LECTURE NOTE 1

LINEAR PREDICTOR AND LEAST-SQUARES FIT

1. LINEAR PREDICTOR

Consider a random sample of n individuals that provides data on their earnings and education. Consider the first individual in the sample, and let Y denote her earnings and let X denote her education. I want you to think of (Y, X) as a pair of *random variables*. The randomness comes from the act of random sampling: before this individual is drawn from the population, we do not know what the earnings and education will turn out to be, but we can assign a joint distribution to (Y, X).

It would be nice if there were a function connecting Y and X: $Y = f(X)$, but no, individuals with the same education may have different earnings. A more promising goal is to establish a relationship in a predictive sense. Given the value of X, we can try to predict the value of Y, and a good place to start is a *linear predictor*:

$$\hat{Y} = \beta_0 + \beta_1 X.$$

Now we have to say how those coefficients β_0 and β_1 are going to get determined. A very convenient criterion is the square of the prediction error, and we choose β_0 and β_1 to minimize its expectation:

$$\min_{\beta_0, \beta_1} E(Y - \hat{Y})^2.$$

So a more complete description of our linear predictor is *minimum mean square error* linear predictor.

Similar minimization problems come up elsewhere in the course, and on the principle that "the same equations have the same solutions," I'd like to once and for all lay out a way to solve these problems. The key is to use *orthogonal projection* in a vector space with an *inner product*. Here the inner product is

$$\langle Y, X \rangle = E(YX).$$

The associated *norm* is

$$||Y|| = \langle Y, Y \rangle^{1/2}.$$

Then we can restate our linear predictor problem as

$$\min_{\beta_0, \beta_1} ||Y - \hat{Y}||^2.$$

The solution is obtained from the orthogonal projection of Y on 1 and X. It is convenient to define X_0 as a degenerate random variable that only takes on the value 1. The orthogonal projection requires that the prediction error $(Y - \hat{Y})$ is orthogonal to X_0 and X:

$$\langle Y - \hat{Y}, X_0 \rangle = 0,$$
$$\langle Y - \hat{Y}, X \rangle = 0.$$

Notation for this orthogonality is

$$Y - \hat{Y} \perp X_0, \quad Y - \hat{Y} \perp X.$$

Writing out the two orthogonality conditions gives

$$\langle Y - \beta_0 X_0 - \beta_1 X, X_0 \rangle = \langle Y, X_0 \rangle - \beta_0 \langle X_0, X_0 \rangle - \beta_1 \langle X, X_0 \rangle = 0,$$
$$\langle Y - \beta_0 X_0 - \beta_1 X, X \rangle = \langle Y, X \rangle - \beta_0 \langle X_0, X \rangle - \beta_1 \langle X, X \rangle = 0.$$

Using our definition for the inner product,

$$E(Y) - \beta_0 - \beta_1 E(X) = 0,$$
$$E(YX) - \beta_0 E(X) - \beta_1 E(X^2) = 0.$$

This gives two linear equations for the two unknowns, β_0 and β_1. These equations can be solved to give

$$\beta_1 = \frac{E(YX) - E(Y)E(X)}{E(X^2) - E(X)E(X)}$$
$$\beta_0 = E(Y) - \beta_1 E(X).$$

The numerator in the expression for β_1 can be taken as the definition of *covariance*:

$$\text{Cov}(Y, X) \equiv E(YX) - E(Y)E(X),$$

and the denominator can be taken as the definition of *variance*:

$$\text{Var}(X) \equiv E(X^2) - E(X)E(X).$$

So we can rewrite the slope coefficient in the linear predictor as

$$\beta_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}.$$

Our notation for the (population) linear predictor is

$$E^*(Y \mid 1, X) = \beta_0 + \beta_1 X.$$

2. LEAST-SQUARES FIT

The data from a sample of size n can be put into two matrices:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

and it convenient to define an additional matrix x_0, which is simply a $n \times 1$ column of 1's:

$$x_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

The fitted value for the i^{th} observation is

$$\hat{y}_i = b_0 + b_1 x_i,$$

and the objective is to choose the coefficients b_0 and b_1 to minimize the sum of squared residuals:

$$\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

(Dividing by n is not necessary, but it does suggest an analogy with minimizing mean square error in the population.)

Define the inner product

$$\langle y, x \rangle = \frac{1}{n} \sum_{i=1}^{n} y_i x_i.$$

Now we have a minimum norm problem:

$$\min_{b_0, b_1} ||y - b_0 x_0 - b_1 x||^2,$$

and the solution, once again, is obtained from the orthogonal projection of y on x_0 and x. This requires that the prediction error $(y - \hat{y})$ be orthogonal to x_0 and x:

$$\langle y - \hat{y}, x_0 \rangle = 0,$$

$$\langle y - \hat{y}, x \rangle = 0.$$

Notation for this orthogonality is

$$y - \hat{y} \perp x_0, \quad y - \hat{y} \perp x.$$

Writing out the orthogonality conditions gives

$$\langle y, x_0 \rangle - b_0 \langle x_0, x_0 \rangle - b_1 \langle x, x_0 \rangle = 0,$$

$$\langle y, x \rangle - b_0 \langle x_0, x \rangle - b_1 \langle x, x \rangle = 0.$$

