1 Review of Probability (Continued)

1.1 The law of iterated expectations

Consider two random variables X and Y. The mean of Y is the weighted average of the conditional expectation of Y given X, weighted by the probability distribution of X. Stated mathematically, if X takes on n values $x_1, \ldots, x_n$, then

$$\mathbb{E}(Y) = \sum_{i=1}^{n} \mathbb{E}(Y|X = x_i) Pr(X = x_i).$$

Stated differently, the expectation of Y is the expectation of the conditional expectation of Y given X,

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|X)] \tag{1}$$

where the inner expectation on the right-hand side of Equation (1) is computed using the conditional distribution of Y given X and the outer expectation is computed using the marginal distribution of X. For example, to get the average test score of a class at UC Berkeley, we can directly compute the average test score for everyone in the class, or we can proceed as follows: (i) compute the average test score for Californians, (ii) compute the average test score for non-Californians, then (iii) weigh the average test score for Californians and non-Californians by the share of Californians and non-Californians in the class, respectively.

Equation (1) is known as the **law of iterated expectations**, or LIE.

1.2 Correlation and conditional mean

If the conditional mean of Y does not depend on X, then Y and X are uncorrelated. That is,

$$\text{if } \mathbb{E}(Y|X) = \mu_Y, \text{then } Cov(Y, X) = 0 \text{ and } Corr(Y, X) = 0. \tag{2}$$

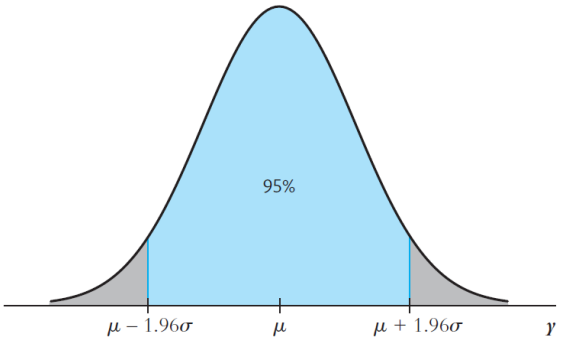
It is *not* necessarily true, however, that if X and Y are uncorrelated, then the conditional mean of Y given X does not depend on X.

Exercise 1:

Suppose Y and X have mean 0. Show that if $\mathbb{E}(Y|X) = \mu_Y$, then $Cov(Y, X) = 0$ and $Corr(Y, X) = 0$.

1.3 Normal, Chi-Squared, and Student t Distributions

A continuous random variable with a **normal distribution** has the familiar bell shaped probability density shown in the following figure. The normal density with mean μ and variance σ^2 is denoted as $N(\mu, \sigma^2)$. It is symmetric around its mean and has 95% of its probability between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$. Let Z be a random variable with a **standard normal distribution**, the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$, which is shown as $Z \sim N(0, 1)$. Its cumulative distribution function (CDF) is denoted by the Greek letter Φ; accordingly, $Pr(Z \leq c) = \Phi(c)$, where c is a constant. Values of the standard normal CDF are tabulated in Appendix Table 1 in the Stock and Watson textbook.



To look up probabilities for a random variable Y that follows a normal distribution with a general mean μ and a variance σ^2, we must first standardize the variable to follow a standard normal distribution. Each value of Y is standardized by subtracting its mean μ from this value and dividing the result by its standard deviation σ. Once this is done, we can use the standard normal distribution table to get probabilities of the form $Pr(Y \leq c)$, where c is a constant.

Exercise 2:

Suppose the random variable Y follows a normal distribution with mean of 1 and a variance of 4. Let $Z = \frac{1}{2}(Y - 1)$. Show that $\mathbb{E}(Z) = 0$ and $Var(Z) = 1$. What is the probability $Pr(Y \leq 2)$?

The **chi-squared distribution** is the distribution of the sum of m squared independent standard normal random variables. This distribution depends on m, denoted as χ^2_m. m is called the degrees of freedom of the chi-squared distribution.

The **Student** t **distribution** with m degrees of freedom is defined to be the distribution of the ratio of a standard normal random variable to the square root of an independently distributed chi-squared random variable with m degrees of freedom divided by m. This distribution is denoted t_m. The Student t distribution depends on the degrees of freedom m. The Student t distribution has a bell shape similar to that of the normal distribution, but it has more mass in the tails. When m is 30 or more, the Student t distribution is well approximated by the standard normal distribution, and the t_∞ distribution equals the standard normal distribution.

