Eco 240 Quantitative Methods and Forecasting

Notes and Examples

Table of Contents I

Table of Contents II

Table of Contents III

Probability

The basics of probability

- The mutually exclusive results of a random process are called the **outcomes**. An example is rolling a die. The outcomes are 1 through 6. Note that they are **mututally exclusive**. If one outcome occurs, others cannot.
- The **probability** of an outcome is the proportion of times it occurs in the long run. If we roll a fair die millions of times, the number 2 will appear close to one-sixth of the time.
- The set of all possible outcomes is the **sample space**. The sample space, S, of rolling a die is

$$S = \{1, 2, 3, 4, 5, 6\}$$

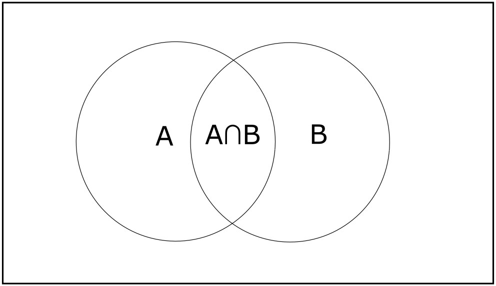
- An **event** is a subset of a sample space. For example, an event, A, might be “at least 3” in the roll of a die. Thus

$$A = \{3, 4, 5, 6\}$$

Probability and Joint Events

Probabilities of joint events can be most easily understood with reference to the diagram below.

Figure 1: Joint Probabilities



The rectangle represents all the outcomes in a sample space. The circles A and B represent events. The probability of event A is the size of the circle A divided by the size of the rectangle.

Joint Probabilities

- The area labeled $A \cap B$ is called “A intersection B” and contains the outcomes in both A and B. If the intersection is zero, the events are **mutually exclusive.** The probability of $A \cap B$ is the size of this area divided by the size of the rectangle. This probability is known as the **joint probability** of A and B.
- What is the probability that an outcome is in A **or** B? The set of such outcomes would include all of A as well as all of B, called “A union B”, and written $A \cup B$. The probability of $A \cup B$ would be the area of A plus the area of B divided by the area of S. But note that $P(A \cup B) \neq P(A) + P(B)$ because this would double count $P(A \cap B)$. Thus

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Joint Probabilities, Examples

- Suppose we roll a fair die once. Let event A be "at least 3". Then
$$A = \{3, 4, 5, 6\}$$
and $P(A) = \frac{4}{6}$.
- Let event B be "even number". Then
$$B = \{2, 4, 6\}$$
and $P(B) = \frac{3}{6}$.
- The intersection of A and B includes the outcomes in **both** A and B.
$$A \cap B = \{4, 6\}$$
The union of A and B includes outcomes in A **or** B.
$$A \cup B = \{1, 3, 4, 5, 6\}$$
and its probability is
$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{4}{6} + \frac{3}{6} - \frac{2}{6} = \frac{5}{6} \end{aligned}$$

Rules of Summation

Let X_1 and X_2 be any two numbers. The sum of these numbers is written

$$X_1 + X_2 = \sum_{i=1}^{2} X_i$$

Σ is the capital Greek letter “sigma” (to indicate “sum”) and i is the **index**. The index starts at the lowest number, shown below sigma, and increases by one until it equals the highest number, shown above sigma. We will often use the following rules of summation. Let Y_1 and Y_2 be two other numbers, and let a and b be constants.

- $\sum aX_i = a \sum X_i$. For example,

$$\sum_{i=1}^{2} 3X_i = 3X_1 + 3X_2 = 3\left(X_1 + X_2\right) = 3\sum_{i=1}^{2} X_i$$

- $\sum_{i=1}^{n} a = na$. For example,

$$\sum_{i=1}^{2} 3 = 3 + 3 = 2 \cdot 3$$

Rules of Summation

- $\sum(X_i + Y_i) = \sum X_i + \sum Y_i$ For example,

$$\sum_{i=1}^{2}(X_i + Y_i) = X_1 + Y_1 + X_2 + Y_2$$

$$= X_1 + X_2 + Y_1 + Y_2 = \sum_{i=1}^{2} X_i + \sum_{i=1}^{2} Y_i$$

- The rules can be combined.

$$\sum_{i=1}^{2}(aX_i + bY_i)^2 = \sum_{i=1}^{2}\left(a^2X_i^2 + b^2Y_i^2 + 2abX_iY_i\right)$$

$$= a^2\sum_{i=1}^{2} X_i^2 + b^2\sum_{i=1}^{2} + 2ab\sum_{i=1}^{2} X_iY_i$$

Random Variables

- A **random variable** is a numerical summary of a random process. For example, suppose we randomly visit a household in New York City. One outcome might be the number of bedrooms in the dwelling. The sample space is $S = \{1, 2, 3, 4, 5, 6, 7\}$. Repeating this experiment many times produces a random variable, which might be called "bedrooms", denoted B.
- This is an example of a **discrete random variable**, because it can take on only a discrete or countable number of values.
- A **continuous** variable takes on a continuum or an uncountable number of values. Income or age are continuous variables because their values vary continuously.
- The possible outcomes of a discrete random variable are customarily arranged in a table as in Table 1. Next to each value is the probability of that value occurring. Thus the probability of choosing an apartment with one child under 6 in NYC is 8.6%. This is expressed $Pr\left(B = 1\right) = 8.6\%$.

Probability Distribution of a Discrete Random Variable

Table 1: Distribution of children under 6 in NYC 2016

Children	Probability Distribution	Cumulative Probability Distribution
0	0.883	0.883
1	0.086	0.969
2	0.027	0.996
3	0.004	1.000
	$\sum Pr = 1$	

Probability Distribution of a Discrete Random Variable

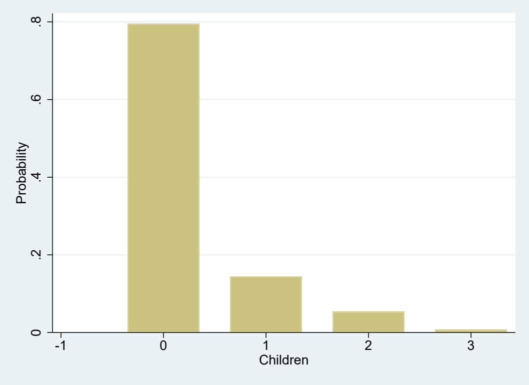
Note three facts about probabilities.

1. $0 \leq Pr\left(C\right) \leq 1$ All probabilities are between zero and one.
2. $Pr\left(C=0\right)+Pr\left(C=1\right)+Pr\left(C=2\right)+Pr\left(C=3\right)=1.$ The sum of the probabilities of every outcome equals 1. Thus the probabilities in column 2 of Table 1 sum to one.
3. $Pr\left(C=1\,or\,C=2\right)=Pr\left(C=1\right)+Pr\left(C=2\right)$ as long as the two outcomes are mutually exclusive. The outcomes in this example are mututally exclusive because if an apartment has one bedroom it does not have two bedrooms.

The list of the probabilities of every outcome, shown in the second column of Table 1, is called the **probability distribution (pdf)**. The list of the probabilities of each outcome *or fewer*, shown in the third column, is the **cumulative probability distribution**. Thus the probability of visiting an apartment with one childe under 6 *or fewer* is
$Pr\left(C=0\,or\,C=1\right)=Pr\left(C=0\right)+Pr\left(C=1\right)=0.883+0.086=0.969.$

Graphing the Probability Distribution of a Discrete Variable

Figure 2: Probability Distribution of Children under 6 in NYC 2016



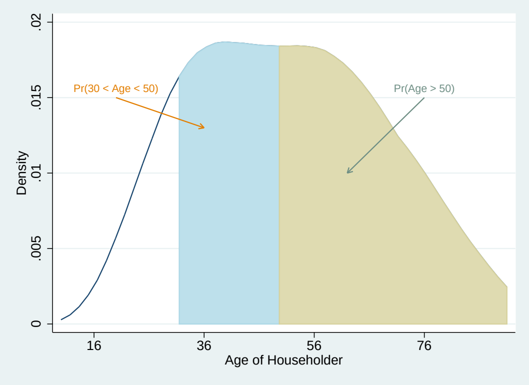
The height of each bar shows the *probability* that a particular value occurs. For example, the third bar shows that the probability of randomly picking an apartment with one child is 8.6%.

Probability Distribution of a Continuous Random Variable

- A continuous random variable has an uncountable number of values, so it's not possible to list them next to their probabilities. In fact, the probability that a continuous variable equals a spcecific value is zero. For example, what's the probability that someone is exactly $35.43829....$ years old? That's one number out of an uncountably large number of possibilities.
- We can, however, calculate the *cumulative* probability distribution of a continuous random variable. It is the probability of obtaining a value *equal to or less than* a particular number. For example, we can calculate the probability that a New York City householder is 40 years old or less.
- We can also calculate the probability that a random value falls within a range. For example, the probability of picking a householder between 30 and 50 in New York City is about 47%.

Graphing the Probability Density of a Continuous Variable

Figure 3: Probability Density Function of Householder Age in NYC 2016



Unlike a discrete variable, the height of the graph of a continuous variable is not the probability but the **probability *density* function** or **pdf**. The *areas* under the curve show probabilities.The total area under the pdf equals 1.

Expected Value

- The **expected value** of a variable is its long-run average over many occurrences. It is the **weighted average** of all possible values, with the weights being the probabilities of each value. For example, the expected value of the number of children is

$$\begin{aligned} E(C) &= Pr(C=0)\cdot 0 + Pr(C=1)\cdot 1 \\ &+ Pr(C=2)\cdot 2 + Pr(C=3)\cdot 3 \\ &= 0.883\cdot 0 + 0.086\cdot 1 + 0.027\cdot 2 + 0.004\cdot 3 \\ &= 0.15 \end{aligned}$$

- It is generally *not* equal to the **simple average**,

$$\frac{0+1+2+3}{4} = \frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{4}\cdot 4 = 1.5$$

because that assumes that the probability of each value is $\frac{1}{4}$.

Expected Value in Excel

The expected value can be easily calculated in Excel, as shown in Table 2. The word “Bedrooms” is in cell A1.

Table 2: Expected Value in Excel

Children (C)	Probability Distribution	$Pr(C) \cdot C$
0	0.883	=A2*B2
1	0.086	=A3*B3
2	0.027	=A4*B4
3	0.004	=A5*B5
		=SUM(C2:C5)

Variance and Standard Deviation

- The **variance** of a variable X is defined as

$$V(X) = E(X - E(X))^2$$

  The variance is generally symbolized by σ^2, which is the Greek letter sigma raised to the second power, called "sigma-squared". The expected value is symbolized by μ, the Greek letter that is pronounced in the U.S. as "myoo". The inside term, $(X - E(X))$, is called the **deviation** of X or "X-centered", written X_c. Using these notations,

$$\sigma_X^2 = E(X_c)^2$$

- The definition gives the recipe to calculate variance.
  1. Calculate $E(X) = \mu_X$.
  2. Calculate $X_c^2 = (X - \mu_X)^2$ for each value of X.
  3. Calculated the weighed average of X_c^2 using the probabilities as weights.

Variance and Standard Deviation

- As an example, let's calculate the variance of the number of children under 6. We've already calculated that $\mu_C = 0.15$. Then

$$\begin{aligned}\sigma_C^2 &= Pr\left(C=0\right)\cdot\left(0-\mu_C\right)^2 + Pr\left(C=1\right)\cdot\left(1-\mu_C\right)^2 \\ &+ Pr\left(C=2\right)\cdot\left(2-\mu_C\right)^2 + Pr\left(C=3\right)\cdot\left(3-\mu_C\right)^2 \\ &= 0.883\cdot\left(0-0.15\right)^2 + 0.086\cdot\left(1-0.15\right)^2 \\ &+ 0.027\cdot\left(2-0.015\right)^2 + 0.004\cdot\left(3-0.15\right) \\ &= 0.21\end{aligned}$$

- The **standard deviation**, symbolized by σ, is the square-root of the variance, σ^2. Here, $\sigma_C = 0.45$ bedrooms. This means that if we pick a dwelling at random in New York City, the number of children under 6 will on average be 0.45 children either above or below the mean of 0.15.

Variance and Standard Deviation in Excel

The variance can easily be calculated using Excel. The dollar signs are used to create an absolute reference to a cell. This allows one to copy a formula across cells without changing one or more cells to which the formula refers. In this calculation, one needs to subtract the mean of B, calculated in cell C9, from each value.

Table 3: Calculating Variance in Excel

Children (C)	$Pr\left(C\right)$	$Pr\left(C\right) \cdot C$	$Pr\left(C\right) \cdot \left(C - \mu_C\right)^2$
0	0.883	=A2*B2	=(A2-C9)^2*B2
1	0.086	=A3*B3	=(A3-C9)^2*B3
2	0.027	=A4*B4	=(A4-C9)^2*B4
3	0.004	=A5*B5	=(A5-C9)^2*B5
		=SUM(C2:C5)	=SUM(D2:D5)

Mean and Variance: a Special Case

Table 4: Expected Value and Variance: a Special Case

	Y	$P(Y)$	$Y - E(Y) = Y_c$	Y_c^2
	2	$\frac{1}{4}$	-4	16
	4	$\frac{1}{4}$	-2	4
	6	$\frac{1}{4}$	0	0
	12	$\frac{1}{4}$	6	36
$E(\bullet)$	6		0	14

Suppose you observe the whole population, as above. Then each element has the *same probability*, and the weighted equals the unweighted average.

$$E(Y) = \frac{1}{4}2 + \frac{1}{4}4 + \frac{1}{4}6 + \frac{1}{4}12 = \frac{2+4+6+12}{4} = 6$$

Similarly, the variance is

$$V(Y) = \frac{(2-6)^2 + (4-6)^2 + (6-6)^2 + (12-6)^2}{4} = 14$$

Mean of a Constant Times a Random Variable

Suppose you earn \$1000 during odd weeks and \$3000 during even weeks. What is your expected earnings, Y? Since half of the weeks are even and half are odd,

$$P(Y = 1000) = P(Y = 3000) = \frac{1}{2}$$

and expected earnings are

$$E(Y) = \frac{1}{2} \cdot 1000 + \frac{1}{2} \cdot 3000 = 2000$$

Now suppose your earnings double each week. What is your new expected earnings? That is, what is $E(2Y)$?

$$E(2Y) = \frac{1}{2} \cdot 2 \cdot 1000 + \frac{1}{2} \cdot 2 \cdot 3000$$

We can factor out the 2s.

$$E(2Y) = 2 \cdot \left(\frac{1}{2} \cdot 1000 + \frac{1}{2} \cdot 3000\right) = 2 \cdot E(Y)$$

Thus in general, if a is a constant and X is a variable, $E(aX) = a \cdot E(X)$

Variance of a Constant Times a Random Variable

Suppose your pre-tax income is Y. You pay a 20% tax, so your post-tax income is $I = 0.8 \cdot Y$. If the variance of your pre-tax income is σ_Y^2, what is the variance of your post-tax income, σ_I^2? Recall the definition of variance.

$$\sigma_I^2 = E\left(I - \mu_I\right)^2$$

Substitute $0.8 \cdot Y$ for I.

$$\sigma_I^2 = E\left(0.8 \cdot Y - 0.8 \cdot \mu_Y\right)^2$$

Factor out the 0.8.

$$\begin{aligned} \sigma_I^2 &= E\left(0.8 \cdot \left(Y - \mu_Y\right)\right)^2 \\ &= E\left(0.8^2 \left(Y - \mu_Y\right)^2\right) \\ &= 0.8^2 \cdot E\left(Y - \mu_Y\right)^2 = 0.8^2 \cdot V\left(Y\right) \end{aligned}$$

Thus in general, if a is a constant and X is a variable,

$$V\left(aX\right) = a^2 V\left(X\right)$$

Joint and Marginal Distributions

Table 5: Education (E) of Husbands and Wives

$\Downarrow$ Husband / Wife $\Rightarrow$	$College_w$= 0	$College_w$= 1	Marginal
$College_h$= 0	0.46	0.13	0.59
$College_h$= 1	0.09	0.32	0.41
Marginal	0.55	0.45	1.00

- A **dummy** or **dichotomous** variable equals one if a condition applies and zero otherwise. In the table above, the variable "$College_w$" equals 1 if the wife graduated from college and 0 if not. Similarly, $College_h$ equals 1 if the husband graduated from college, zero if not.
- The **joint probability distribution** of two variables is the probability that the variables simultaneously take on specific values. For example, the probability that a wife graduated from college *and* the husband also graduated from college is 32%. Symbollically, $Pr\left(C_w = 1, C_h = 1\right) = 0.32$. It is the intersection of two events, $C_w = 1$ and $C_h = 1$.

Joint and Marginal Distributions

- In this example there are four joint probabilities. The probability that
  1. neither graduates from college: $Pr\left(C_w = 0, C_h = 0\right) = 0.46$
  2. wife graduates, husband does not: $Pr\left(C_w = 1, C_h = 0\right) = 0.13$
  3. husband graduates, Wife does not: $Pr\left(C_w = 0, C_h = 1\right) = 0.09$
  4. both graduate: $Pr\left(C_w = 1, C_h = 1\right) = 0.32$.
- These outcomes are mutually exclusive and account for all the possibilities. Therefore they constitute the sample space and their probabilities sum to 1.
- The **marginal** or **unconditional distribution** is simply the probabity distribution of a single variable. The marginal distribution of husbands' education is shown in the fourth column, and that of wives in the fourth row. For example, the marginal or unconditional probability that a wife graduated from college is 45%. Note that the marginal probabilities are the sums of the joint probabilities. For example, $Pr\left(C_w = 1\right) = Pr\left(C_w = 1, C_h = 0\right) + Pr\left(C_w = 1, C_h = 1\right)$.

