s_e} = s_e^{-3} \sum (Y_i - (b_0 + b_1 X_i))^2 - \frac{n}{s_e}$$

Setting these partial derivatives equal to 0 and solving for the parameters yields the same values as the OLS approach. Furthermore, the error variance is found to be:

$$s_e^2 = \frac{\sum e_i^2}{n}$$

However, due to estimation, we lose 2 ‘degrees of freedom,’ making the unbiased denominator  $n - 2$ , rather than  $n$ . Realize that this end result is really nothing more than the average squared error. If we take the square root, we get the ‘standard error of the regression,’ which we can use to construct measures of model fit.

Using the results above, the regression for the death rate data yields the following line:

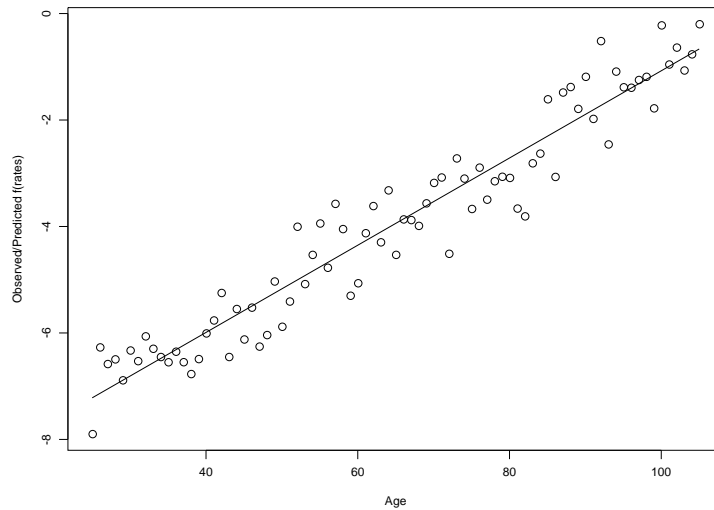


Figure 3. Observed and Predicted F(death rates)