2 Review of Optimization

Optimization is the process of finding the maximum or minimum value of an objective function under given constraints. In econometrics, optimization techniques are extensively used for tasks such as estimating model parameters, testing hypothesis, etc. Here we consider unconstrained optimization.

Consider a smooth and continuous function $f(x)$ defined over some domain D. We want to find a point x^* such that $f(x^*)$ is a minimum or maximum within D. This involves two steps:

1. *Finding Critical Points (First-order Condition)*: A point x^* is a critical point of $f(x)$ if $f'(x^*) = 0$ or $f'(x^*)$ is undefined. We compute the derivative $f'(x)$, then solve the equation $f'(x) = 0$ to find potential critical points, and verify if these points lie in the domain D.

2. *Check for Optimality (Second-order Condition)*: To determine if a critical point is a maximum, minimum, or neither, we compute and evaluate the second derivative $f''(x)$ at the critical points: If $f''(x) > 0$, then $f'(x^*)$ is a local minimum; if $f''(x) < 0$, then $f'(x^*)$ is a local maximum; if $f''(x) = 0$, then the test is inconclusive.

For functions of two or more variables, the concept is essentially the same, except for the fact that we are now working with partial derivatives. Consider a function $f(x_1, x_2, \ldots, x_n)$ defined over some domain D. We want to find a point $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ such that $f(\mathbf{x}^*)$ is a minimum or maximum within D.

1. *Finding Critical Points (First-order Condition)*: A point $\mathbf{x}^*$ is a critical point of $f(\mathbf{x})$ if the gradient $\nabla f(\mathbf{x}^*) = \mathbf{0}$ or if the gradient is undefined. We compute the gradient $\nabla f(\mathbf{x}) = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}]$. Solve the system of equations $\frac{\partial f}{\partial x_1} = 0$, $\frac{\partial f}{\partial x_2} = 0$,...,$\frac{\partial f}{\partial x_n} = 0$ to find potential critical points.

2. *Check for Optimality (Second-order Condition)*: To determine if a critical point is a maximum, minimum, or saddle point, we compute and evaluate the Hessian matrix H consisting of second-order partial derivatives,

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

   Then we apply the following conditions: If the Hessian is positive-definite at $\mathbf{x}^*$, then $f(\mathbf{x}^*)$ is a local minimum; if the Hessian is negative-definite at $\mathbf{x}^*$, then $f(\mathbf{x}^*)$ is a local maximum; if the Hessian has both positive and negative eigenvalues, then $\mathbf{x}^*$ is a saddle point for f; otherwise the test is inconclusive. The condition for a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ to be positive-definite is $det(A) = ad - bc > 0$ and $a > 0$, and for A to be negative-definite, the condition is that $det(A) > 0$ and $a < 0$.

Exercise 3:

Consider minimizing the following function with respect to the arguments (b_0, b_1)

$$f(b_0, b_1) = vb_0^2 + wb_1^2 + xb_1b_0 + yb_1 + zb_0,$$

where (v, w, x, y, z) are known constants.

1. Write down the first order conditions for this problem. These should give you two equations in two unknowns (remember the unknowns are b_0 and b_1). Solve these equations for the optimal values of (b_0, b_1) as a function of the variables (v, w, x, y, z).

2. How would you check whether the candidate values you found in part (a) above are correct?

3 Motivating Ordinary Least Squares (OLS)

Assume that the Conditional Expectation Function (CEF) or Regression Function (RF) is **linear**

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X, \tag{3}$$

where β_0 and β_1 are unknown fixed constants known as the **population parameters**. The problem of learning the relationship between X and Y is now reduced to learning about (β_0, β_1). We would like to know the values of (β_0, β_1) or at least find a good guess for it. The only information we have is data $(X_i, Y_i)_{i=1}^n$. We split this problem into two steps:

1. **First Step:** Learn about population parameters under the assumption that the distribution f is known.
2. **Second Step**: Use Step 1 to learn about population parameters using only the data (without knowing the distribution of f).