Conditional Distributions

- Suppose we know that a wife graduated from college. Given this, what's the probability that her husband graduated from college? This is a **conditional probability**. It is the probability that a husband graduated from college *conditional* on the wife's having graduated from college, and is written $Pr\left(C_h = 1 | C_w = 1\right)$, where the vertical bar, $|$, means "conditional on" or "given that".
- The formula for calculating conditional probabilities can be understood using Figure 1. Let event A be that the wife is a college graduate, and event B, that the husband is a college graduate. Since we know that A has happened, the sample space is not the whole rectange, but just the A-circle. The probability of B given A is the proportion of the circle A that includes the circle B. In other words, we want the intersection of A and B as a proportion of A. Denote the joint probability of A and B as $Pr\left(A, B\right)$. Then

$$Pr\left(B|A\right) = \frac{Pr\left(A, B\right)}{Pr\left(A\right)}$$

Conditional Probability, Examples

- What is the probability that a husband has a college degree given that his wife has a college degree?

  $$Pr\left(C_h = 1 | C_w = 1\right) = \frac{Pr\left(C_w = 1, C_h = 1\right)}{Pr\left(C_w = 1\right)} = \frac{0.32}{0.45} = 0.71$$

  Thus among women with a college degree, 71% of the husbands have college degrees.

- What is the probability that a husband does not have a college degree, given that the wife has a college degree?

  $$Pr\left(C_h = 0 | C_w = 1\right) = \frac{Pr\left(C_w = 1, C_h = 0\right)}{Pr\left(C_w = 1\right)} = \frac{0.13}{0.45} = 0.29$$

  Thus among women with a college degree, 29% of the husbands lack college degrees.

Conditional Expectations

Table 6: Doctor Visits in Past Two Weeks and Health Insurance

	Number of Visits			
Coverage	0	1	2	Marginal
Covered = 0	0.12	0.01	0.00	0.13
Covered = 1	0.72	0.11	0.04	0.87
Marginal	0.84	0.12	0.04	1.00

- Table 6 shows the number of doctor visits in the past two weeks among people 25 and over stratified by health insurance coverage.
- Recall that the expected value of a variable is the weighted average of its possible values, using the probability distribution as weights. For example, the expected value of the number of doctor visits (D) in the past two weeks is

$$\begin{aligned} E\left(D\right) &= Pr\left(D=0\right)\cdot 0 + Pr\left(D=1\right)\cdot 1 + Pr\left(D=2\right)\cdot 2 \\ &= 0.84\cdot 0 + 0.12\cdot 1 + 0.04\cdot 2 = 0.2 \end{aligned}$$

Conditional Expectations

- The **conditional expected value** of a variable is its average *among a subgroup of the population*. It is the weighted average of possible values using *conditional* probabilities as weights. For example, the average number of doctor visits (D) in the past two weeks *among people with health insurance* ($C=1$) is

$$
\begin{aligned}
E\left(D|C=1\right) &= Pr\left(D=0|C=1\right)\cdot 0\\
&\quad + Pr\left(D=1|C=1\right)\cdot 1\\
&\quad + Pr\left(D=2|C=1\right)\cdot 2\\
&=\frac{Pr\left(D=0,C=1\right)}{Pr\left(C=1\right)}\cdot 0\\
&+\frac{Pr\left(D=1,C=1\right)}{Pr\left(C=1\right)}\cdot 1\\
&+\frac{Pr\left(D=2,C=1\right)}{Pr\left(C=1\right)}\cdot 2\\
&=0.82\cdot 0+0.13\cdot 1+0.05\cdot 2=0.22
\end{aligned}
$$

Mean Independence

- If the overall or unconditional expected value of a variable, Y, equals its expected value for each subgroup defined by another variable, X, it is said that Y is **mean-independent** of X. Symbollically, this is true when

$$E(Y) = E(Y|X = x)$$

for every possible value of X.

- For example, the overall or unconditional expected value of doctor visits is 0.2 visits every two weeks. That is, $E(D) = 0.2$. The average number of doctor visits *among people with health insurance* is 0.22, or 10% higher. That is, $E(D|C = 1) = 0.22$. Since

$$E(D|C = 1) = 0.22 \neq E(D) = 0.2$$

we can conclude that the variable “doctor visits in past two weeks” is **mean-dependent** on insurance.

Conditional Variance

- Recall the definition of variance.

$$V(X) = E(X - \mu_x)^2$$

Let's apply this to doctor visits (D). We know from above that $E(D) = 0.2$.

$$V(D) = 0.84 \cdot (0 - 0.2)^2 + 0.12 \cdot (1 - 0.2)^2 + 0.04 \cdot (2 - 0.2)^2 = 0.24$$

- The **conditional variance** formula simply replaces the unconditional mean with the conditional mean, and the unconditional probabilities with the conditional probabilities. Let's calculate $V(D|C=1)$ using $E(D|C=1)$ and $Pr(D|C=1)$ from above.

$$\begin{aligned} V(D|C=1) = & Pr(D=0|C=1) \cdot (0-0.2)^2 \\ & + Pr(D=1|C=1) \cdot (1-0.2)^2 \\ & + Pr(D=2|C=1) \cdot (2-0.2)^2 \\ = & 0.82 \cdot 0.047 + 0.13 \cdot 0.075 + 0.05 \cdot 0.155 = 0.28 \end{aligned}$$

Independence

- Let X and Y be random variables, and x and y be specific values of the variables. Two variables are **independent** if their unconditional and conditional probabilities are the same. Symbollically,

$$Pr\left(X=x|Y=y\right)=P\left(X=x\right)$$

  For example, we found that the probability of the husband having a college degree is 41% overall but 71% if his wife has a college degree. Thus wives' and husbands' education are **dependent**.

- In contrast, suppose we flip a fair coin. Let event A be getting heads on the first toss, and event B, getting a heads on the second toss. $Pr\left(A\right)=0.5$. What is the probability of heads *given* you just got a heads, $Pr\left(B|A\right)$? Still 50%. Thus

$$Pr\left(B\right)=Pr\left(B|A\right)$$

  Therefore the different results of a coin toss are **independent**.

Independence, example

Recall the formula for the conditional probability of event A given event B.

$$Pr\left(A|B\right) = \frac{Pr\left(A,B\right)}{Pr\left(B\right)}$$

Multiply both sides by $Pr\left(B\right)$.

$$Pr\left(A,B\right) = Pr\left(B\right)Pr\left(A|B\right)$$

If A and B are *independent* events,

$$Pr\left(A|B\right) = Pr\left(A\right)$$

Therefore, for *independent* events,

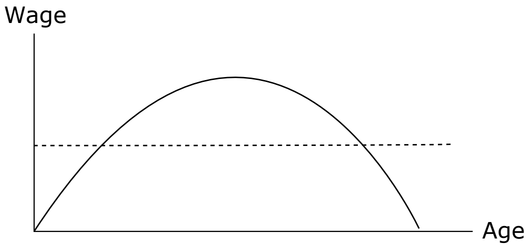
$$Pr\left(A,B\right) = Pr\left(A\right)\cdot Pr\left(B\right)$$

For example, the probability of getting two heads in a row in two tosses of a fair coin is

$$Pr\left(H\right)\cdot Pr\left(H\right) = 0.5\cdot 0.5 = 0.25$$

Covariance

Figure 4: Age and Wage



If the variables Y and X are independent, there is no relationship between them whatsoever. If they are mean-independent, there is no relationship between X and the *average* of Y at different values of X. **Covariance** tells us whether there is a *linear* relationship between X and Y. In Figure 4 there is clearly a non-linear relationship between age and wage, but there is no *linear* relationship, as shown by the flat dotted line. Thus the covariance between the two variables is zero.

Covariance

- Covariance between Y and X, symbolized by σ_{XY}, is defined as

$$\sigma_{XY} \equiv E\left(X-\mu_X\right)\left(Y-\mu_Y\right) = E\left(X_c Y_c\right).$$

  Like variance, this definition contains step-by-step directions.

  1. Calculate $E\left(X\right)$ and $E\left(Y\right)$.
  2. Calculate X_c and Y_c.
  3. Calculate the average of $X_c \cdot Y_c$, using the joint probabilities as weights.

- The covariance between husbands' and wives' college is

$$\begin{aligned}
E\left(C_w\right) &= Pr\left(C_w=0\right)\cdot 0 + Pr\left(C_w=1\right)\cdot 1 = 0.45\\
E\left(C_h\right) &= Pr\left(C_h=0\right)\cdot 0 + Pr\left(C_h=1\right)\cdot 1 = 0.41\\
Cov\left(C_w, C_h\right) &= Pr\left(C_w=0, C_h=0\right)\cdot\left(0-\mu_w\right)\cdot\left(0-\mu_h\right)\\
&+ Pr\left(C_w=0, C_h=1\right)\cdot\left(0-\mu_w\right)\cdot\left(1-\mu_h\right)\\
&+ Pr\left(C_w=1, C_h=0\right)\cdot\left(1-\mu_w\right)\cdot\left(0-\mu_h\right)\\
&+ Pr\left(C_w=1, C_h=1\right)\cdot\left(1-\mu_w\right)\cdot\left(1-\mu_h\right)\\
&= 0.46\cdot 0.45\cdot 0.41 + 0.09\cdot\left(-0.45\right)\cdot 0.59\\
&+ 0.13\cdot 0.55\cdot\left(-0.59\right) + 0.32\cdot 0.55\cdot 0.59 = 0.12
\end{aligned}$$

Correlation

- The units of measurement of the covariance of Y and X are the units of Y times the units of X. Thus the covariance of husbands' and wives' college attainment is 12 percentage-points-squared. That's hard to interpret. Also, is 12 big or small? That's also hard to judge. All the covariance really tells us is whether the linear relationship between Y and X is positive or negative. The calculation above tells us that husbands' and wives' college attainment is positively related.
- The covariance is easier to interpret when it is divided by the standard deviations of both variables. This produces the **correlation**, defined as

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- Correlation is a unit-free number, because the units in the numerator (units-of-X *times* units-of-Y) cancel out the units in the denominator (also units-of-X *times* units-of-Y).
- Correlation is bounded by -1 and 1. If $\rho_{XY} = 1$, X and Y are perfectly correlated. If $\rho_{XY} = -1$, X and Y are perfectly negatively correlated.

Correlation, example

- The variance of C_w is

$$\sigma_w^2 = 0.55 \cdot (0 - 0.45)^2 + 0.45 \cdot (1 - 0.45)^2 = 0.2475$$

and of C_h is

$$\sigma_h^2 = 0.59 \cdot (0 - 0.41)^2 + 0.41 \cdot (1 - 0.41)^2 = 0.2419$$

Therefore

$$\rho_{w,h} = \frac{\sigma_{w,h}}{\sigma_w \sigma_h} = \frac{0.12}{\sqrt{0.2475}\sqrt{0.2419}} = 0.49$$

- There are no logical rules deciding when the correlation is low or high. Nonetheless, statisticians would say that a correlation of 0.49 is a moderately strong positive correlation.

Covariance and Correlation: A Special Case

Table 7: Covariance and Correlation with an Entire Population

	Y	X	$P(Y,X)$	$Y - E(Y) = Y_c$	Y_c^2	X_c	Y_cX_c	X_c^2
	2	2	$\frac{1}{4}$	-4	16	-7	28	49
	4	10	$\frac{1}{4}$	-2	4	1	-2	1
	6	6	$\frac{1}{4}$	0	0	-3	0	9
	12	18	$\frac{1}{4}$	6	36	9	54	81
$E(\bullet)$	6	9		0	14	0	20	35

Given the whole population, the weighted equals the unweighted averages.

$$\sigma_Y^2 = \sum Y_c^2/n = 14$$

$$\sigma_X^2 = \sum X_c^2/n = 35$$

$$\sigma_{YX} = \sum X_cY_c/n = 20$$

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = \frac{20}{\sqrt{14}\sqrt{35}} = 0.90$$

Mean of Sums of Random Variables

- If a and b are constants and X and Y are variables,

  $$E\left(aX + bY\right) = aE\left(X\right) + bE\left(Y\right)$$

  For example, denote the income in \$1000s of husbands, Y_h, and of wives, Y_w. Suppose $E\left(Y_h\right) = \$75$ and $E\left(Y_w\right) = \$37$, what is the expected *couple*'s income, $E\left(Y_h + Y_w\right)$? We can use the rule above by letting $a = b = 1$.

  $$E\left(Y_h + Y_w\right) = E\left(Y_h\right) + E\left(Y_w\right) = \$112$$

- The average *personal* income is $\frac{Y_w + Y_h}{2}$. Its expectation is

  $$\begin{aligned} E\left(\frac{Y_w + Y_h}{2}\right) &= E\left(\frac{1}{2}Y_w + \frac{1}{2}Y_h\right) \\ &= E\left(\frac{1}{2}Y_w\right) + E\left(\frac{1}{2}Y_h\right) \\ &= \frac{1}{2}E\left(Y_w\right) + \frac{1}{2}E\left(Y_h\right) = 56 \end{aligned}$$

Variance of Sums of Random Variables

- Suppose the standard deviation of wives' incomes is $\sigma_w = \$54$, of husbands', $\sigma_h = \$88$, and their correlation, $\rho_{w,h} = 0.14$, what is the variance of *couples*' income $V(Y_w + Y_h)$? The general rule is

$$V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab\sigma_X\sigma_Y\rho_{XY}$$

In this example, $a = b = 1$. Then

$$V(Y_w + Y_h) = 54^2 + 88^2 + 2 \cdot 54 \cdot 88 \cdot 0.14 = 11991$$

- What is the variance of the average *personal* income, $\frac{Y_w+Y_h}{2}$?

$$\begin{aligned}
V\left(\frac{Y_w + Y_h}{2}\right) &= V\left(\frac{1}{2}Y_w + \frac{1}{2}Y_h\right) \\
&= V\left(\frac{1}{2}Y_w\right) + V\left(\frac{1}{2}Y_h\right) + 2 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \sigma_w \cdot \sigma_h \cdot \rho_{w,h} \\
&= \left(\frac{1}{2}\right)^2 \sigma_w^2 + \left(\frac{1}{2}\right)^2 \sigma_h^2 + 2 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \sigma_w \cdot \sigma_h \cdot \rho_{w,h} \\
&= 0.25 \cdot 54^2 + 0.25 \cdot 88^2 + \frac{1}{2} \cdot 54 \cdot 88 \cdot 0.14 = 2997.64
\end{aligned}$$

Mean and Variance of a Stock Portfolio

Over the past 40 years the average annual rate of return of Apple stock (A) was 0.25 with a standard deviation of 0.50, and of Wells Fargo was 0.18 with a standard deviation of 0.2. The correlation of their rates of return was -0.3. Suppose we create a portfolio of equal dollar amounts of the two stocks. The expected rate of return on the portfolio (P) is

$$E\left(P\right)=E\left(\frac{1}{2}A+\frac{1}{2}W\right)=E\left(\frac{1}{2}A\right)+E\left(\frac{1}{2}W\right)=\frac{1}{2}0.25+\frac{1}{2}0.18=0.215$$

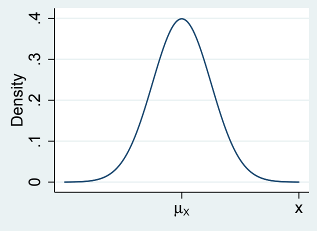
The variance of the portfolio is

$$\begin{aligned} V\left(P\right) &= V\left(\frac{1}{2}A+\frac{1}{2}W\right) \\ &= \left(\frac{1}{2}\right)^2\sigma_A^2+\left(\frac{1}{2}\right)^2\sigma_W^2+2\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\sigma_A\cdot\sigma_W\cdot\rho_{AW} \\ &= 0.25\cdot 0.5^2+0.25\cdot 0.2^2+\frac{1}{2}\cdot 0.5\cdot 0.2\cdot(-0.3)=0.0575 \end{aligned}$$

The standard deviation of the portfolio is $\sqrt{0.0575}=0.24$.

Normal Distribution

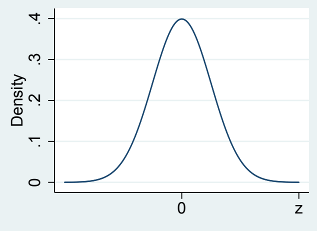
Figure 5: The Normal Distribution



- A Normal distribution is described by exactly two numbers: its mean, $E(X)$ or μ_X, and its variance, $V(X)$ or σ_X^2.
- The notation $X \sim N(\mu_X, \sigma_X^2)$ means "X is distributed as a Normal variable, with mean μ_X and variance σ_X^2."
- The Normal density is symmetric about its mean. If $\mu_X = 5$, then $P(X > 7) = P(X < 3)$ because both 3 and 7 are both 2 from μ_X.

Standard Normal Distribution

Figure 6: Standard Normal Distribution

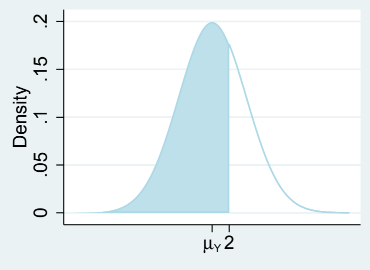


- An important special case of the Normal is the Standard Normal, denoted Z. $Z \sim N(0,1)$. That is, Z is normally-distributed with a mean of zero and a variance of 1.
- Any Normal variable can be converted to a Z by **standardizing** it: subtract the mean and divide by the standard deviation.

$$Z = \frac{X - \mu_X}{\sigma_X}$$

Normal Probabilities

Figure 7: Cumulative Normal Probability



The **Cumulative Standard Normal** distribution is represented by the Greek letter Φ; that is, $P\left(Z \leq c\right) = \Phi\left(c\right)$, where c is a constant. To find probabilities for a normal variable, we must standardize the variable by first subtracting the mean, then dividing the result by the standard deviation. Then plug the standardized value into the cumulative normal function in Excel, Stata, R, or other statistical program.

