Stat 400, section 12.1 Simple Linear Regression Model

notes by Tim Pilachowski

Recall back to Chapter 5.

Definition: Given two discrete random variables *X* and *Y* defined on a sample space *S*, the joint probability mass function is defined for each ordered pair of numbers (*x*, *y*) by

$$p(x, y) = P_{X,Y}(x, y) = P(X = x, Y = y).$$

Our investigations included independence of jointly distributed random variables, covariance and the calculation of the correlation coefficient. We also had a Theorem: If two random variables *X* and *Y* are independent, then Cov(*X*, *Y*) = 0.

But, what if the two random variables are not independent, and there is a correlation (read "connection", not "cause") between *X* and *Y*? Can we specify the nature of the relationship in some meaningful way?

The simplest deterministic mathematical relationship between two variables *x* and *y* is a linear relationship $y = \beta_0 + \beta_1 x$. That is, the set of pairs (*x*, *y*) which relate *Y* to *X* determines a line with slope β_1 and *y*-intercept at $y = \beta_0$. If the two variables *are* **deterministically** related, then for a fixed value of *x*, the observed value of the associated *y* will always be the same. That is, *y* will always be $\beta_0 + \beta_1 x$.

If the two variables *are not* deterministically related, then for a fixed value of *x*, there is uncertainty in the value of the second variable. We'll call a non-deterministic model a **probabilistic model**.

In Example A below, we designate Math Anxiety Score as random variable *X*, and Math Grade as *Y*. If we decide to select a middling value of *x* = 50 for Math Anxiety, then the observed value *y* might take on any number of values. That is, before the selection of Math Anxiety Score *X* is made, Math Grade is a random variable *Y*.

More generally, the variable whose value is fixed by the experimenter will be denoted by *x* and will be called the **independent, predictor,** or **explanatory variable.** For fixed *x*, the second variable will be random. We denote this random variable *Y*, and denote its observed value by *y*, and refer to it as the **dependent** or **response variable.**

Usually observations will be made for a number of settings of the independent variable.
We'll use $x_1, x_2, \ldots, x_n$ to denote values of the independent variable for which observations are made, and let Y_i and y_i, respectively, denote the random variable and observed value associated with x_i. Our data will then consists of *n* coordinate pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

A graph of this data is called a **scatter plot**, and gives preliminary impressions about the nature of any relationship. (Go to Example A-1 below.)

For the deterministic model $y = \beta_0 + \beta_1 x$, the actual observed value of *y* is a linear function of *x*.
The appropriate generalization of this to a probabilistic model assumes that *the expected value of Y is a linear function of x*, but that for fixed *x* the variable *Y* differs from its expected value by a random amount.

Definition: The Simple Linear Regression Model
There are parameters β_0, β_1, and σ^2, such that for any fixed value of the independent variable *x*, the dependent variable is a random variable related to *x* through the **model equation** $Y = \beta_0 + \beta_1 x + \in$.
The quantity $\in$ in the model equation is a random variable, assumed to be normally distributed, with $E(\in) = 0$ and $V(\in) = \sigma^2$.

The variable $\in$ is usually referred to as the **random deviation** or **random error term** in the model. In a deterministic model, any observed pair (x, y) would correspond to a point falling exactly on the line $y = \beta_0 + \beta_1 x$, called the **true** (or **population**) **regression line.** In a probabilistic model, the inclusion of the random error term $\in$ allows an observed (x, y) pair to fall either above the true regression line (when $\in > 0$) or below the line (when $\in < 0$). In other words, for a fixed value x, there can be more than one observed corresponding value for y, some of which may be larger and some of which may be smaller than the y-value calculated using the regression equation. The n coordinate pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ resulting from n independent observations will thus be scattered about the true regression line. (Go to Example A-2 below.)

Theory: Let x^* denote a particular value of the independent variable x and

$\mu_{Y \cdot x^*} =$ the expected (or mean) value of Y when x has value x^* [alternative notation $E(Y \mid x^*)$]

$\sigma^2_{Y \cdot x^*} =$ the variance of Y when x has value x^* [alternative notation $V(Y \mid x^*)$].

If we think of an entire population of (x, y) pairs, then $\mu_{Y \cdot x^*}$ is the mean of all y values for which $x = x^*$, and $\sigma^2_{Y \cdot x^*}$ is a measure of how much these values of y spread out above and below the mean value. (Go to Example A-3.)

Once x is fixed, the only randomness on the right-hand side of the model equation $Y = \beta_0 + \beta_1 x + \in$ is in the random error $\in$, whose mean value and variance are 0 and σ^2, respectively, whatever the value of x.

notes on the proof:

$$\begin{aligned} \mu_{Y \cdot x^*} &= E(Y \text{ when } x = x^*) \\ &= E(\beta_0 + \beta_1 x^* + \in) \\ &= \beta_0 + \beta_1 x^* + E(\in) \\ &= \beta_0 + \beta_1 x^* \end{aligned}$$

$$\begin{aligned} \sigma^2_{Y \cdot x^*} &= V(Y \text{ when } x = x^*) \\ &= V(\beta_0 + \beta_1 x^* + \in) \\ &= V(\beta_0 + \beta_1 x^*) + V(\in). \\ &= \quad 0 \quad + \sigma^2 \\ &= \sigma^2 \end{aligned}$$

The change of variables formulas were developed back in Lecture 3.6b.

Replacing x^* in $\mu_{Y \cdot x^*}$ by x gives the relation $\mu_{Y \cdot x} = \beta_0 + \beta_1 x$, which says that the *mean value* of Y, rather than Y itself, is a linear function of x.

The true regression line $y = \beta_0 + \beta_1 x$ is thus the *line of mean values* or *line of expected values*; its height above any particular x value is the expected value of Y for that value of x.

The slope β_1 of the true regression line is interpreted as the *expected* change in Y associated with a 1-unit increase in the value of x. The relation $\sigma^2_{Y \cdot x} = \sigma^2$ tells us that the amount of variability in the distribution of Y values is the same at each different value of x (homogeneity of variance).

For fixed x, Y is the sum of a constant $\beta_0 + \beta_1 x$ and a normally distributed random variable $\in$, so Y itself has a normal distribution. Thus, we can use the true regression equation and the normal distribution table to determine probabilities that the observed y-values resulting from a selected x^* will fall within a given interval. (Go to Example A-4 below.)

Example A: A paper in *Measurement and Evaluation in Counseling and Development* (Oct 90, pp. 121–127) discussed a survey instrument called the *Mathematics Anxiety Scale for Children* (MASC). Suppose the MASC was administered to ten fifth graders with the following results:

MASC Score	67	37	70	40	35	65	40	35	30	40
Math grade (%)	75	85	60	90	80	75	70	90	95	80

1. Draw a scatter diagram in the space to the right below. (MASC is the independent variable.)

What preliminary impressions do you have of the relationship (assuming there is one) between MASC score and Math grade?

2. Suppose that the true regression line that relates MASC score (x) to Math grade (Y) is $Y = 100 - 0.5x$. Calculate the value of $\in$ for the 6th through the 10th data points.

3. Explore the meaning of $\mu_{Y \cdot x^*}$ in the context of MASC score (x) to Math grade (Y) for $x^* = 40$.

4. Suppose that the true regression line that relates MASC score (x) to Math grade (Y) is $Y = 100 - 0.5x$ with $\sigma = 3$. Find $P(Y > 86$ when $x = 40)$, $P(Y > 86$ when $x = 45)$, and $P(Y_2$ exceeds Y_1 when $X_2 = 45, X_1 = 40)$.