3.1 First Step

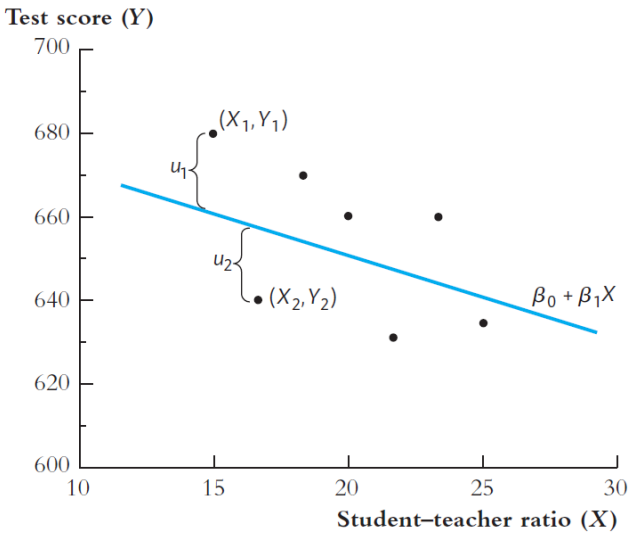
It turns out that under (3) we can find β as follows

$$(\beta_0, \beta_1) = \arg \min_{(b_0, b_1)} \mathbb{E}\left(Y - b_0 - b_1 X\right)^2$$

So if we knew how to take these expectations (remember, we need to know the joint distribution to do that), we could find (β_0, β_1) by solving the problem above.

FIGURE 4.1 Scatterplot of Test Score vs. Student-Teacher Ratio (Hypothetical Data)

The scatterplot shows hypothetical observations for seven school districts. The population regression line is $\beta_0 + \beta_1 X$. The vertical distance from the i^{th} point to the population regression line is $Y_i - (\beta_0 + \beta_1 X_i)$, which is the population error term u_i for the i^{th} observation.



Proof: Consider finding the minimum of the function

$$\begin{aligned} q\left(b_0, b_1\right) &\equiv \mathbb{E}\left(Y - b_0 - b_1 X\right)^2 \\ &= \mathbb{E}\left(Y^2\right) + b_0^2 + b_1^2 \mathbb{E}\left(X^2\right) + 2 b_0 b_1 \mathbb{E}\left(X\right) \\ &\quad -2\mathbb{E}\left(Y\left(b_0 + b_1 X\right)\right) \end{aligned}$$

Recall that to find an optimum we take (partial) derivatives and find the values of (b_0, b_1) that sets them equal to zero

$$\frac{\partial q}{\partial b_0} = 2b_0 + 2b_1\mathbb{E}(X) - 2\mathbb{E}(Y)$$
$$\frac{\partial q}{\partial b_1} = 2b_1\mathbb{E}\left(X^2\right) + 2b_0\mathbb{E}(X) - 2\mathbb{E}(XY)$$

Now, find $\left(b_0^*, b_1^*\right)$ such that

$$2b_0 + 2b_1\mathbb{E}(X) - 2\mathbb{E}(Y) = 0$$
$$2b_1\mathbb{E}\left(X^2\right) + 2b_0\mathbb{E}(X) - 2\mathbb{E}(XY) = 0$$

This is a system of two linear equations with two unknowns.

$$\underbrace{\begin{bmatrix} 1 & \mathbb{E}(X) \\ \mathbb{E}(X) & \mathbb{E}\left(X^2\right) \end{bmatrix}}_{A} \begin{bmatrix} b_0^* \\ b_1^* \end{bmatrix} = \begin{bmatrix} \mathbb{E}(Y) \\ \mathbb{E}(XY) \end{bmatrix}$$

Recall that you can solve this system of equations for (b_0, b_1) as long at A is invertible.

$$\begin{bmatrix} b_0^* \\ b_1^* \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{E}(X) \\ \mathbb{E}(X) & \mathbb{E}\left(X^2\right) \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}(Y) \\ \mathbb{E}(XY) \end{bmatrix}$$

and recall that for a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, we can find the inverse directly using the formula:

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

(Note that A^{-1} exists only when $ad - bc \neq 0$.)