Normal Probabilities

- For example, suppose $Y \sim N(1,4)$. What is $Pr(Y \leq 2)$? It is the shaded area in Figure 7. To calculate this, first standardize Y, then plug the number into Excel, "=normsdist(0.5)".

$$P(Y \leq 2) = P\left(\frac{Y - \mu_Y}{\sigma_Y} < \frac{2-1}{2}\right) = P(Z \leq 0.5) = \Phi(0.5) = 0.6914$$

- One variable that is roughly normally distributed in the population is height (H). Among U.S. men, approximately $H \sim N\left(69.1, 3.1^2\right)$ measured in inches. What is $P(H \geq 6'4")$?

$$\begin{aligned} P(H \geq 76) &= P\left(Z \geq \frac{76 - 69.1}{3.1}\right) \\ &= P(Z \geq 2.26) \end{aligned}$$

Remember, $\Phi(z)$ gives $P(Z \leq z)$. We want $P(Z \geq z)$ or $1 - \Phi(z)$. Because Z is symmetrical around the mean, $1 - \Phi(z) = \Phi(-z)$.

$$P(Z \geq 2.26) = 1 - \Phi(2.26) = \Phi(-2.26) = normsdist(-2.26) = 0.012$$

Normal Probability of a Range

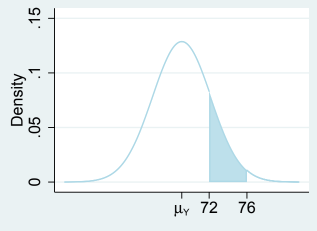


Figure 8: Probability of a Range

Let's find the percent of men whose heights are *between* 72 and 76 inches.

$$\begin{aligned} P\left(72 \le H \le 76\right) &= P\left(H \le 76\right) - P\left(H \le 72\right) \\ &= P\left(Z < \frac{76 - 69.1}{3.1}\right) - P\left(Z < \frac{72 - 69.1}{3.1}\right) \\ &= \Phi\left(2.26\right) - \Phi\left(0.93\right) \\ &= normsdist\left(2.26\right) - normsdist\left(0.93\right) = 0.164 \end{aligned}$$

Sums of Normal Variables

- Sums and Differences of Normal variables are also Normal. For example, suppose $X \sim N(16, 877)$, $Y \sim N(36, 381)$, and $\rho_{XY} = 0.17$. Thus $E(X) = 16$, $V(X) = 877$, $E(Y) = 36$, and $V(Y) = 381$. Then $X + Y$ and $X - Y$ are also Normal.
- Let $W = 5X + 5Y$. What is $E(W)$? Recall

$$E(aX + bY) = aE(X) + bE(Y)$$

Then

$$E(W) = 5 \cdot E(X) + 5 \cdot E(Y) = 5 \cdot 16 + 5 \cdot 36 = 260$$

- What is $V(W)$? Recall

$$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2 \cdot a \cdot b \cdot \sigma_X \cdot \sigma_Y \cdot \rho_{XY}$$

Then

$$V(W) = 5^2 \cdot 877 + 5^2 \cdot 381 + 2 \cdot 5 \cdot 5 \cdot \sqrt{877} \cdot \sqrt{381} \cdot 0.17 = 36363.39$$

Student t Distribution

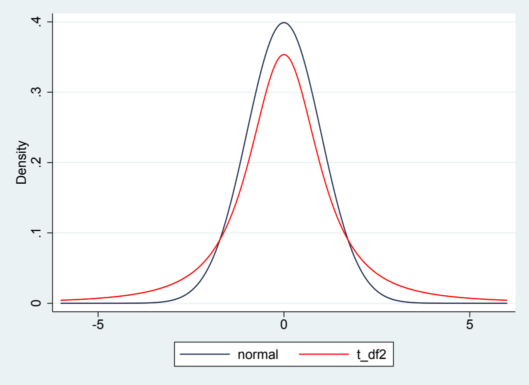
- In calculating the Z-value we use σ, the true, population standard deviation. If we do not know σ we must use s, the sample standard deviation. But when we use s we are not calculating a Z but a t-statistic.

$$t = \frac{X - \mu_X}{s_X}.$$

- Probabilities associated with the t-statistic depend on **degrees of freedom**, which equals the sample size minus the number of parameters estimated. For example, when we calculate the sample variance of X, we first must calculate the sample mean of X or $\overline{X}$, which is an estimate of the single parameter μ_X. Therefore the degrees of fredom is $n - 1$. We use t-statistics rather than Z statistics when the following are true.
  1. We know that X is normally distributed, and
  2. We use an *estimated* standard deviation.

Student t Distribution

Figure 9: Normal vs. t



When the degrees of freedom is small, the t-distribution has much fatter tails than the Normal. But as the degrees of freedom increases, the t-distribution and the Normal distribution become indistinguishable.

Student t Distribution

Suppose $X \sim N$ but we don't know its mean and variance. We do, however, have a sample of 7 observations, with a *sample variance* of 4, $s^2 = 4$. If $E(X) = 3$, what is $P(X \leq 2)$? Since we only have the sample variance, we have to use a t-distribution, in the following steps.

1. Convert X to a t

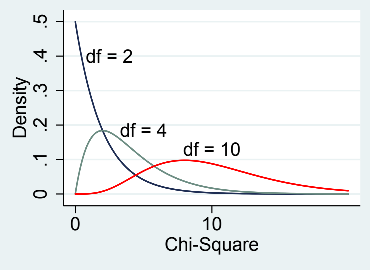
$$\begin{aligned} t &= \frac{X - \mu_X}{s} \\ &= \frac{2-3}{2} \\ &= -1 \end{aligned}$$

2. Calculate the probability, $P(t < -1)$, assuming the degrees of freedom is $n - 1 = 6$. In Excel, this is

$$\begin{aligned} &= t.dist(-1, 6, TRUE) \\ &= 0.178 \end{aligned}$$

Chi-squared Distribution

Figure 10: Chi-square Density



A chi-square variable with n degrees of freedom is defined as the sum of n squared independent standard normal variables.

$$\chi_n^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

Chi-squared Distribution

- As the number of degrees of freedom increases the shape of a chi-square density approaches that of a Normal, with the difference that the chi-square itself can never be negative.
- Chi-squares are used to describe the distribution of sums of squares. An example is the variance, which is a sum of squared deviations, divided by the degrees of freedom. Since a chi-square is a sum of squared terms, it is always positive.
- We will be interested in the probabilities in the right tail of the distribution. To find this, one needs the degrees of freedom. For example, suppose we want to find the probability that a chi-square with 5 degrees of freedom is greater than 3. In Excel, the formula is =CHISQ.DIST.RT(3,5). In Stata, chi2tail(5,3) = 0.7.

The F-distribution

- The t distribution is used to test a *single* hypothesis about a sample average when the standard deviation is estimated. When the goal is to test *multiple* hypotheses using estimated standard deviations, the F distribution is used.
- Formally,

  $$F_{n,m} = \frac{\chi^2_n/n}{\chi^2_m/m}$$

  written $F\left(n, m\right)$. It is the ratio of two chi-squares. Hence, like the chi-square, it cannot be negative.
- To look up probabilities of variable distributed as an F, one needs both the numerator degrees of freedom, n, and the denominator degrees of freedom, m. For example, what is the probability that an $F_{20,2} > 4$? In Excel, it is =F.DIST.RT(4,20,2) = 0.21.

Random Sampling

Table 8: Sample Dataset

Observation Number	Variable (Y)
1	Y_1
2	Y_2
$\vdots$	$\vdots$
n	Y_n

All the data we ever get come as a **sample** from some larger **population**. We shall assume that our data come from a **simple random sample** (**SRS**), in which every member of the population has the same chance of being selected. Each element in the sample is called an **observation**. The first observation on a variable Y is denoted Y_1, the second is Y_2, etc., up to Y_n, where n is the sample size. For example, a data set might look like that in Table 6.

Random Sampling

- Since the units are picked randomly, the first observation in one sample is unlikely to be the first observation in the next sample. Thus Y_1, Y_2, etc. are all random variables.
- We shall also assume that the distribution of Y depends on certain **population parameters**, which are fixed features of the population and so do not change from sample to sample. For example, the distribution of heights depends on the population mean, μ_H, and variance, σ_H^2. The mean and variance of the *population* do not change if we merely collect a new *sample.*
- Because Y_1, Y_2, etc. are picked randomly, they are independent of one another and reflect the same underlying distribution. They reflect the same population mean, the same population variance, and any other feature of the population. They are said to be **independently and identically distributed** or **i.i.d.**

Estimates and Estimators

- Our goal is to make an **estimate** of the population parameters using the sample. The formula used to obtain an estimate is an **estimator**. For example, suppose we take a random sample from the population. Let Y_1 be Y-value of the first person, Y_2 be the Y value of the second person, and so on. An estimator of $E\left(Y\right)$ or μ_Y is the **sample mean**, denoted $\overline{Y}$ or "Y-bar".
- An estimator of the **sample variance** is

$$s_Y^2 = \frac{\left(Y_1 - \overline{Y}\right)^2 + \left(Y_2 - \overline{Y}\right)^2 + \cdots + \left(Y_n - \overline{Y}\right)^2}{n-1}$$

- The division by $n-1$ rather than n is called a **degrees-of-freedom correction**. It is necessary because instead of μ_Y we are using $\overline{Y}$, which produces values of Y_c^2 that are too small.
- An estimator of the sample covariance is

$$s_{YX} = \frac{\left(Y_1 - \overline{Y}\right)\left(X_1 - \overline{X}\right) + \cdots + \left(Y_n - \overline{Y}\right)\left(X_n - \overline{X}\right)}{n-1}$$

Estimators, example

- Let's calculate the sample mean, sample variance, sample covariance in a small sample.

Table 9: Sample mean, variance, and covariance

Obs	Y	X	X_c	Y_c	Y_c^2	X_c^2	$X_c \cdot Y_c$
1	2	1	-3	-1	9	1	3
2	3	6	2	0	4	0	0
3	4	5	1	1	1	1	1
Σ	9	12	0	0	14	2	4

- The sample variance of Y is

$$s_Y^2 = \frac{Y_{1c}^2 + Y_{2c}^2 + Y_{3c}^2}{n-1} = \frac{14}{2} = 7$$

- The sample covariance of Y and X is

$$s_{XY} = \frac{Y_{1c}X_{1c} + Y_{21c}X_{2c} + Y_{3c}X_{3c}}{n-1} = \frac{4}{2} = 2$$

Sampling Distributions

- If we collect a new sample, we will get new values of Y_1 and Y_2 and therefore new values of $\overline{Y}$, s_Y^2, and s_{XY}. Thus $\overline{Y}$, s_Y^2, and s_{XY} are *variables* and have their own probability distributions, which are called the **sampling distributions** of $\overline{Y}$, s_Y^2, and s_{XY}.
- Two important features of the sampling distribution of an estimator such as $\overline{Y}$ are its mean and variance. As demonstrated below,

$$E\left(\overline{Y}\right) = \mu_Y$$
$$V\left(\overline{Y}\right) = \frac{V(Y)}{n}$$

- Recall the rules of expectation. If X and Y are two random variables, then

$$E(X+Y) = E(X) + E(Y)$$

Expected value of $\overline{Y}$

- Also, if X is a variable and a is a constant,

$$E\left(aX\right) = aE\left(X\right)$$

- Therefore

$$E\left(\overline{Y}\right) = E\left(\frac{1}{2}\left(Y_1 + Y_2\right)\right) = \frac{1}{2}E\left(Y_1 + Y_2\right) = \frac{1}{2}\left(E\left(Y_1\right) + E\left(Y_2\right)\right)$$

Recall that we pick all observations from the *same* population. Thus they all have the same expected value, μ_Y.

$$E\left(\overline{Y}\right) = \frac{1}{2}\left(\mu_Y + \mu_Y\right) = \frac{1}{2}2\mu_Y = \mu_Y$$

Whenever $E\left(estimator\right) = true\, parameter$ the estimator is **unbiased**. Thus $\overline{Y}$ is an unbiased estimator of μ_Y.

Variance of $\overline{Y}$

Recall

$$V\left(aX+bY\right)=a^2\sigma_X^2+b^2\sigma_Y^2+2ab\sigma_X\sigma_Y\rho_{XY}$$

Since

$$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

we can replace $\sigma_X\sigma_Y\rho_{XY}$ with σ_{XY}.

$$V\left(aX+bY\right)=a^2\sigma_X^2+b^2\sigma_Y^2+2ab\sigma_{XY}$$

We can apply this formula to find the variance of $\overline{Y}$. If $n=2$,

$$\begin{aligned}\overline{Y}\equiv\frac{Y_1+Y_2}{2}&=\frac{1}{2}Y_1+\frac{1}{2}Y_2\\ V\left(\overline{Y}\right)=\left(\frac{1}{2}\right)^2V\left(Y_1\right)&+\left(\frac{1}{2}\right)^2V\left(Y_2\right)\\ &+2\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\sigma_{XY}\end{aligned}$$

Variance of $\overline{Y}$ example

Table 10: Variance of the Mean

Observation number	Sample 1	Sample 2	Sample variance
1	1	2	$s^2_{Y_1} = 0.5$
2	5	0	$s_{Y_2} = 12.5$
$\overline{Y}$	3	1	$s^2_{\overline{Y}} = 2$
$V\left(Y\right)$	8	2	$s_{Y_1,Y_2} = -2.5$

Note that there are two $\overline{Y}$s, $\overline{Y}_1 = 3$ and $\overline{Y}_2 = 1$ and their average is $\overline{\overline{Y}} = 2$. Therefore,

$$V\left(\overline{Y}\right) = \frac{1}{1}\left(3-2\right)^2 + \frac{1}{1}\left(1-2\right)^2 = 2$$

Using the formula gives the same answer.

$$V\left(\overline{Y}\right) = \left(\frac{1}{2}\right)^2 \cdot 0.5 + \left(\frac{1}{2}\right)^2 \cdot 12.5 - 2 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot 2.5 = 2$$

Variance of $\overline{Y}$ in a Simple Random Sample

- We will often use a simpler version of the formula. Since this is a simple random sample from a given population, Y_1 and Y_2 have the same variance, $V(Y)$. This equality is called **homoskedasticity**. The observations are also independent of each other, implying $Cov(Y_1, Y_2) = 0$. Therefore

$$
\begin{aligned}
V\left(\overline{Y}\right) &= \left(\frac{1}{2}\right)^2 V(Y) + \left(\frac{1}{2}\right)^2 V(Y) \\
&= \left(\frac{1}{2}\right)^2 \cdot 2 \cdot V(Y) = \frac{V(Y)}{2}
\end{aligned}
$$

  Recall that $n = 2$. Therefore, more generally,

$$
Var\left(\overline{Y}\right) = \frac{V(Y)}{n}
$$

- Suppose $X \sim N(10, 192)$ and we have a sample of 12 observations. Thus $E(X) = 10$, $V(X) = 192$, and $n = 12$. Then $V\left(\bar{X}\right) = 192/12 = 16$ and $\sigma_{\bar{X}} = \sqrt{16} = 4$.

Consistency

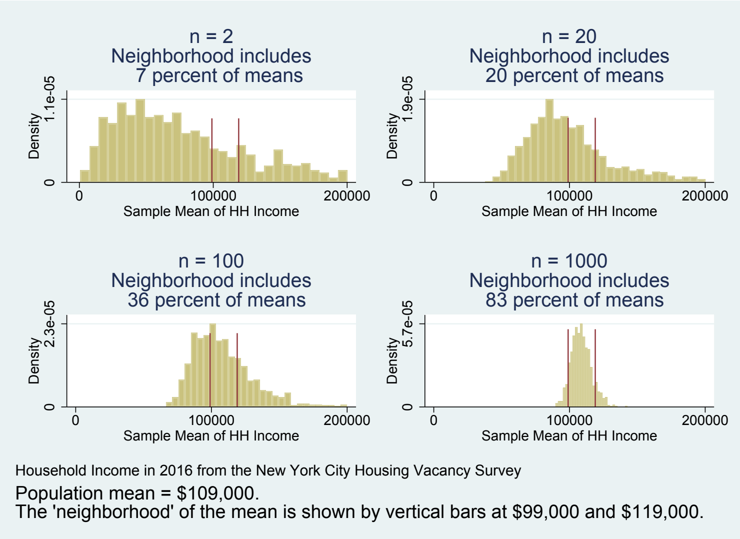
- Asymptotic properties of estimators are those that hold only in large samples. Two such properties are especially important in econometrics, **consistency** and **asymptotic normality**.
- An estimator is **consistent** if it becomes closer and closer to the population parameter as the sample size increases. More precisely, it is consistent when
  1. its variance shrinks to zero as the sample size increases, and
  2. its value equals the population parameter when the sample size is infinitely large.
- The sample mean, $\overline{Y}$, is consistent for μ because, first, its variance shrinks to zero as the sample size increases, and second, it is unbiased. As the sample size increases, the distribution of $\overline{Y}$ collapses to a single value, the population mean, μ. The fact that $\overline{Y}$ becomes μ as the sample size increases to infinity is called the **law of large numbers** or **LLN**.

Consistency and Asymptotic Normality

- The strict definition of consistency is illustrated in the next figure. First, pick a neighborhood around the population mean. The neighborhood chosen in the figure is \$10,000 above and below the population mean. The population mean household income in New York City in 2016 was \$109,000. (The median, or middle, income was \$68,000. The mean is higher than the median because a few housholds have incomes in the millions.)
- Second, note that as the sample size increases, the variance or spread of the sample means decreases. Eventually, all of the sample means will be inside the neighborhood of the mean, regardless of the neighborhood chosen. This is **consistency**.
- Also note that the shape of the distribution becomes more and more Normal as the sample size increases. This is called **asymptotical normality**. It is illustrated more clearly in the following figure, where the sample mean is standardized. The statement that the mean is asymptotically normal is the **central limit theorem**.

Consistency

Figure 11: Distribution of the Sample Mean as Sample Size Increases



n = 2
Neighborhood includes
7 percent of means

n = 20
Neighborhood includes
20 percent of means

n = 100
Neighborhood includes
36 percent of means

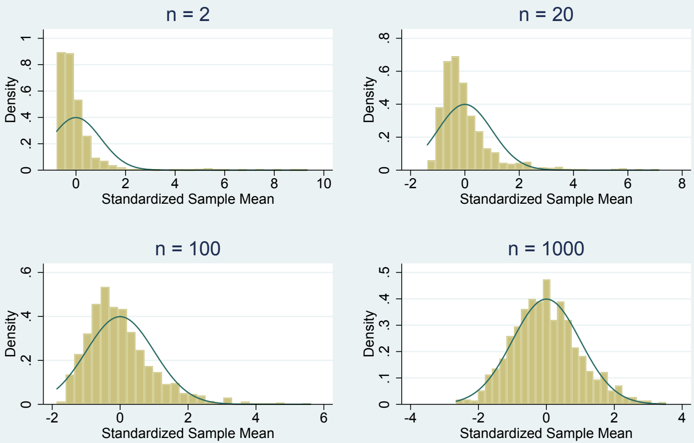
n = 1000
Neighborhood includes
83 percent of means

Household Income in 2016 from the New York City Housing Vacancy Survey

Population mean = $109,000.
The 'neighborhood' of the mean is shown by vertical bars at $99,000 and $119,000.