Using our definition for the (least-squares) inner product, we have

$$\bar{y} - b_0 - b_1 \bar{x} = 0,$$

$$\overline{yx} - b_0 \bar{x} - b_1 \overline{x^2} = 0,$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \overline{yx} = \frac{1}{n} \sum_{i=1}^{n} y_i x_i, \quad \overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2.$$

The two linear equations for the two unknowns, b_0 and b_1, can be solved to give

$$b_1 = \frac{\overline{yx} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}\bar{x}},$$

$$b_0 = \bar{y} - b_1 \bar{x}.$$

Our notation for the least-squares fit (or sample linear predictor) is

$$\hat{y}_i \mid 1, x = b_0 + b_1 x_i.$$

3. GOODNESS OF FIT

Note that

$$0 \leq \frac{||Y - E^*(Y \mid 1, X)||^2}{||Y - E^*(Y \mid 1)||^2} \leq 1.$$

This ratio is less than or equal to 1 because using X to predict Y cannot increase the mean square error—β_1 is allowed to be 0. (The linear predictor using just a constant is $E^*(Y \mid 1) = E(Y)$.) We define a measure of goodness of fit in the population as

$$R^2_{\text{pop}} = 1 - \frac{||Y - E^*(Y \mid 1, X)||^2}{||Y - E^*(Y \mid 1)||^2}.$$

This measure is scale free in that it is not affected if Y is multiplied by a constant (for example, changing the units from dollars to cents). It is easy to interpret since

$$0 \leq R^2_{\text{pop}} \leq 1.$$

The sample counterpart is

$$R^2 = 1 - \frac{||y - (\hat{y} \,|\, 1, x)||^2}{||y - (\hat{y} \,|\, 1)||^2}.$$

(The least-squares fit using just a constant is $(\hat{y} \,|\, 1) = \bar{y}$.) It is also scale free with

$$0 \leq R^2 \leq 1.$$

4. OMITTED VARIABLES

Consider an individual chosen at random from a population. Let Y denote her earnings, and let X_1 and X_2 denote her education and her score on a test administered when she was in the third grade. The random variables (Y, X_1, X_2) have a joint distribution. There is a (population) linear predictor for Y given X_1 and X_2 (and a constant):

$$E^*(Y \,|\, 1, X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{\textit{Long}}$$

and there is a (population) linear predictor for Y just given X_1 (and a constant):

$$E^*(Y \,|\, 1, X_1) = \alpha_0 + \alpha_1 X_1. \tag{\textit{Short}}$$

I want to develop the relationship between these two linear predictors. This requires the auxiliary linear predictor of X_2 given X_1 (and a constant):

$$E^*(X_2 \,|\, 1, X_1) = \gamma_0 + \gamma_1 X_1. \tag{\textit{Aux}}$$

Let U denote the prediction error using the long predictor:

$$U \equiv Y - E^*(Y \,|\, 1, X_1, X_2),$$

so that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U. \tag{1}$$

Because U is a prediction error, it is orthogonal to the variables used in the predictor:

$$U \perp 1, \quad U \perp X_1, \quad U \perp X_2.$$

In particular, U is orthogonal to 1, X_1, which implies that

$$E^*(U \mid 1, X_1) = 0. \tag{2}$$

Use equations (1) and (2) to write the short predictor as

$$\begin{aligned} E^*(Y \mid 1, X_1) &= \beta_0 + \beta_1 X_1 + \beta_2 E^*(X_2 \mid 1, X_1) + E^*(U \mid 1, X_1) \\ &= \beta_0 + \beta_1 X_1 + \beta_2(\gamma_0 + \gamma_1 X_1) + 0 \\ &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)X_1. \end{aligned}$$

So we have proved the following

Claim 1. $\alpha_0 = \beta_0 + \beta_2\gamma_0, \quad \alpha_1 = \beta_1 + \beta_2\gamma_1.$

The coefficient α_1 on X_1 in the short predictor is the coefficient β_1 from the long predictor plus an additional term. This additional term is the product of the coefficient β_2 on the omitted variable and the coefficient γ_1 on X_1 in the auxiliary predictor. This result is often called the *omitted variable bias* formula. If the goal is the coefficient on X_1 in the linear predictor that includes X_1 and X_2, then the coefficient on X_1 in the short predictor differs from this goal by $\beta_2\gamma_1$. Note that this bias term is 0 if $\gamma_1 = 0$, which holds if $\mathrm{Cov}(X_1, X_2) = 0$.

There is a similar result for the least-squares fit using sample data. Our notation for the long, short, and auxiliary least-squares fit is

$$\hat{y}_i \mid 1, x_{i1}, x_{i2} = b_0 + b_1 x_{i1} + b_2 x_{i2},$$

$$\hat{y}_i \mid 1, x_{i1} = a_0 + a_1 x_{i1},$$

$$\hat{x}_{i2} \mid 1, x_{i1} = c_0 + c_1 x_{i1}.$$

The argument above using the population predictors translates directly into an argument using sample predictors (least-squares fits). Just change the inner product from $E(XY)$ to $\sum_{i=1}^{n} y_i x_i / n$. This gives

Claim 2. $a_0 = b_0 + b_2 c_0, \quad a_1 = b_1 + b_2 c_1.$

This least-squares version of the omitted variable bias formula is a computational identity, which can be checked on a data set using a least-squares computer program.