$$\begin{bmatrix} 1 & \mathbb{E}(X) \\ \mathbb{E}(X) & \mathbb{E}\left(X^2\right) \end{bmatrix}^{-1} = \underbrace{\frac{1}{\mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2}}_{Var(X)} \begin{bmatrix} \mathbb{E}\left(X^2\right) & -\mathbb{E}(X) \\ -\mathbb{E}(X) & 1 \end{bmatrix}$$

$$\begin{aligned}
\begin{bmatrix} b_0^* \\ b_1^* \end{bmatrix} &= \frac{1}{Var(X)} \begin{bmatrix} \mathbb{E}\left(X^2\right) & -\mathbb{E}(X) \\ -\mathbb{E}(X) & 1 \end{bmatrix} \begin{bmatrix} \mathbb{E}(Y) \\ \mathbb{E}(XY) \end{bmatrix} \\
&= \frac{1}{Var(X)} \begin{bmatrix} \mathbb{E}\left(X^2\right)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(XY) \\ -\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(XY) \end{bmatrix} \\
&= \frac{1}{Var(X)} \begin{bmatrix} \mathbb{E}\left(X^2\right)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(XY) \\ \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \end{bmatrix} \\
&= \frac{1}{Var(X)} \begin{bmatrix} \mathbb{E}\left(X^2\right)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(XY) \\ Cov(X, Y) \end{bmatrix}
\end{aligned}$$

The final formula is as follows:

$$b_1^* = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\mathbb{E}(X^2) - (\mathbb{E}(X))^2} \equiv \frac{Cov(X, Y)}{Var(X)}$$

$$b_0^* = \mathbb{E}(Y) - b_1^*\mathbb{E}(X)$$

To verify this is a minimum, we can look at the matrix of the second derivative.

Since we know that (β_0, β_1) are the unique minimizers to this problem by definition, we conclude that $(\beta_0, \beta_1) = \left(b_0^*, b_1^*\right)$

Relationship between β_1 and the correlation between X and Y.

$$Corr\left(X,Y\right)=\frac{Cov\left(X,Y\right)}{\sqrt{Var\left(X\right)Var\left(Y\right)}}$$

Exercise 4:

Consider two random variables X and Y each of which can take on two values $\{0,1\}$. Their joint distribution is given by

$$\begin{aligned}\mathbb{P}\left(X=0,Y=0\right) &\equiv p \\ \mathbb{P}\left(X=0,Y=1\right) &\equiv q \\ \mathbb{P}\left(X=1,Y=0\right) &\equiv r\end{aligned} \tag{1}$$

1. Express the conditional mean of Y given X, $\mathbb{E}\left(Y|X\right)$, as a function of the form $\beta_0+\beta_1 X$ and express (β_0,β_1) as a function of (p,q,r).

2. What would the value of β_1 be if the correlation between X and Y was equal to zero? What would be the value of β_0?

3.2 Second Step

We don't know the joint distribution f of (X,Y), so we cannot compute the expectations, variances, and covariances defined above. We only have the data $(X_i,Y_i)_{i=1}^n$, which are an independent and identically distributed (**i.i.d**) sample from this (unknown) distribution.

- **Independently Distributed**: All observations are mutually independent.
- **Identically Distributed**: Distribution of (X_1,Y_1) is the same as distribution of (X_k,Y_k) for any k and all have the common distribution f.

The **analogy principle** proposes to replace unknown expectations with sample averages. Specifically, we replace the unknown quantity

$$Q_0(b_0,b_1)\equiv\mathbb{E}\left(Y-b_0-b_1X\right)^2$$

by

$$Q_n(b_0, b_1) \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$$

and minimize this (known) function with respect to (b_0, b_1). The minimizers of this function are known as the least squares ("**ordinary least squares**" or "OLS") estimators of the unknown parameters (β_0, β_1). OLS minimizes the average squared difference between the actual values of Y_i and the predicted values based on the estimated line. OLS estimator picks (b_0, b_1) to make the predicted values as close to the actual values as possible.

Let $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ be the minimizers to the problem above. Recall that

$$\beta_1 = \frac{Cov\left(X, Y\right)}{Var\left(X\right)}$$

and

$$\beta_0 = \mathbb{E}\left(Y\right) - \beta_1 \mathbb{E}\left(X\right)$$

Similarly,

$$\hat{\beta}_1 = \frac{\widehat{Cov\left(X, Y\right)}}{\widehat{Var\left(X\right)}} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \ , \ \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\widehat{Cov\left(X, Y\right)} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right) \quad \widehat{Var\left(X\right)} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$