Asymptotic Normality

Figure 12: Distribution of the Standardized Sample Mean as Sample Size Increases



Household Income in 2016 from the New York City Housing Vacancy Survey

Standardized Mean = (Sample Mean - Mean of All Samples)/Stdev of Sample Mean

Estimators and their Properties

- Recall that an **estimator** is a formula to calculate an estimate from a *sample* of a *population* parameter.
- An esimator is unbiased if $E\left(estimator\right) = population\, parameter.$
- $\overline{Y}$ is an unbiased estimator of μ_Y because $E\left(\overline{Y}\right) = \mu_Y$.
- Another estimator of μ_Y is simply the first observation in each sample, Y_1. Remember that Y_1 is random because every new sample has a new Y_1. Also, we assume every sample comes from the *same* population. Therefore, $E\left(Y_1\right) = \mu_Y$ and Y_1 is also unbiased.
- But the *variances* of $\overline{Y}$ and Y_1 are different.

$$V\left(\overline{Y}\right) \equiv \sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{n}$$
$$V\left(Y_1\right) \equiv \sigma^2_Y$$

  Because $V\left(\overline{Y}\right)$ shrinks to zero as n grows to infinity, $\overline{Y}$ is consistent. Because $V\left(Y_1\right)$ *does not* shrink to zero as n grows to infinity, it is **inconsistent**, even though it is unbiased.
- Because $V\left(\overline{Y}\right) < V\left(Y_1\right)$, $\overline{Y}$ is called the more **efficient** estimator.

Null and Alternative Hypotheses

- Suppose we want to test the hypothesis that the average hourly earnings of a college graduate between 25 and 29 years old in New York City is $23. The hypothesis that we actually test is called the **null hypothesis**. It is written H_0, and the hypothesized value of the population parameter is the **null value**, written $\mu_{Y,0}$, assuming the parameter is the population mean. In this example, $\mu_{Y,0} = 23$.

$$H_0 : \mu_Y = 23$$

- The **alternative hypothesis**, written H_A, tells us what is true if the null hypothesis is false. A **two-sided alternative hypothesis** is true if the population parameter is *not equal* to the null value, regardless of whether the population value is above *or* below the null value.

$$H_A : \mu_Y \neq 23$$

- A **one-sided alternative hypothesis** is true if the population parameter is either above or below the null value but not both, e.g.

$$H_A : \mu_Y > 23$$

Hypothesis Testing

- Hypotheses are always about *population parameters*, because we never know them. We never test hypotheses about *sample values*. For example, the following is wrong.

  $$H_0 : \overline{Y} = 23$$

  This hypothesizes about the sample mean, $\overline{Y}$. *But we **know** the sample mean*. We never hypothesize about what we know. What we don't know is the *population* mean, μ_Y.
- The goal of the researcher is to use data from *sample* to test a hypothesis about the *population*. Data from a random sample are always uncertain, so we can never know for sure if a hypothesis is false, but we can tell if it is *unlikely*.
- If the sample data conflict with the null hypothesis, we **reject the null.** If not, we **fail to reject**. We never *accept* the null hypothesis.

Absolute Value

Absolute value is defined as

$$|c| = \begin{cases} c & if \quad c > 0 \\ 0 & if \quad c = 0 \\ -c & if \quad c < 0 \end{cases}$$

Suppose $a > 0$ and $|c| < a$.

1. Suppose $c > 0$. Then clearly $c = |c| < a$. Also $-c < a$ because $-c < c = |c| < a$. For example, let $c = 2$ and $a = 3$. Clearly, $c < a$ because $2 < 3$. Also, $-c < a$ because $-2 < 3$.
2. Now suppose $c < 0$. Then $c < a$ because $a > 0$. Also, $-c < a$ because $-c = |c| < a$. For example, let $c = -2$ and $a = 3$. Clearly, $c < a$ because $-2 < 3$. Also $-c < a$ because $-(-2) < 3$.

Thus if $a > 0$ and $|c| < a$, both $c < a$ and $-c < a$. Multiply $-c < a$ by -1 to get $c > -a$. Therefore

$$-a < c < a$$

P-values

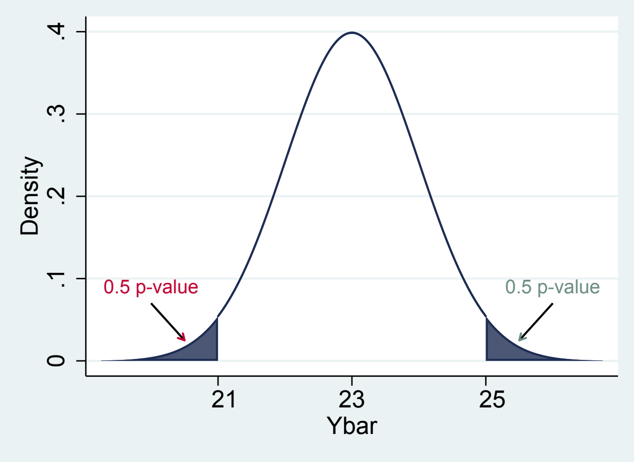
- Suppose we collect wage data on a random sample of college graduates between 25 and 29 in New York City, and that the average wage in the actual sample we draw, $\overline{Y}^{actual}$, is $25. Should we reject H_0 that the average wage in the population, $\mu_{Y,0}$, is $23?
- Unclear. It's *possible* to draw a random $\overline{Y}$ of $25 even though μ_Y is $23. The question is, how likely is it? This is given by the **p-value.** For a two-sided test, the p-value is the probability of drawing a statistic *at least as far* from the null value as the one drawn, assuming the null hypothesis is true.
- "Far" is a measure of distance, and distance is always positive. Thus we need to use *absolute values*.

$$distance = \left| \overline{Y}^{actual} - \mu_{Y,0} \right| = |25 - 23| = 2$$

  Thus the p-value is the probability of drawing a random $\overline{Y}$ that is at least $2 above $\mu_{Y,0}$ *or* at least $2 below $\mu_{Y,0}$, that is, *below* $21 *or above* $25, since both $21 and $25 are $2 away from the null value, $23. This shown in the following figure.

P-values

Figure 13: P-values



P-values

- Formally, the p-value, as shown in the shaded areas in Figure 13, equals the probability that a random $\overline{Y}$ is further from $\mu_{Y,0}$ than the actual sample average, $\overline{Y}^{actual}$, that we draw.

$$\begin{aligned} pvalue &= P\left(\left|\overline{Y}-\mu_{Y,0}\right| > \left|\overline{Y}^{actual}-\mu_{Y,0}\right|\right) \\ &= P\left(\left|\overline{Y}-\mu_{Y,0}\right| > 2\right) \\ &= P\left(\overline{Y} < 21 \, or \, \overline{Y} > 25\right) \\ &= P\left(\overline{Y} < 21\right) + P\left(\overline{Y} > 25\right) \end{aligned}$$

- Note also that

$$P\left(\overline{Y} < 21\right) = P\left(\overline{Y} > 25\right)$$

  because 21 and 25 are the same distance from the mean, 23. Thus we only need the simpler calculation

$$pvalue = 2 \cdot P\left(\overline{Y} < 21\right)$$

Calculating P-values

- To calcuate the p-value we need to know the distribution of $\overline{Y}$ if H_0 is true. If the sample is large enough (conventionally, over 30 observations) then we can assume that $\overline{Y}$ is Normally distributed. We then need to know its mean (that is, the expected value) and variance.
- We get $E\left(\overline{Y}\right)$ from the null hypothesis. In this example, $\mu_{Y,0} = 23$.
- We then need the standard deviation of $\overline{Y}$. There are two possibilities.

1. If we know the true, population variance of Y, we calculate the p-value using a Z-statistic.

$$p\left(\overline{Y} < 21\right) = P\left(\frac{\overline{Y} - \mu_{\overline{Y}}}{\sigma_{\overline{Y}}} < \frac{21 - 23}{\sigma_{\overline{Y}}}\right) = P\left(Z < \frac{21 - 23}{\sigma_{\overline{Y}}}\right)$$

2. If we know only the *sample estimate* of the variance of Y, we calculate the p-value using a t-statistic.

$$p\left(\overline{Y} < 21\right) = P\left(\frac{\overline{Y} - \mu_{\overline{Y}}}{s_{\overline{Y}}} < \frac{21 - 23}{s_{\overline{Y}}}\right) = P\left(t < \frac{21 - 23}{s_{\overline{Y}}}\right)$$

Calculating P-values

- We rarely know true population variance. In this example we only know the estimated sample variance of Y and the sample size.

$$s_Y^2 = 390$$
$$n = 400$$

Recall the variance of the sample mean.

$$s_{\overline{Y}}^2 = \frac{s_Y^2}{n} = \frac{390}{400} = 0.975$$

- Therefore the p-value is

$$pvalue = 2 \cdot P\left(\overline{Y} < 21\right) = 2 \cdot P\left(t < \frac{21 - 23}{\sqrt{0.975}}\right)$$
$$= 2 \cdot t.dist\left(-2.025, 399\right) = 0.044$$

The value -2.025 is called the **sample t-statistic**, symbolized t_{sample}. The general Excel function to find the p-value is

$$pvalue = 2 \cdot t.dist\left(-\left|t_{sample}\right|, d.f., true\right)$$

Reject or Fail to Reject

- Thus the probability of randomly picking a sample whose average is at least \$2 away from the hypothesized value of \$23 is 4.4%.
- Does this mean that picking such a sample is so unlikely if the null is true that we should reject the null?
- We make this decision by comparing the p-value with a preselected threshold called the **significance level**, denoted α. If $pvalue < \alpha$, we conclude that obtaining a $\overline{Y}$ as far from $\mu_{Y,0}$ as we did is so unlikely if H_0 is true that we must **reject the null hypothesis**. If $pvalue \geq \alpha$, we **fail to reject the null hypothesis**.
- Suppose we set $\alpha = 0.05$. Since $0.044 < 0.05$, we reject the null. If we had chosen $\alpha = 0.01$, we would fail to reject the null.
- We never know if we've made the right decision. If the null is true and we reject it, we make a **type I error**. If the null is false and we fail to reject it, we make a **type II error**.
- Suppose the null is true. If we set $\alpha = 0.05$ and reject the null whenever $pvalue < \alpha$, it means we will wrongly reject the null 5% of the time. Thus α is the probability of a type I error.

P-values, another example

- Using the same data as above, test the hypothesis that $\mu_{Y,0} = 24$. Assume $\alpha = 0.05$. As before, $\overline{Y} = 25$ and $\sigma_{\overline{Y}} = \sqrt{0.975}$.
- The hypotheses are

$$\begin{aligned} H_0 &: \mu_Y = 24 \\ H_A &: \mu_Y \neq 24 \end{aligned}$$

- Note that $\overline{Y}$ is $1 away from the null value. Therefore the p-value is

$$\begin{aligned} pvalue &= P\left(\overline{Y} < 23\right) + P\left(\overline{Y} > 25\right) \\ &= 2 \cdot P\left(\overline{Y} < 23\right) \\ &= 2 \cdot P\left(t < \frac{23 - 24}{\sqrt{0.975}}\right) = 2 \cdot t.dist\left(-1.013, 399, true\right) = 0.31 \end{aligned}$$

- Since $pvalue = 0.31 > \alpha = 0.05$, we fail to reject the null.

Hypothesis Testing with a Prespecified Significance Level

- Recall that when $|t_{sample}| = 2.025$, the p-value in our two-sided test is 0.044, and when $|t_{sample}| = 1.013$, the p-value in our two-sided test is 0.31. So somewhere between 2.025 and 1.031 there must be t-value whose p-value in a two-sided test is exactly 0.05, our chosen α. That t-value is called the **critical t-value**, denoted t_α. If $|t_{sample}| > t_\alpha, pvalue < \alpha$ and we reject. If $|t_{sample}| \leq t_\alpha, pvalue \geq \alpha$, and we fail to reject.
- We can find t_α in Excel with

$$t_\alpha = t.inv.2t(\alpha, d.f.) = t.inv.2t(0.05, 399) = 1.97$$

  where d.f. means "degrees of freedom".
- Thus a simple procedure to test a hypothesis is
  1. Choose H_0, H_A, and α.
  2. Calculate t_{sample} and t_α
  3. If $|t_{sample}| > |t_\alpha|$, reject. If $|t_{sample}| \leq |t_\alpha|$, fail to reject.

Hypothesis Testing with a Prespecified α, example

- Suppose we test a new hypothesis about the average wage of college graduates in New York City.

$$H_0 : \mu_Y = 24$$
$$H_A : \mu_Y \neq 24$$

Also, let's set $\alpha = 0.05$

- In a sample of 400 college graduates we find $\overline{Y} = 25$ and $\sigma^2_Y = 390$. First, calculate $\sigma^2_{\overline{Y}}$.

$$s^2_{\overline{Y}} = \frac{s^2_Y}{n} = \frac{390}{400} = 0.975$$

- Now calculate the sample t-statistic.

$$t_{sample} = \frac{\overline{Y} - \mu_{Y,0}}{\sigma^2_{\overline{Y}}} = \frac{25 - 24}{\sqrt{0.975}} = 1.013$$

- The critical t-value is 1.97. Since $|1.013| < 1.97$, we fail to reject.

Confidence Intervals

- Recall that we fail to reject when $|t_{sample}| < t_\alpha$. This implies

$$-t_\alpha < t_{sample} < t_a$$

- The probability of t_{sample} falling in this range when d.f. $=$ 399 is

$$prob\,(-1.97 < t_{sample} < 1.97) = 0.95$$

- We know

$$t_{sample} = \frac{\overline{Y} - \mu_{Y,0}}{s_{\overline{Y}}}$$

- Plug this into the above equation, then solve to get μ_Y by itself in the center.

$$prob\left(-1.97 < \frac{\overline{Y} - \mu_{Y,0}}{s_{\overline{Y}}} < 1.97\right) = 0.95$$

Confidence Intervals

- First, multiply both inequalities by $\sigma_{\overline{Y}}$.

$$prob\left(-1.97s_{\overline{Y}} < \overline{Y} - \mu_{Y,0} < 1.97s_{\overline{Y}}\right) = 0.95$$

- Then subtract $\overline{Y}$.

$$prob\left(-1.97s_{\overline{Y}} - \overline{Y} < -\mu_{Y,0} < -\overline{Y} + 1.97s_{\overline{Y}}\right) = 0.95$$

- Then multiply by -1. Recall that multiplying an inequality by -1 reverses the inequality signs.

$$prob\left(\overline{Y} + 1.97s_{\overline{Y}} > \mu_{Y,0} > \overline{Y} - 1.97s_{\overline{Y}}\right) = 0.95$$

Confidence Intervals

- $\overline{Y} - 1.97s_{\overline{Y}}$ is the lower bound of all values of $\mu_{Y,0}$ we would not reject, and $\overline{Y} + 1.97s_{\overline{Y}}$ is the upper bound. The convention is to put the lower bound on the left and the upper bound on the right

$$prob\left(\overline{Y} - 1.97s_{\overline{Y}} < \mu_{Y,0} < \overline{Y} + 1.97s_{\overline{Y}}\right) = 0.95$$

- Because $E\left(\overline{Y}\right) = \mu_Y$ we can replace $\mu_{\overline{Y}}$ with μ_Y.

$$prob\left(\overline{Y} - 1.97s_{\overline{Y}} < \mu_{Y,0} < \overline{Y} + 1.97s_{\overline{Y}}\right) = 0.95$$

- This interval includes all the value of $\mu_{Y,0}$ that *would not* be rejected in a two-sided hypothesis test.
- The interval is called a **95% percent confidence interval (CI)** for μ_Y. More generally, it is a $(1 - \alpha) \cdot 100\%$ confidence interval. This says that if we draw 100 samples and calculate the lower bound and upper bound according to the formulas above, they will bracket the true population mean 95 times.

Confidence Intervals, example

- Let's use the data from the sample above to calculate a 95% confidence interval for the average wage of college graduates 25 to 29 in New York City.

$$\begin{aligned} prob\left(\overline{Y} - 1.97 s_{\overline{Y}} < \mu_Y < \overline{Y} + 1.97 s_{\overline{Y}}\right) &= 0.95 \\ prob\left(25 - 1.97 \cdot \sqrt{0.975} < \mu_Y < 25 + 1.97 \cdot \sqrt{0.975}\right) &= 0.95 \\ prob\left(23.1 < \mu_Y < 26.9\right) &= 0.95 \end{aligned}$$

- Let's calculate a 90% confidence interval for μ_Y. Then $\alpha = 0.1$ and $t_\alpha = i.inv.2t\left(0.1, 399\right) = 1.65$. The confidence interval is

$$\begin{aligned} prob\left(\overline{Y} - 1.65 s_{\overline{Y}} < \mu_Y < \overline{Y} + 1.65 s_{\overline{Y}}\right) &= 0.95 \\ prob\left(25 - 1.65 \cdot \sqrt{0.975} < \mu_Y < 25 + 1.65 \cdot \sqrt{0.975}\right) &= 0.95 \\ prob\left(23.4 < \mu_Y < 26.6\right) &= 0.95 \end{aligned}$$

This interval is narrower because we are *less certain* (90% versus 95%) that it includes the true population parameter.

Comparing Means of Different Populations

- Suppose we want to compare the wages of male and female college graduates between 25 and 29 in New York City. Let μ_w be the population mean of women's wages, and μ_m be the population mean of men's wages. Our null hypothesis is that the difference between them is zero. Set $\alpha = 0.05$.

$$H_0 : \mu_m - \mu_w = 0$$
$$H_A : \mu_m - \mu_w \neq 0$$

- A sample of 218 women gives $\overline{Y}_w = \$23.3$ and $s_w^2 = 108$; an independent sample of 182 men gives $\overline{Y}_m = 27$ and $s_m^2 = 658$. Thus

$$s^2_{\overline{w}} = \frac{s_w^2}{n_w} = \frac{108}{218} = 0.5$$
$$s^2_{\overline{m}} = \frac{s_m^2}{n_m} = \frac{658}{182} = 3.6$$

Comparing Means of Different Populations

- The sample difference in wages is $\overline{Y}_m - \overline{Y}_w = 27 - 23.3 = 3.7$. We want to test whether this difference is significantly different from zero. For this we need the variance of the difference. Because the samples of men and women were independently collected, the correlation of their means is zero.
- Recall

$$\begin{aligned} V\left(aX + bY\right) &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_X\sigma_Y\rho_{XY} \\ V\left(\overline{Y}_m - \overline{Y}_w\right) &= 1^2 \cdot s_{\overline{m}}^2 + (-1)^2\, s_{\overline{w}}^2 + 2 \cdot 1 \cdot (-1) \cdot s_{\overline{m}} \cdot s_{\overline{w}} \cdot 0 \\ &= 0.5 + 3.6 = 4.1 \end{aligned}$$

- The sample t-statistic is

$$t_{sample} = \frac{\left(\overline{Y}_m - \overline{Y}_w\right) - \left(\mu_m - \mu_w\right)}{s_{\overline{Y}_m - \overline{Y}_w}} = \frac{3.7 - 0}{\sqrt{4.1}} = 1.83$$

- The p-value is $pvalue = 2 \cdot t.dist\left(-\left|t_{sample}\right|, d.f., true\right) = 2 \cdot t.dist\left(-1.83, 199, true\right) = 0.07$. Since $pvalue > \alpha$, we fail to reject.

The Linear Regression Model

- Suppose a college sophomore is considering whether to continue in school for two more years. She wants to know, if she *changes* her education by two years, what will be the *change* in her hourly wage?
- Denote the change in the hourly wage by $\Delta wage$ and the change in the number of years of education by $\Delta education$. Suppose that every additional year of education changes the wage by a fixed dollar amount. Call this amount $\beta_{education}$. Then

$$\Delta wage = \beta_{education} \cdot \Delta education$$

For example, suppose $\beta_{education} = 2$. Then two more years of college brings

$$\Delta wage = 2 \cdot 4 = \$4$$

Thus two more years of college raises her wage by about $4 per hour.

The Linear Regression Model

- The equation above can be rewritten

$$\frac{\Delta wage}{\Delta education} = \beta_{education}$$

- This equation is the definition of the **slope** of a straight line. The full equation of the line itself is

$$wage = \beta_0 + \beta_{education} \cdot education$$

  where β_0 is the y-**intercept**. β_0 shows how much a person makes if her education is zero.

- The equation says that if you know a person's education, you know her exact wage. This cannot be true. A person's wage reflects *many* other factors, such as age, experience, scholastic ability, and gender. Group these other factors in the variable, u. Thus the true equation is

$$wage = \beta_0 + \beta_{education} \cdot education + u$$

The Linear Regression Model

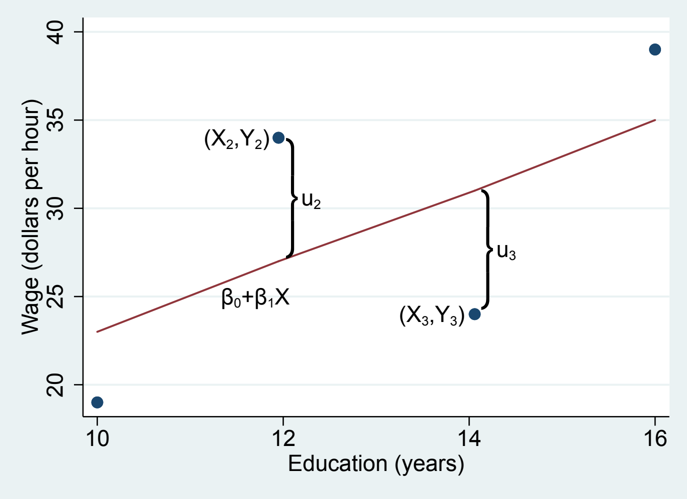
- In the model above, wage is the **dependent** variable, education is the **independent** variable, u is the **error term**, β_0 is the **intercept**, and $\beta_{education}$ is the **slope** as well as the **coefficient** on education.
- The dependent variable is also known as the **outcome**, the **regressand,** the **response variable**, the **explained variable**, the **predicted variable**, or the **left-hand-side variable**.
- The independent variable is also known as the **regressor**, the **explanatory variable**, the **predictor variable**, or the **right-hand-side variable.**
- More generally, the dependent variable is denoted Y and the independent variable, X.

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where Y_i is the value of the dependent variable of the i'th person, X_i is the value of the independent variable of the i'th person, and u_i is the collection of other factors that influence the wage of the i'th person. β_0 and β_1 are the **parameters** of the model.

The Population Regression Function and Errors

Figure 14: The Population Regression Function and Errors



The Population Regression Function and Errors

- The first part of the model

$$\beta_0 + \beta_1 X_i$$

  is the **population regression function (PRF)**. It is the straight line in the diagram above.

- The second part of the model, the error term,

$$u_i = Y_i - (\beta_0 + \beta_1 X_i)$$

  is the vertical distance between the PRF and the actual value of Y_i.

- Because of factors *other than education* that affect the wage, the second person's wage is above the value predicted by the PRF and therefore $u_2 > 0$. Similarly, factors other than education reduced the wage of the third person's wage below that predicted by the PRF and therefore $u_3 < 0$.
- Because we do not observe the u_is, we never know β_0 and β_1. But we can *estimate* them from a sample.

Predicted Values and Residuals

- Just as we estimated the population mean, μ_Y, using the sample mean, $\overline{Y}$, we can estimate the population parameters, β_0 and β_1 using sample data as well.
- Suppose we arbitrarily pick estimates of β_0 and β_1. Let's denote these estimates as $\hat{\beta}_0$ and $\hat{\beta}_1$. How do we know if they are "good" estimates?
- One criterion is how well they predict Y. Define the **predicted value** of Y, which we will call $\hat{Y}$ ("yhat"), as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_i$$

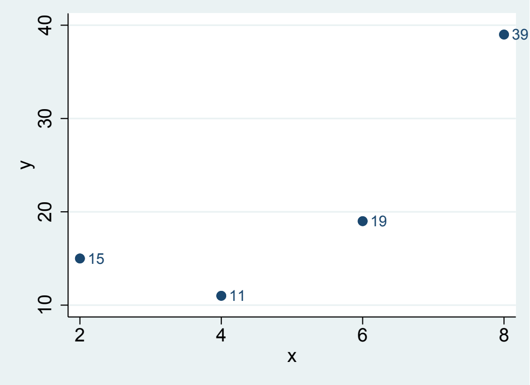
- Define the difference between the actual and predicted values of Y as $\hat{u}$ ("uhat"), called **residuals**.

$$\hat{u}_i = Y_i - \hat{Y}_i$$

- Note: residuals, $\hat{u}_i$, are *not* errors, u_i. Errors are Y_i minus the *true PRF*, which we never know. Residuals are Y_i minus *our estimate* of the PRF, called the **sample regression function (SRF)**.

Predicted Values and Residuals, example

Figure 15: Finding β_0 and β_1



For example, suppose we want to find the values of β_0 and β_1 that produced the points in Figure 15.

Predicted Values and Residuals, example

Let's arbitrarily pick the values $\hat{\beta}_0 = 6$ and $\hat{\beta}_1 = 3$. Then predicted values and residuals are:

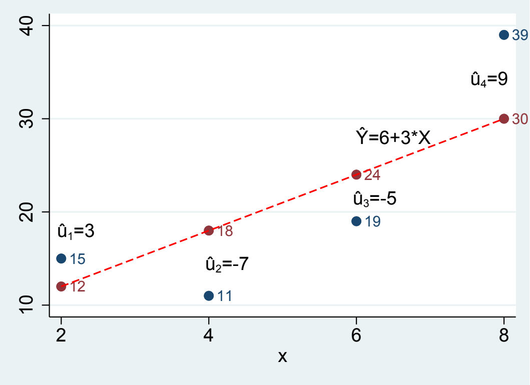
Table 11: Initial Predicted Values and Residuals

Y	X	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$	$\hat{u} = Y - \hat{Y}$
15	2	$\hat{\beta}_0 + \hat{\beta}_1 X_1 = 6 + 3 \cdot 2 = 12$	$Y_1 - \hat{Y}_1 = 3$
11	4	$\hat{\beta}_0 + \hat{\beta}_1 X_2 = 6 + 3 \cdot 4 = 18$	$Y_2 - \hat{Y}_2 = -7$
19	6	$\hat{\beta}_0 + \hat{\beta}_1 X_3 = 6 + 3 \cdot 6 = 24$	$Y_3 - \hat{Y}_3 = -5$
39	8	$\hat{\beta}_0 + \hat{\beta}_1 X_4 = 6 + 3 \cdot 8 = 30$	$Y_4 - \hat{Y}_4 = 9$

The predicted values and the residuals are graphed on the following slide.

Predicted Values and Residuals, example

Figure 16: Initial guesses of β_0 and β_1



The actual Y-values are shown in blue. The predicted values, $\hat{Y}$, are in red and connected by the dotted line, whose equation is $\hat{Y}_i = 6 + 3 \cdot X_i$. The residuals, $\hat{u}_i$, equal the *vertical distance* between the *actual* Y-values and the *predicted* Y-values. The $\hat{u}_i$s are shown between Y_i and $\hat{Y}_i$.

Predicted Values and Residuals, example

- Let's repeat this process with different values of $\hat{\beta}_0$ and $\hat{\beta}_1$. Suppose $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 4$. Then predicted values and residuals are:

Table 12: Initial Predicted Values and Residuals

Y	X	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$	$\hat{u} = Y - \hat{Y}$
15	2	$\hat{\beta}_0 + \hat{\beta}_1 X_1 = 1 + 4 \cdot 2 = 9$	$Y_1 - \hat{Y}_1 = 6$
11	4	$\hat{\beta}_0 + \hat{\beta}_1 X_2 = 1 + 4 \cdot 4 = 17$	$Y_2 - \hat{Y}_2 = -6$
19	6	$\hat{\beta}_0 + \hat{\beta}_1 X_3 = 1 + 4 \cdot 6 = 25$	$Y_3 - \hat{Y}_3 = -6$
39	8	$\hat{\beta}_0 + \hat{\beta}_1 X_4 = 1 + 4 \cdot 8 = 33$	$Y_4 - \hat{Y}_4 = 6$

- How do we know which set of predictions are better, those produced with $\hat{\beta}_0 = 6$ and $\hat{\beta}_1 = 3$ or those produced with $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 4$?
- "Good" estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ should preduce "good" predictions and therefore "small" residuals. But how do we measure the smallness of the residuals?

Estimating β_0 and β_1

- One way is simply to sum them. For example, the sum of the residuals produced when $\hat{\beta}_0 = 6$ and $\hat{\beta}_1 = 3$ is

  $$\sum \hat{u}_i = 3 - 7 - 5 + 9 = 0$$

  The sum of the residuals produced when $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 4$ is

  $$\sum \hat{u}_i = 6 - 6 - 6 + 6 = 0$$

  Thus summing the residuals does not tell us which predictions are better, because the positive and negative residuals cancel out in both cases.

- The solution used most often is to *square* the residuals, and pick esimates of β_0 and β_1 that minimize the **sum of squared residuals**, or SSR. That is find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

  $$SSR = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

Estimating β_0 and β_1

- When $\hat{\beta}_0 = 6$ and $\hat{\beta}_1 = 3$, $SSR = 3^2 + (-7)^2 + (-5)^2 + 9^2 = 164$. When $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 4$, $SSR = 6^2 + (-6)^2 + (-6)^2 + 6^2 = 144$, implying we should choose $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 4$.
- Because we want estimates of β_0 and β_1 that minimize the sum of squared residuals (SSR), the method is called **ordinary least squares (OLS)**.
- Using calculus, one can show that the formulas for β_0 and β_1 that minimize the SSR are

$$\hat{\beta}_1 = \frac{\sum_i Y_{ic} X_{ic}}{\sum_i X_{ic}^2} = \frac{\sum_i Y_{ic} X_{ic} / (n-1)}{\sum_i X_{ic}^2 / (n-1)} = \frac{cov(X,Y)}{var(X)}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- If these formulas are applied correctly, the following will *always* be true.
  1. $\sum_{i=1}^{n} \hat{u}_i = 0$
  2. $\sum_{i=1}^{n} \hat{u}_i X_{ci} = 0$

Estimating β_0 and β_1, example

Table 13: OLS Calculations

Y	X	u	X_c	Y_c	X_cY_c	X_c^2	Y_c^2	$\hat{Y}$	$\hat{u}$	$\hat{u}^2$
8	2	-2	-1	-6.5	6.5	1	42.25	11	-3	9
14	2	4	-1	-0.5	0.5	1	0.25	11	3	9
15	4	-3	1	0.5	0.5	1	0.25	18	-3	9
21	4	3	1	6.5	6.5	1	42.25	18	3	9
$\frac{\Sigma}{n}$ =14.5	3	0.5	0	0	3.5	1	21.25	14.5	0	9
Σ =58	12	2	0	0	14	4	85	58	0	36

Using the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ gives

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_{ic} X_{ci}}{\sum_{i=1}^{n} X_{ic} X_{ci}} = \frac{14}{4} = 3.5$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 14.5 - 3.5 \cdot 3 = 4$$

Estimating β_0 and β_1, example

- Using these values we can calculate the predicted values of Y,$\hat{Y}$.

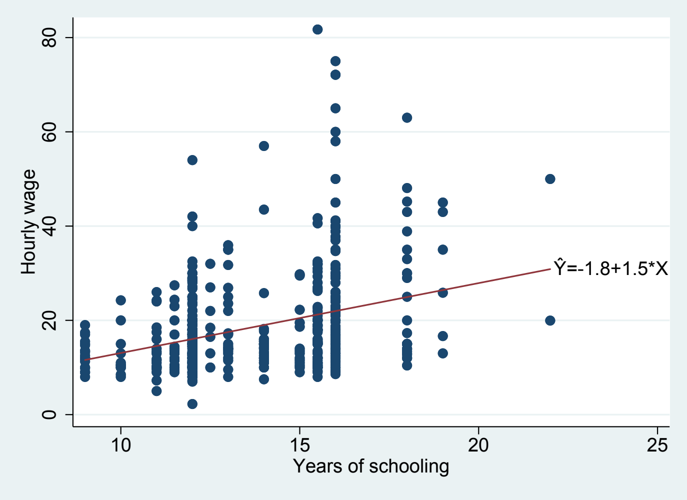
$\hat{Y}_i$	$=$	$\hat{\beta}_0 + \hat{\beta}_1 X_i$		
$\hat{Y}_1$	$=$	$4 + 3.5 \cdot 2$	$=$	11
$\hat{Y}_2$	$=$	$4 + 3.5 \cdot 2$	$=$	11
$\hat{Y}_3$	$=$	$4 + 3.5 \cdot 4$	$=$	18
$\hat{Y}_4$	$=$	$4 + 3.5 \cdot 4$	$=$	18

- Using the values of $\hat{Y}_i$ we can calculate $\hat{u}_i$, $\hat{u}_i^2$, and SSR. Note that $\sum \hat{u}_i = 0$ and $\sum \hat{u}_i X_{ci} = 0$.

$\hat{u}_i$	$=$	$Y_i - \hat{Y}_i$		$\hat{u}_i$	$\hat{u}_1^2$	$\hat{u}_i X_{ci}$
$\hat{u}_1$	$=$	$8 - 11$	$=$	-3	9	3
$\hat{u}_2$	$=$	$14 - 11$	$=$	3	9	-3
$\hat{u}_3$	$=$	$15 - 18$	$=$	-3	9	-3
$\hat{u}_4$	$=$	$21 - 18$	$=$	3	9	3
				$\Sigma = 0$	$\Sigma = 36 = SSR$	$\Sigma = 0$

OLS Example with Actual Data

Figure 17: Scatterplot of Wages and Education



OLS Example with Actual Data

- The scatterplot in Figure 17 shows a sample of 500 observations on the wages and education of New Yorkers between 25- and 54-years-old in 2017 and 2018. The sample regression function (SRF) is

$$\widehat{wage} = -1.8 + 1.5 \times education$$

- The coefficient on education, $\hat{\beta}_1 = 1.5$. This means that for each extra year of education, the predicted wage rises by $\$1.5$. For four years more of education, the predicted wage rises by $\$6$ per hour.
- The intercept, $\hat{\beta}_0 = -1.8$. To interpret this, suppose $education = 0$. Then

$$\widehat{wage} = -1.8 + 1.5 \times 0 = -1.8$$

This means that persons with zero years of education make $-\$1.8$ per hour. Obviously this is nonsense. The method of OLS can produce nonsensical intercepts if $X = 0$ for no observations. In this example, no people in the sample have zero years of education.

OLS Predictions

- The sample regression function enables us to make predictions. For example, what is the predicted wage for someone with 10 years of education?

$$\widehat{wage} = -1.8 + 1.5 \times 10 = \$13.2$$

- What is the predicted wage of someone with 16 years of education (16 years is a college degree)?

$$\widehat{wage} = -1.8 + 1.5 \times 16 = \$22.2$$

- What is the predicted wage of someone with 20 years of education?

$$\widehat{wage} = -1.8 + 1.5 \times 20 = \$28.2$$

- What is the predicted wage of someone with 50 years of education?

$$\widehat{wage} = -1.8 + 1.5 \times 50 = \$73.2$$

But this is unreliable because no one in the sample has anything like 50 years of education.

R^2

- The **total sum of squares** or TSS is $\sum_i Y_{ci}^2$.
- The **explained sum of squares** or ESS is $\sum_i \hat{Y}_{ci}^2$.
- It can be shown that

$$TSS = ESS + SSR$$

- Divide both sides by TSS.

$$1 = \frac{ESS}{TSS} + \frac{SSR}{TSS}$$

- R^2 is defined as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- R^2 is the fraction of the sample variation in Y explained by X. In the four-observation example above, $TSS = 85$, $SSR = 36$, therefore $R^2 = 36/85 = 0.42$, or X explains 42% of the variation in Y. In the wage-education example, $R^2 = 0.11$, meaning that education explains 11% of the variation in the wage.

R^2

- R^2 ranges from zero to one. If $\hat{\beta}_1 = 0$ then X explaines none of the variation in Y. Therefore $ESS = 0$ and $R^2 = 0$.
- If X explaines *all* of the variation in Y, then the predictions are perfect, meaning $\hat{Y}_i = Y$, $\hat{u}_i = 0$, $SSR = 0$, and $R^2 = 1$. If this happens, either the researcher made a mistake or the research question is uninteresting.
- A high R^2 does not mean the model is good, and a low R^2 does not mean the model is bad. Remember, we want to know the effect of education on earnings. If the model provides a credible answer to that question, it is a good model, regardless of the R^2.
- R^2 indicates how well X predicts Y *in this sample*. It says nothing about how well X predicts Y *in the population*. If $\hat{\beta}_0$ and $\hat{\beta}_1$ estimated from one sample does a good job of predicting Y *in another sample*, that is indeed evidence that these estimates constitute a good model.

The Standard Error of the Regression

- The **standard error of the regression** or **SER** is an estimate of the standard deviation of u_i, σ_u. It has the same units of measurement as the dependent variable. For example, if Y is in dollars per hour, σ_u is measured in dollars per hour.
- If we knew the true errors, the u_is, we would estimate σ_u^2 with

$$\sigma_u^2 = \frac{\sum \left(u_i - E\left(u_i\right)\right)^2}{n} = \frac{\sum \left(u_i - 0\right)^2}{n} = \frac{\sum u_i^2}{n}$$

- But we only have the residuals, which are only estimates of the errors. Since we have to estimate two parameters, β_0 and β_1, in order to calculate the residuals, we lose two degrees of freedom. Therefore we must divide SSR by $n-2$.

$$\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-2} = \frac{SSR}{n-2}$$

- In the four-observation example, $\hat{\sigma}_u^2 = 36/\left(4-2\right) = 18$. Therefore $SER = \sqrt{18} = 4.2$

OLS Assumption 1

The first OLS assumption is that at each value of X, the average error term is zero. Mathematically this is written

$$E\left(u_i|X\right)=0$$

To illustrate, let's calculate Y_i in a small population. We assume that the true values of the coefficients are $\beta_0 = 2$ and $\beta_1 = 4$. Given data on X_i and u_i, we can calculate the following values of Y.

Table 14: Small population

Y	X	u
8	2	-2
20	2	10
22	6	-4
46	6	20

OLS Assumption 1

- Let's check whether the first OLS assumption is true of the data in table 14. When $X_i = 2$, the average of u_i is

$$E\left(u_i | X_i = 2\right) = \frac{-2+10}{2} = 4 \neq 0$$

When $X_i = 6$, the average of u_i is

$$E\left(u_i | X_i = 4\right) = \frac{-4+20}{2} = 8 \neq 0$$

Since $E\left(u_i | X_i = x\right) \neq 0$ for at least one value of X_i, the first OLS assumption is false in this population.

- To understand why this matters, let's calculate the OLS estimates of β_0 and β_1, which turn out to be $\hat{\beta}_0 = 4$ and $\hat{\beta}_1 = 5$. Now let's plot both the true line (using $\beta_0 = 2$ and $\beta_1 = 4$) as well as the OLS line (using the values just given).

OLS Assumption 1

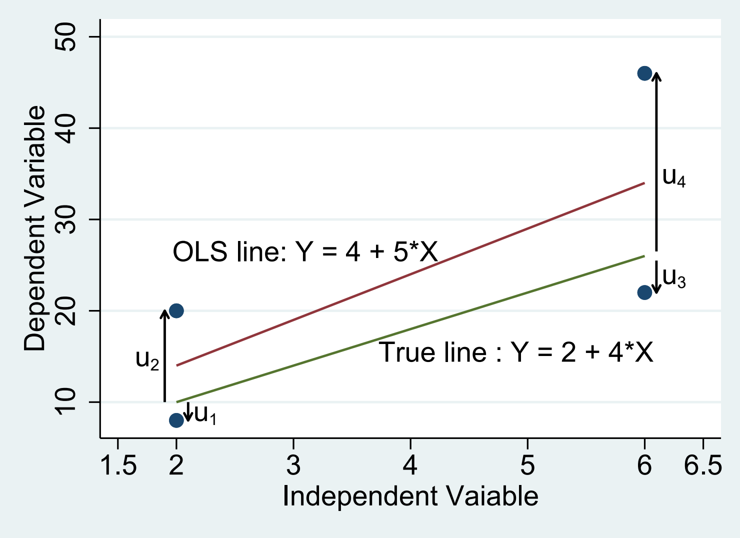


Figure 18: OLS versus true slope

OLS Assumption 1

- As shown in figure 18, the OLS line always runs through the center of the Y values. This is true because the OLS line minimizes the squared deviations from the Y_i values.
- But the *true* line *does not* always run through the center of the Y_i values. The true line runs through the center of the Y_i values *only* when u_i averages zero at each value of X_i. In the example above, this is not true.
- Whenever the $E\left(u_i|X_i\right) \neq 0$, the OLS estimate of β_0 is biased. In the example above, the average error is positive. (For example, $E\left(u_i|X_i = 2\right) = 4$). As a result, $\hat{\beta}_0 > \beta_0$.
- Whenever $E\left(u_i|X_i\right)$ varies with X, the OLS estimate of β_1 is biased. In the example above, $E\left(u_i|X_i\right)$ increases with X, so $\hat{\beta}_1 > \beta_0$.

Interpreting OLS Assumption 1

- Recall that u is mean-indepent from X if $E\left(u|X\right)=E\left(u\right)$. Suppose that Y is the wage, u is ability and X is education. Then the first OLS assumption means that $E\left(ability|Education=college\right)=E\left(ability\right)$, that is, the average ability among college students equals the average ability in the overall population.
- Suppose Y is the crime rate in the county, X is the number of policepersons per capita in the county, and u is county income per capita. Then $E\left(income|high\,police/capita\right)=E\left(income\right)$ means that the average income per capita in areas with a high number of police per capita is the same as the average income in the overall population.
- Suppose Y is the lifespan in years, X cigarettes smoked per day, and u is exercise minutes per day. Then $E\left(exercise|heavy\,smoker\right)=E\left(exercise\right)$ means that the average exercise minutes among heavy smokers equals the average exercise minutes in the overall population.

Interpreting β_0 and β_1 under OLS assumption 1

- Recall the basic model.

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Also recall the rules of expectation.

$$E\left(aX + bY\right) = aE\left(X\right) + bE\left(Y\right)$$

  where a, b are constants and X, Y are variables.
- Take expectations (or averages) of both sides conditional on X, assuming $E\left(u_i | X\right) = 0$.

$$\begin{aligned} E\left(Y_i | X\right) &= E\left(\beta_0 + \beta_1 X_i + u_i | X\right) \\ &= \beta_0 + \beta_1 X_i + E\left(u_i | X\right) \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

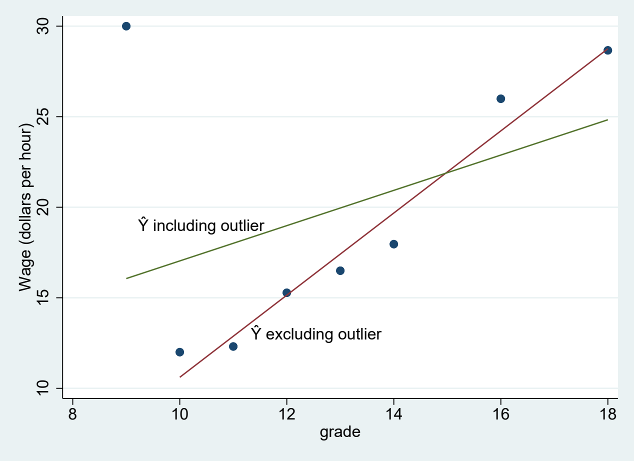
- Therefore, for each one-unit increase in X_i, the population average of Y_i changes by β_1.
- Also, when $X_i = 0$, the population average of $Y_i = \beta_0$.

OLS Assumption 2

- The second OLS assumption is that the data are a simple random sample from a large population.
- Since every observation comes from the *same* population, each has the same expected value and variance.
- Since the observations are picked at random, each is independent of all other observations.
- Suppose Y is the wage and X is the years of education. Then Y_1 is the wage of the first person in a sample and X_1 is the education of the first person in a sample. Both Y_1 and X_1 are random variables because they change from sample to sample.
- OLS assumption 2 means that the mean and variance of Y_1 from sample to sample is the same as the mean and variance of Y_2 from sample to sample, and the same for X_1 and X_2.
- Also, the covariance of Y_1 and X_1 from sample to sample is the same as the covariance of Y_2 and X_2 from sample to sample.

OLS Assumption 3

Figure 19: OLS Assumption 3: Outliers are rare



OLS Assumption 3

- The third OLS assumption is that outliers are rare.
- Large outliers distort the estimated correlation between variables.
- In Figure 18 the wage for persons with 9 years of education is recorded as $30 per hour. This is clearly an outlier, as it is far from the sample regression function (SRF) with this observation excluded, shown as the steeper line. Including the outlier severly twists the estimated SRF, as shown in the flatter line.
- Often outliers are data-entry errors. Perhaps the respondent said he makes $3 per hour but the survey taker wrote down $30.
- Large outliers distort not only the estimated slope but also the *variance* of the estimated slope, which makes hypothesis tests unreliable.
- Because of the problems caused by outliers, we shall assume that they are rare.

Purpose of the OLS assumptions

- The first assumption implies that the estimators are unbiased.

  $$E\left(\hat{\beta}_1\right) = \beta_1$$
  $$E\left(\hat{\beta}_0\right) = \beta_0$$

  On average, the sample OLS line will equal the true, population regression function.
- The assumptions also imply that, in large samples, OLS estimates have a Normal distribution. This allows us to conduct hypothesis tests.
- If the second assumption is false, then we may have to modify the calculations of the coefficients or the standard errors.
- The third assumption shows that OLS estimates are sensitive to large outliers.

Sampling Distribution of OLS estimators

Table 15: Small Population

	Y	X	u	X_c	$X_c u$	X_c^2
	6	2	-4	-2	8	4
	14	2	4	-2	-8	4
	22	6	-4	2	-8	4
	30	6	4	2	8	4
$E(\bullet)$	18	4	0	0	0	4
$V(\bullet)$	80	4	16	4	64	0

Consider the small population above. The true values of the parameters are $\beta_0 = 2$ and $\beta_1 = 4$. We can check whether the first OLS assumption is true of the population above.

$$E\left(u|X=2\right) = \frac{-4+4}{2} = 0\, and\, E\left(u|X=6\right) = \frac{-4+4}{2} = 0$$

Thus OLS asumption 1 is true, and the OLS estimates of β_0 and β_1 are unbiased.

Sampling Distribution of OLS Estimators

Of course we never have the whole population. We only have *samples*. With each new sample, we get new estimates of β_0 and β_1. Suppose we draw all possible samples of three observations from the population in Table 15. The four possible samples along with the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ for each sample are given below.

Table 16: All Samples of Three Observations

	Sample 1		Sample 2		Sample 3		Sample 4	
	Y	X	Y	X	Y	X	Y	X
	6	2	14	2	6	2	6	2
	22	6	22	6	14	2	14	2
	30	6	30	6	22	6	30	6
$\hat{\beta}_0$	-4		8		4		0	
$\hat{\beta}_1$	5		3		3		5	

Sampling Distribution of OLS estimators

- Based on Table 16, the expected value of $\hat{\beta}_0$ is

$$E\left(\hat{\beta}_0\right) = \frac{-4+8+4+0}{4} = 2 = \beta_0$$

- The expected value of $\hat{\beta}_1$ is

$$E\left(\hat{\beta}_1\right) = \frac{5+3+3+5}{4} = 4 = \beta_1$$

Thus on average both estimators equal their true, population values. Thus both are indeed unbiased.

- The variance of $\hat{\beta}_1$ is

$$V\left(\hat{\beta}_1\right) = \frac{(5-4)^2+(3-4)^2+(3-4)^2+(5-4)^2}{4} = 1$$

Variance of OLS slope

- Thus one way to calculate the variance of $\hat{\beta}_1$ is to collect multiple samples and calculate the variance of $\hat{\beta}_1$ across the samples, as above.
- The more common method is to apply a formula to the data from *one* sample. The formula is

$$V\left(\hat{\beta}_1\right) = \frac{1}{n}\frac{var\left(X_{ci}u_i\right)}{\left[var\left(X\right)\right]^2}$$

- For example, using the data in table 15,

$$V\left(\hat{\beta}_1\right) = \frac{1}{4}\frac{64}{4^2} = 1$$

  which matches the variance calculated from the four samples.
- The formula for $V\left(\hat{\beta}_1\right)$ simplifies when the error terms are **homorskedastic**, as explained below.

Homoskedasticity

- Let's calculate the variance of the error term in Table 15 *for each value of X.* X takes on two values, 2 and 6. That is, we want $V(u|X=2)$ and $V(u|X=6)$.

$$
\begin{aligned}
V(u|X=2) &= E\left(u_i - E(u_i|X=2)\right)^2 \\
&= \frac{\left(u_1 - E(u_i|X=2)\right)^2 + \left(u_2 - E(u_i|X=2)\right)^2}{2} \\
&= \frac{(-4)^2 + (4)^2}{2} = 16
\end{aligned}
$$

$$
\begin{aligned}
V(u|X=6) &= E\left(u_i - E(u_i|X=6)\right)^2 \\
&= \frac{\left(u_3 - E(u_i|X=6)\right)^2 + \left(u_4 - E(u_i|X=6)\right)^2}{2} \\
&= \frac{(-4)^2 + (4)^2}{2} = 16
\end{aligned}
$$

- Thus $V(u|X)$ is *the same* for each value of X. This is called **homoskedasticity**. If the variances differ, it is **heteroskedasticity**.

$V\left(\hat{\beta}_1\right)$ under homoskedasticity

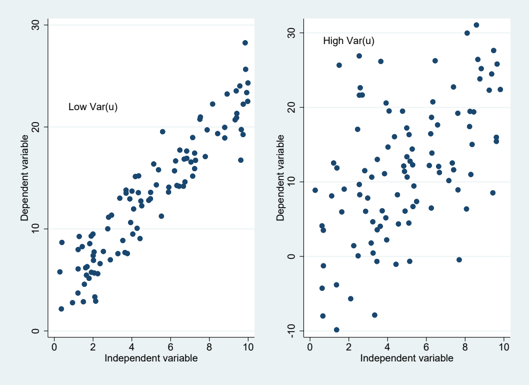
- Under homoskedasticity, the formula for $V\left(\hat{\beta}_1\right)$ simplifies to

  $$V\left(\hat{\beta}_1\right) = \frac{V\left(u_i\right)}{\sum X_{ci}^2} = \frac{1}{n}\frac{V\left(u_i\right)}{V\left(X\right)} = \frac{16}{16} = 1$$

  using the population data in Table 15.
- This is the same result as calculated above using the more complicated formula. That formula works under *both* heteroskedasticity *and* homoskedasticity. The simpler formula works *only* under homoskedasticity. Since in this example the error terms really are homoskedastic, both formulas give the same result.
- We want our estimate of β_1 to be precise. In other words, we want $V\left(\hat{\beta}_1\right)$ to be small. The formula above shows there are three ways of making $V\left(\hat{\beta}_1\right)$ small.
  1. Reduce $V\left(u_i\right)$
  2. Increase $V\left(X\right)$
  3. Increase n

$V\left(\hat{\beta}_1\right)$: Low $V(u)$ vs. High $V(u)$

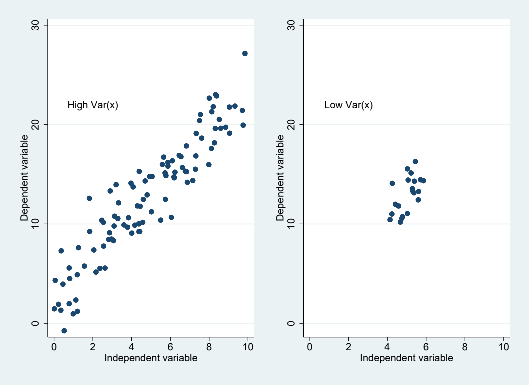
Figure 20: Effect of Low versus High $Var(u)$ on $Var\left(\hat{\beta}_1\right)$



When the variance of u is small, the points cluster close to the line of best fit, clearly revealing its slope, as shown in the left diagram, implying a small variance of $\hat{\beta}_1$.

$V\left(\hat{\beta}_1\right)$: Low $V(X)$ vs. High $V(X)$

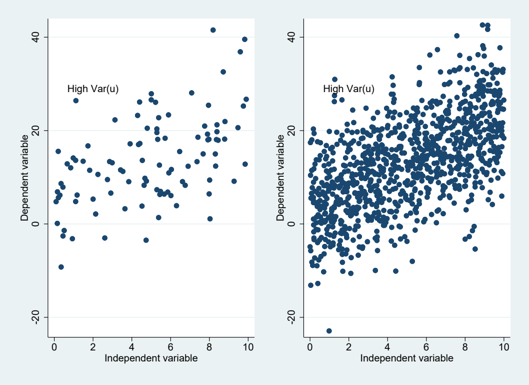
Figure 21: Effect of Low versus High $Var(X)$ on $Var\left(\hat{\beta}_1\right)$



The high variance of X in the left diagram reveals the slope more clearly than the limited variance of X in the right diagram, implying that the variance of $\hat{\beta}_1$ is smaller in the sample on the left than on the right.

$V\left(\hat{\beta}_1\right)$: Low n vs. High n

Figure 22: Effect of Low versus High n on $Var\left(\hat{\beta}_1\right)$



The large sample size in the right diagram reveals the slope more clearly than the small sample size in the left diagram, implying that the variance of $\hat{\beta}_1$ is smaller in the sample on the right than on the left.

$V\left(\hat{\beta}_1\right)$ under homoskedasticity from a sample

- Because we only observe samples, we must use *an estimate* of $V\left(u_i\right)$, which has been shown to be

$$\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-2} = \frac{SSR}{n-2}$$

- Hence

$$V\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}_u^2}{\sum X_{ci}^2} = \frac{32}{10\frac{2}{3}} = 3$$

using data from the first sample in Table 16.

	Y	X	X_c^2	$\hat{Y}$	$\hat{u}^2$
	6	2	7.11	6	0
	22	6	1.78	26	16
	30	6	1.78	26	16
$\sum(\bullet)$			10.67		32

Distribution of $\hat{\beta}_1$

The previous slides have established the following essential facts about $\hat{\beta}_1$.

- If $E(u|X) = 0$, $E\left(\hat{\beta}_1\right) = \beta_1$ and $E\left(\hat{\beta}_0\right) = \beta_0$. That is, $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased estimators of β_1 and β_0.
- $\hat{\beta}_1$ is Normally distributed. Recall the formula for $\hat{\beta}_1$.

$$\hat{\beta}_1 = \frac{\sum X_{ci} Y_{ci}}{\sum X_{ci}^2} = \frac{\sum X_{ci} Y_{ci}/n}{\sum X_{ci}^2/n}$$

  The numerator, $\sum X_{ci} Y_{ci}/n$, *is a sample average*. Since sample averages are Normally distributed, so is $\hat{\beta}_1$

- The variance of $\hat{\beta}_1$ is

$$V\left(\hat{\beta}_1\right) = \frac{1}{n} \frac{V\left(X_{ci} u_i\right)}{V\left(X_i\right)^2}$$

  whether u_i is heteroskedastic or homoskedastic. If it is homoskedastic, then

$$V\left(\hat{\beta}_1\right) = \frac{1}{n} \frac{V\left(u_i\right)}{V\left(X_i\right)}$$

Hypothesis Tests with OLS

- Suppose we want to test the hypothesis that studying has no effect on grades. We assume the following model.

  $$Grade_i = \beta_0 + \beta_1 Studyhours_i + u_i$$

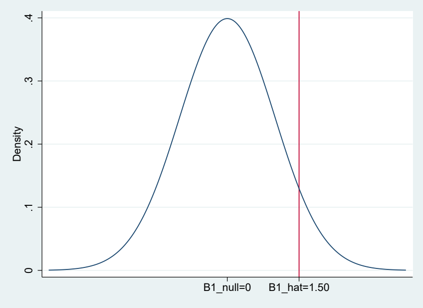
  The coefficient on Studyhours, β_1, tells us that an hour increase in study is associated with a change in the predicted grade of β_1 points. If $\beta_1 = 0$, then studying has no effect. Thus we may want to test whether $\beta_1 = 0$.
- The hypothesis that we want to test is called the **null hypothesis**, or H_0. The contrasting hypothesis is the **alternative hypothesis**, or H_A. In this example H_A is that studying has *some* effect. Because in this example either a positive or negative effect disproves the null hypothesis, we are conducting a **two-sided** hypothesis test. The null and alternative hypotheses are usually stated together as follows.

  $$H_0 : \beta_1 = 0$$
  $$H_A : \beta_1 \neq 0$$

Hypothesis Tests with OLS

- Recall that we estimate $\hat{\beta}_1$ from *one sample*. To test hypotheses, we must know how $\hat{\beta}_1$ varies across *many samples*. This is called the **sampling distribution** of $\hat{\beta}_1$.
- We shall assume that $\hat{\beta}_1$ is Normally distributed across samples with a mean of $\beta_{1,null}$ and a standard deviation of $se\left(\hat{\beta}_1\right)$.
- In the example above, $H_0 : \beta_1 = 0$. Also, suppose $\hat{\beta}_1 = 1.5$. Then:

Figure 23: Distribution of $\hat{\beta}_1$



Calculating the pvalue

Using $E\left(\hat{\beta}_1\right)$ from H_0 and $V\left(\hat{\beta}_1\right)$ calculated from the sample, we calculate and apply the pvalue in three steps.

1. Calculate the sample t-statistic, using the null hypothesis:

$$t_{sample} = \frac{\hat{\beta}_1 - E\left(\hat{\beta}_1\right)_{H_0}}{se\left(\hat{\beta}_1\right)}$$

2. Calculate $pvalue = P\left(t > |t_{sample}|\right) \cdot 2$, assuming it is a two-sided alternative hypothesis.
3. Compare pvalue with α, the level of significane, chosen before we do the test. If $pvalue < \alpha$, reject H_0. If not, fail to reject.

Calculating the pvalue

- Based on the model of Grade above, $\beta_{1,null} = 0$, $\hat{\beta}_1 = 1.5$, $se\left(\hat{\beta}_1\right) = 1$. Also, let's assume the sample size, n, is 100. Since we are estimating two parameters (β_0, β_1) the degrees of freedom is 98. Also, let's choose $\alpha = 0.05$. Therefore

$$\begin{aligned} t_{sample} &= \frac{\hat{\beta}_1 - E\left(\hat{\beta}_1\right)_{H_0}}{se\left(\hat{\beta}_1\right)} \\ &= \frac{1.5 - 0}{1} = 1.5 \end{aligned}$$

- Using the Excel t.dist.rt($|t_{sample}|$,d.f.) function,

$$P\left(t > |t_{sample}|\right) \cdot 2 = 0.1368.$$

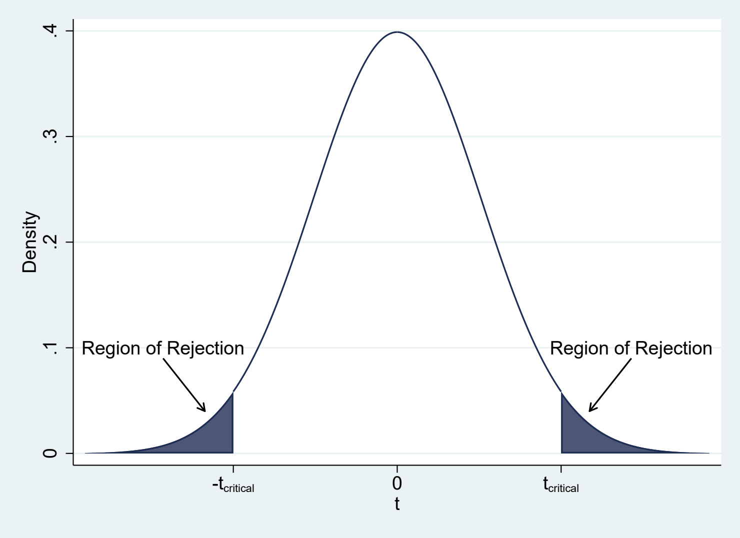
- Since $0.1368 > 0.05$, we *fail to reject* H_0.

Hypothesis Testing with a Prespecified Significance Level

- Another way to test hypotheses is to find the value of t_{sample} at which the p-value equals α. This is called the **critical t-value** or $t_{critical}$.
- For a two-sided text, any t_{sample} *larger* than $t_{critical}$ or smaller than $-t_{critical}$ will produce a p-value *smaller* than α, so we reject H_0.
- Any t_{sample} *between* than $-t_{critical}$ will and $t_{critical}$ produces a p-value *larger* than α, so we fail to reject H_0.
- To understand this, recall that the p-value is the area in the *tails* of the distribution of t_{sample}. As t_{sample} gets bigger, the area in the tails (and therefore the p-value) gets smaller.
- We can find the critical t-statistic for a two-sided test using the Excel function =t.inv.2t(α, d.f.). For example, for the hypothesis test above, $t_{critical}$ = t.inv.2t(0.05, 98) = 1.98. Any sample-t *larger* than this will produce a p-value *smaller* than 0.05, and any sample-t *smaller* than this will produce a p-value *larger* than 0.05.
- In the example above, since $t_{sample} = 1.5 < 1.98 = t_{critical}$, we fail to reject.

Hypothesis Testing with a Prespecified Significance Level

Figure 25: Hypothesis testing with critical t-values



Confidence Intervals

- Using $t_{critical}$, we can calculate confidence intervals for β_0 and β_1.
- In the example above, $t_{critical} = 1.98$. Therefore we know that, for a random t,

$$P\left(-t_{critical} < t < t_{critical}\right) = 0.95$$

Replace t with the formula for the t-statistic, assuming $E\left(\hat{\beta}_1\right) = \beta_1$, and rearrange to get β_1 in the center.

$$\begin{aligned}
P\left(-t_{critical} < \frac{\hat{\beta}_1 - \beta_1}{se\left(\hat{\beta}_1\right)} < t_{critical}\right) &= 0.95 \\
P\left(-t_{critical} \cdot se\left(\hat{\beta}_1\right) < \hat{\beta}_1 - \beta_1 < t_{critical} \cdot se\left(\hat{\beta}_1\right)\right) &= 0.95 \\
P\left(-\hat{\beta}_1 - t_{critical} \cdot se\left(\hat{\beta}_1\right) < -\beta_1 < -\hat{\beta}_1 + t_{critical} \cdot se\left(\hat{\beta}_1\right)\right) &= 0.95 \\
\left(\hat{\beta}_1 - t_{critical} \cdot se\left(\hat{\beta}_1\right) < \beta_1 < \hat{\beta}_1 + t_{critical} \cdot se\left(\hat{\beta}_1\right)\right) &= 0.95
\end{aligned}$$

Confidence Intervals

- Thus the lower bound (LB) of the confidence interval is

$$LB = \hat{\beta}_1 - t_{critical} \cdot se\left(\hat{\beta}_1\right)$$

  and the upper bound is

$$\hat{\beta}_1 + t_{critical} \cdot se\left(\hat{\beta}_1\right)$$

- Let's calculate a confidence interval for the Grade example.

$$LB = 1.5 - 1.98 \cdot 1 = -0.48$$
$$UB = 1.5 + 1.98 \cdot 1 = 2.48$$

- A confidence interval is usually reported with the coefficient as follows.

$$1.5\,(-0.48, 2.48)$$

Confidence Intervals

- Note that every time we collect a new sample, we would have a new $\hat{\beta}_1$ and therefore a new confidence interval. Thus $\hat{\beta}_1$ and the confidence interval are random variables.
- But the true β_1 would be the same each time.
- Thus we cannot say that a given confidence interval has a 95% chance of including the true value. A given confidence interval either includes the true β_1 or it doesn't.
- Rather, the correct interpretation of a confidence interval is to state that if we collect 100 samples and calculate 100 confidence intervals, 95 of them would include the true β_1.
- An important feature of a confidence interval is that it shows all the values of a null hypothesis that cannot be rejected at the significance level of α. For example, a confidence interval of $(-0.48, 1.48)$ shows that we cannot reject the hypothesis that $\beta_1 = 0$ at the 5% level because 0 is inside the confidence interval.

Confidence Intervals

- Suppose we sample 100 young adults in New York City and regress the wage (in dollars per hour) on education (in years), obtaining the following. (Standard errors are in parentheses below the coefficients.)

$$Wage = \underset{(9.8)}{-19} + \underset{(0.65)}{2.9} \cdot Education$$

  Let's calculate a 99% confidence interval around $\hat{\beta}_1$.

- First, find the $t_{critical}$. Using Excel's =t.inv.2t(0.01,98) gives 2.63. Thus

$$LB = 2.9 - 2.63 \cdot 0.65 = 1.2$$
$$UB = 2.9 + 2.63 \cdot 0.65 = 4.6$$

- Can we reject $H_0 : \beta_1 = 1$ at the 1% level? Yes! because 1 is outside the 99% confidence interval.

Binary or Dummy Variables

- A **binary** or **dummy** variable takes two values, one or zero, depending on whether a condition is true or false. For example, the variable *female* equals one if the person is a female, and zero otherwise.
- Consider a model of wages.

$$wage_i = \beta_0 + \beta_1 \cdot female_i + u_i$$

  This model can be estimated using OLS just like any other model, but the interpretation of β_1 is different.
- A *continuous* independent variable is interpreted as follows: A one unit increase in X is associated with a change in the predicted value of Y of $\hat{\beta}_1$ units of Y.
- But a dummy variable is not continuous. It has only two values. One cannot say, "A one unit change in female..." What is a one-unit change in female?

Binary or Dummy Variables

- Instead, to interpret a dummy variable, one makes two predictions and compares them. For example, using the wage model above, first predict the wage when female = 1.

$$\widehat{wage}_1 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1$$

- Now predict the wage when female = 0.

$$\begin{aligned} \widehat{wage}_0 &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 \\ &= \hat{\beta}_0 \end{aligned}$$

- Now take the difference.

$$\Delta w\hat{a}ge = \hat{\beta}_1$$

Thus $\hat{\beta}_1$ is the *difference* in the average wages of males and females.

Binary or Dummy Variables

- For example, we can estimate the model above using the dataset nlsy97wages2015, available on Blackboard. The data include people who were between 30 and 36 in 2015. The results are as follows (standard errors in parentheses).

$$\widehat{wage}_i = \underset{(0.312)}{22.8} - \underset{0.44}{3.23} \cdot Female$$

- The predicted wage for a female is

$$\widehat{wage} = 22.8 - 3.23 \cdot 1 = \$19.6$$

- The predicted wage for a male is

$$\widehat{wage} = 22.8 - 3.23 \cdot 0 = \$22.8$$

- Thus among 30- to 36-year-olds in 2015, men made about $3.23 per hour more than women.

Definition of Omitted Variable Bias

- Reconsider the model of education and earnings.

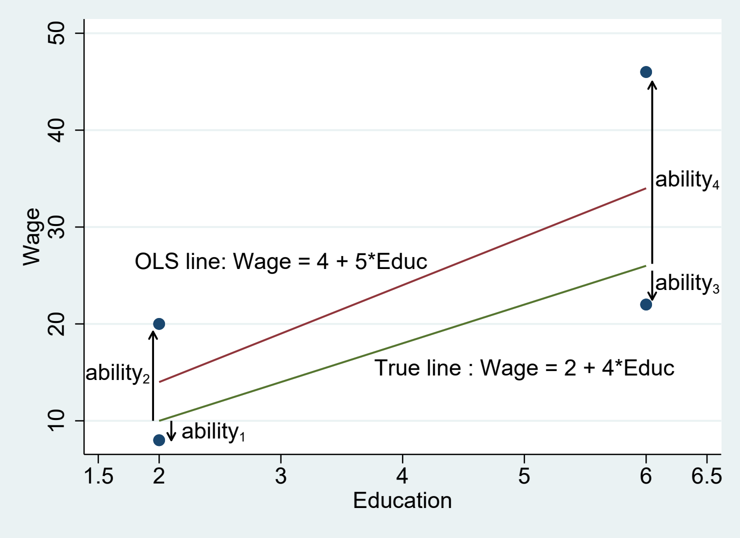
$$wage_i = \beta_0 + \beta_1 \cdot education_i + u_i$$

  where $education$ is measured in years. Recall that the error term, u_i, include all the factors that affect wages *other than* education.

- One of these factors may be scholastic ability. A scholastically able person will
  1. go to school longer because she is good at it, and
  2. make more income even if she doesn't go to school.
- Thus more educated people make more income partly because of education and partly because of ability. The effects of education and ability are mixed together.
- But β_1 is supposed to measure the pure effect of education. Failure to isolate the effect of education from the effect of factors omitted from the analysis will bias the estimate of β_1. This is called **omitted variable bias** or **OVB.**

Visual Example of Omitted Variable Bias

Figure 26: Omitted Variable Bias



Visual Example of Omitted Variable Bias

- Figure 26 shows how the correlation between ability and education biases the estimated effect of education on earnings.
- Assume u_i is $ability_i$.
- The true effect of education on earnings is $\beta_1 = 4$, as shown in the true line.
- But as education increases, so does average ability. That is, $cov\left(Education, Ability\right) > 0$.
- Since OLS always passes through the center of the values of the dependent variable, the OLS slope is steeper than the true line. In example above, the OLS slope is 5.
- Hence the omitted variable bias in this example is $\hat{\beta}_{OLS} - \beta_{true} = 5 - 4 = 1$.
- Now let's illustrate this point with actual data.

Formula for Omitted Variable Bias

- Suppose the true model is

  $$wage_i = \beta_0 + \beta_1 \cdot education_i + \beta_2 \cdot ability_i + u_i.$$

  Let's call this the **long model**.

- But suppose we lack data on ability, so instead we estimate the **short model**.

  $$wage_i = \alpha_0 + \alpha_1 \cdot education_i + v_i$$

- Note that ability is in the error term of the short model.

  $$v_i = \beta_2 \cdot ability + u_i$$

- Will $\hat{\alpha}_1 = \hat{\beta}_1$? We can find out using the **omitted variable bias formula**.

  $$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2 \frac{cov\left(ability, education\right)}{var\left(education\right)}$$

Example of Omitted Variable Bias

- Using the data set nlsy97wages2015.dta, we estimate both the short and long models. For ability, we use the variable ASVAB, a scholastic aptitude test. Note that some observations on ASVAB are missing. We use only those observations with non-missing values for wage, education, and ASVAB.
- The results of the short- and long-model are (standard errors in parentheses)

$$\widehat{wage}_i = \underset{(0.8)}{12} + \underset{(0.06)}{0.69} Educ$$

$$\widehat{wage}_i = \underset{(0.8)}{10} + \underset{(0.06)}{0.35} Educ + \underset{(0.009)}{0.14} ASVAB$$

  Also, $cov(educ, ability) = 50.57$ and $var(educ) = 20.8$.
- Using the the OVB formula, the bias in $\hat{\alpha}_1$ is

$$bias = \hat{\beta}_2 \frac{cov(ability, educ)}{var(educ)} = 0.14 \frac{50.57}{20.8} = 0.34$$

Key points on OVB

- Memorize the OVB formula.

$$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2 \frac{cov\left(X_{included}, X_{omitted}\right)}{var\left(X_{included}\right)}$$

  where $\hat{\alpha}_1$ is the OLS slope ignoring the omitted variable, $\hat{\beta}_1$ is an unbiased estimate of β_1, the true effect of education, and $\hat{\beta}_2$ is an unbiased estimate of β_2, the true effect of ability.

- OVB is only a problem if
  1. $cov\left(X_{included}, X_{omitted}\right) \neq 0$
  2. $\beta_2 \neq 0$, meaning that the omitted variable has an effect on Y_i.

Multiple Regression

- The most natural way of solving the omitted variable problem is to include the omitted variable. Then it will not be omitted. This leads to the **multiple regression model**.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

  where X_{1i} is the first independent variable, X_{2i} is the second independent variable, β_1,β_2 are the true, constant population coefficients and u_i is the error term.

- Take expectations of both sides conditional on X_1 and X_2, meaning that we treat X_1 and X_2 as known constants, as in $X_{1i} = 5$.

$$E\left(Y_i | X_1, X_2\right) = E\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i | X_1, X_2\right)$$

- Recall the rules of expectations. Assume a, b are constants and X, Y are variables.
  1. $E\left(a\right) = a$
  2. $E\left(a \cdot X + b \cdot Y\right) = a \cdot E\left(X\right) + b \cdot E\left(Y\right)$

Multiple Regression

- Thus β_0, β_1, X_{1i}, and X_{2i} are all constants *given* specific values of X_{1i}, and X_{2i}

$$\begin{aligned} E\left(Y_i|X_1, X_2\right) &= E\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i|X_1, X_2\right) \\ &= E\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}|X_{1i}, X_{2i}\right) + E\left(u_i|X_{1i}, X_{2i}\right) \\ &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E\left(u_i|X_{1i}, X_{2i}\right) \end{aligned}$$

- The first OLS assumption for multiple regression is

$$E\left(u_i|X_1, X_2\right) = 0$$

If this is true, then the **Population Regression Function** is

$$E\left(Y_i|X_{1i}, X_{2i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

where β_0 is the Y-intercept, β_1 is the coefficient on X_1, and β_2 is the coefficient on X_2.

Multiple Regression

- The interpretation of β_1 is different in the multiple regression model from the simple regression model. Suppose $X_{1i} = 4$ and $X_{2i} = 6$. Then

$$E\left(Y_i | X_{1i} = 4, X_{2i} = 6\right) = \beta_0 + \beta_1 \cdot 4 + \beta_2 \cdot 6$$

- Now suppose X_{1i} increases by one unit to 5 while X_{2i} remains at 6. Then

$$E\left(Y_i | X_{1i} = 4, X_{2i} = 6\right) = \beta_0 + \beta_1 \cdot 5 + \beta_2 \cdot 6$$

- Now take the difference. Note that we keep X_2 constant. Therefore both β_0 *and* $\beta_2 \cdot 6$ cancel out.

$$\Delta E\left(Y_i | X_{1i}, X_{2i}\right) = \beta_1$$

  Thus β_1 is the change in the expected or average value of Y when X_1 increases by one unit *and X_2 is held constant.*

Interpreting Population Slopes in Multiple Regression

- Let's interpret the coefficients in the wage model.

$$wage_i = \beta_0 + \beta_1 \cdot education_i + \beta_2 \cdot ability + u_i.$$

Assuming $E\left(u_i | education, ability\right) = 0$, the PRF is

$$E\left(wage_i | education_i, ability_i\right) = \beta_0 + \beta_1 \cdot education_i + \beta_2 \cdot ability_i$$

- β_0 is the $E\left(wage_i\right)$ when $education = 0$ and $ability = 0$
- β_1 is the $\Delta E\left(wage_i\right)$ when education increases by one unit, *holding ability constant*.
- β_2 is the $\Delta E\left(wage_i\right)$ when ability increases by one unit, *holding education constant.*

Calculating Coefficients in the Multiple OLS model

- As with a simple OLS, define the **residual** as

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}$$

- We want to find the values of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ that minimize the sum of squared residuals.

$$\underset{\hat{\beta}_0 \hat{\beta}_1 \hat{\beta}_2}{min} \sum_{i=1}^{n} \hat{u}_i^2$$

- Once we have $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, we can calculate the **predicted values** of Y_i.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

- If we calculate the coefficients correctly, the following will be true.

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

$$\sum_{i=1}^{n} \hat{u}_i X_i = 0$$

Multiple Regression example

- Let's return to the NLSY regression of wages on education and ability.

$$\widehat{wage}_i = \underset{(0.8)}{12} + \underset{(0.06)}{0.69}\,Educ$$

$$\widehat{wage}_i = \underset{(0.8)}{10} + \underset{(0.06)}{0.35}\,Educ + \underset{(0.009)}{0.14}\,ASVAB$$

- Wages are measured in dollars per hour, education in years, and ability in points from 0 to 100. The interpretation of $\hat{\beta}_1$ in the simple regression is, "A one-year increase in education is associated with an increase in wages of 0.69 dollars per hour."
- Now interpret $\hat{\beta}_1$ in the multiple regression. "A one-year increase in education is associated with an increase in the wage of 0.35 dollars per hour, *holding ASVAB constant*."
- In other words, if we compare two people with the same ASVAB score, and one of the persons has a year more of education, she is *predicted* (not "expected", which refers to the *population* regression function) to earn 0.35 dollars more per hour.

More Multiple OLS examples

- Ability is likely not the only omitted variable in our wage model. Let's add gender and work experience in years. Define work experience as the cumulative number of hours worked divided by 2080 (52 weeks * 40 hours per week).

$$wage_i = \beta_0 + \beta_1 education_i + \beta_2 asvab_i + \beta_3 experience_i + \beta_4 female_i + u_i$$

- Using the data set nls797wages2015 we get

$$\widehat{wage}_i = \underset{(1.05)}{3.73} + \underset{(0.057)}{0.40}\, education_i + \underset{(0.009)}{0.13}\, asvab_i + \underset{(0.053)}{0.62}\, experience_i - \underset{(.49)}{2.68} female_i$$

- Interpret each of the coefficients. The correct interpretations are on the next slide.

Interpreting Multiple OLS coefficients

- Each extra year of education is associated with an increase in the wage of 0.4 dollars per hour, holding asvab, experience, and gender constant.
- Each additional point in the asvab score is associated with an increase in the wage of 0.13 dollars per hour, holding education, experience, and gender constant.
- Each additional year of experience is associated with an increase in the wage of 0.62 dollars per hour, holding education, asvab, and gender constant.
- On average, women make 2.68 dollars per hour less than men, holding education, asvab, and experience constant.
- On average, a man with no education, no ability, and no experience makes 3.73 dollars per hour.
- Never say “holding everything else constant”. We can never hold everything else constant.

The Standard Error of the Regression

- The standard error of the regression (SER) estimates the standard deviation of the error term, u_i. It measures the spread of the distribution of Y around the regression line.
- It is estimated using the Sum of Squared Residuals (SSR) divided by the degrees of freedom of the residual.

$$SER = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{d.f.}}$$

- The degrees of freedom equals the number of observations, n, minus the number of estimated slopes, k, minus one for the intercept. Thus $d.f. = n - k - 1$.
- In the wage regression above, $SSR = 1408408$, $n = 4934$, and $k = 4$. Therefore

$$SER = \sqrt{\frac{1408408}{4929}} = \$16.9$$

R^2

- R^2 is the proportion of the sample variation in Y_i explained by the variation in the independent variables.
- In the wage regression above, $R^2 = 0.11$. Thus education, asvab, experience, and gender explain about 11% of the variation in wages.
- The formula is the same for multiple as for simple regression.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

where $TSS = \sum \left(Y_i - \bar{Y}\right)^2$ and $ESS = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$.

- R^2 always increases when an additional independent variable is added.
- Sometimes a preferred measure of fit is the **adjusted $\boldsymbol{R^2}$**, or $\overline{R}^2$.

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

$\overline{R}^2$ increases only when the additional independent variables has a t-statistic greater than one in absolute value.

OLS assumption 1

Consider the following small population, in which $\beta_0 = 2$, $\beta_1 = 3$, $\beta_2 = 4$.

Table 17: $E\left(u|X\right)$ in Multiple Regression

Y_i	X_1	X_2	u_i	v_i
10	2	1	-2	2
14	2	1	2	6
18	2	3	-2	10
22	2	3	2	14
32	4	5	-2	18
36	4	5	2	22
40	4	7	-2	26
44	4	7	2	30

OLS Assumption 1

- The true model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- The false model is

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + v_i$$

Thus the error term in the false model is

$$v_i = \beta_2 X_{2i} + u_i$$

- Calculate $E\left(v_i | X_{1i}\right)$.

$$E\left(v_i | X_{1i} = 2\right) = \frac{2+6+10+14}{4} = 8$$
$$E\left(v_i | X_{1i} = 4\right) = \frac{18+22+26+30}{4} = 24$$

Since $E\left(v_i | X_{1i}\right)$ is not constant, $\hat{\alpha}_1$ is biased for β_1; since $E\left(v_i | X_{1i}\right) \neq 0$, $\hat{\alpha}_0$ is biased for β_0.

OLS Assumption 1

- Calculate $E\left(u_i|X_{1i}, X_{1i}\right)$.

$$E\left(u_i|X_{1i}=2, X_{2i}=1\right) = \frac{-2+2}{2} = 0$$
$$E\left(u_i|X_{1i}=2, X_{2i}=3\right) = \frac{-2+2}{2} = 0$$
$$E\left(u_i|X_{1i}=4, X_{2i}=5\right) = \frac{-2+2}{2} = 0$$
$$E\left(u_i|X_{1i}=4, X_{2i}=7\right) = \frac{-2+2}{2} = 0$$

- Thus $E\left(u_i|X_{1i}, X_{2i}\right) = 0$ for all combinations of X_{1i} and X_{2i}. Therfore $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are unbiased estimators of β_0, β_1, and β_2.
- In othe words, at each combination of X_{1i} and X_{2i} the error terms above the OLS line offset the error terms below the OLS line. Hence the OLS and true line are on average the same.

OLS Assumptions 2 and 3

- OLS assumption 2: The values of each observation in a sample is independent of the values of every other observation in the sample. For example, the age of the first person sampled is unrelated to the age of the second person sampled.This holds automatically if each observation is randomly chosen from the same population.
- OLS assumption 3: Outliers–values far outside the usual range of the data–are rare. OLS estimates are sensitive to extreme values. This sensitivity reduces the precision of the estimates and can obscure important relationships among the variables.
- For example, suppose we want to find the effect of schooling on earnings. Some datasets include billionaires, but years of education has virtually no effect on the likelihood of becoming a billionaire. Thus including billionaires in the regression will pull the coefficient towards zero. It may be wise to leave out billionaires from the analysis since we are interested in the effect of education on typical workers.

OLS Assumption 4: No Perfect Multicollinearity

- Variables are perfectly collinear when one is a linear combination of the others. For example, temperature in farenheit (F) is a linear function of the temperature in Celsius (C).

$$F = 32 + \frac{9}{5}C$$

- Thus we cannot estimate

$$Exercise\, minutes_i = \beta_0 + \beta_1 F_i + \beta_2 C_i + u_i$$

  because β_1 is supposed to show the effect on exercise of a change in F *holding C constant*. But we cannot hold C constant if we change F.

- Suppose $family\, income$ is the sum of $mom\, income$ plus $dad\, income$. Then we cannot estimate

$$\begin{aligned} Vacation\, spending_i = \beta_0 + \beta_1 mom\, income_i + \beta_2 dad\, income_i \\ + \beta_3 family\, income_i + u_i \end{aligned}$$

  because we cannot change $family\, income$ without changing either $mom\, income$ or $dad\, income$.

Dummy Variable Trap

- Suppose we want to estimate the effect of gender on earnings and include dummy variables for both female and male.

  $$wage_i = \beta_0 + \beta_1 female_i + \beta_2 male_i + u_i$$

  Will this work? No, because $female_i = 1 - male_i$. That is, $female_i$ is a linear function of $male_i$.

- Suppose we want to estimate the effect of marital status on earnings, using dummy variables for every category.

  $$\begin{aligned} wage_i = \beta_0 + \beta_1 married_i + \beta_2 never\,married_i + \beta_3 cohabiting_i \\ + \beta_4 separated_i + \beta_5 divorced_i + \beta_6 widowed_i + u_i \end{aligned}$$

  Will this work? No, because

  $$\begin{aligned} married = 1 - never\,married - cohabiting \\ - separated - divorced - widowed \end{aligned}$$

  Including dummies for all categories is the **dummy variable trap**.

Interpreting Dummy Variables

- The solution to the dummy variable trap is to omit one category. The omitted category is called the **base category**. Usually the base category is the most numerous. For example, let's leave out $married$ in the model above and estimate it using the NLSY97wages2015 dataset. The results are

$$\begin{aligned}\widehat{wage}_i = \underset{(0.4)}{24.6} - \underset{(0.6)}{6.1}\, nevermarried_i - \underset{(0.8)}{2.9}\, cohabiting_i \\ - \underset{(2.1)}{6.7}\, separated_i - \underset{(1.0)}{4.5}\, divorced_i - \underset{(5.1)}{11.2} widowed_i\end{aligned}$$

- Let's predict the wage of a never-married person.

$$\widehat{wage}_i = 24.6 - 6.1 \cdot 1 - 2.9 \cdot 0 - 6.7 \cdot 0 - 4.5 \cdot 0 - 11.2 \cdot 0$$

Now predict the wage of a married person.

$$\widehat{wage}_i = 24.6$$

Now take the difference. Hence never-married persons make $\$6.1$ less per hour than married persons.

Imperfect Multicollinearity

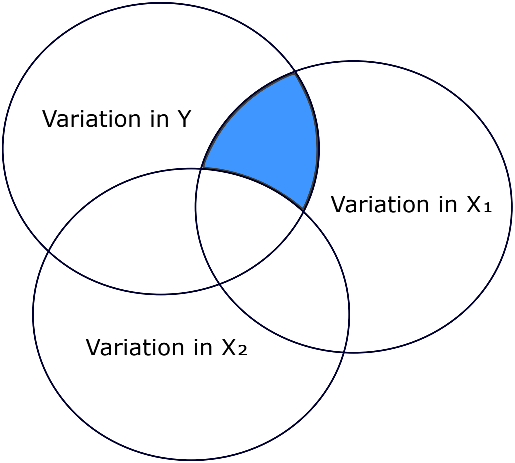
- Imperfect multicollinearity occurs when independent variables are correlated, but not perfectly.
- For example, suppose we want to estimate the effect on wages of one's own education as well as that of one's parents.

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 mom\, educ_i + \beta_3 dad\, educ_i + u_i$$

  The education of the child, the mother, and the father tend to be highly but not perfectly correlated; thus they are imperfectly collinear.
- Because the correlation is not perfect, all three variables can be included in an OLS regression.
- The drawback of collinearity is that it increases standard errors and widens confidence intervals. In other words, coefficients of highly collinear variables tend to be estimated imprecisely. The reason for this can be illustrated with Venn diagrams.

Imperfect Multicollinearity

Figure 27: Imperfect Multicollinearity



Imperfect Multicollinearity

- Each circle represents the variation in a variable. The overlap of the circles represents the *covariation* between those variables. Since Y overlaps both X_1 and X_2 it is correlated with both. Similarly, the overlap between X_1 and X_2 shows that they are also correlated.
- β_1 shows the effect of X_1 on Y, *holding* X_2 *constant*. Thus β_1 reflects none of the variation X_1 shares with X_2, which is the area where X_1 and X_2 overlap. β_1 is calculated using *only* the area X_1 shares uniquely with Y (shaded in blue) excluding the area X_1 shares with X_2.
- The more X_1 and X_2 overlap, the less variation remains that is unique to X_1 and Y. That is, the blue area becomes smaller. This will produce an imprecise estimate of β_1.
- Figure 27 also shows R^2. It is the area where X_1 and X_2 overlap with Y divided by the size of the Y circle.
- Next let's estimate the model wages and parental education.

Imperfect Multicollinearity example

Table 18: Imperfect Multicollinearity

	(1) Wage	(2) Wage	(3) Wage
Education (years)	0.801	0.674	0.639
	(0.0663)	(0.0682)	(0.0692)
Mom's education (years)		0.737	0.524
		(0.103)	(0.126)
Dad's education (years)			0.334
			(0.114)
Constant	11.03	3.386	2.358
	(1.000)	(1.457)	(1.498)
Observations	4162	4162	4162

Standard errors in parentheses

Imperfect Multicollinearity example

- The first colum shows the effect solely of a person's own education on her wage. The second column adds the mother's education, and the third adds the father's education as well.
- The correlation between one's own education and one's parents' education is about 0.3. The correlation between the parents' education is about 0.6.
- Because of the collinearity among the three independent variables, the standard error on one's own education rises as the mother's and then the father's education is added to the model. In this example, the increase in the standard error is pretty small, but often that is not the case.
- Imperfect multicollinearity is not a problem in itself. If the standard errors are small enough to obtain precise estimates of the coefficients, it can be ignored. If it greatly increases standard errors, a good solution is to get a bigger sample. Otherwise one might reconsider the necessity of including all the collinear variables.

Hypothesis Tests for a Single Coefficient

- Testing hypotheses about individual coefficients in multiple regression is the same as in simple regressions.
- Using the results in table 18, let's test whether $\beta_{mom\,education} = 0.3$.
  1. State the null and alternative hypotheses and choose α. Let's set $\alpha = 0.05$.
     $$H_0 : \beta_2 = 0.3$$
     $$H_A : \beta_2 \neq 0.3$$
  2. Calculate the sample t-statistic.
     $$t_{sample} = \frac{\hat{\beta}_2 - \beta_{2,H_0}}{se\left(\hat{\beta}_2\right)} = \frac{0.52 - 0.3}{0.13} = 1.69$$
  3. Find the p-value. Note that $d.f. = n - k - 1 = 4162 - 3 - 1 = 4158$
     $$pvalue = 2 \cdot t.dist\left(-abs\left(t_{sample}\right), d.f., true\right) = 0.09$$
  4. Since $pvalue > \alpha$, we fail to reject.

Confidence Intervals for Single Coefficients

- Now we create a confidence interval for β_2. Let's choose $\alpha = 0.05$.
  1. Find the critical t-statistic, which is always positive.

$$t_{critical} = t.inv.2t\left(\alpha, d.f.\right) = 1.96$$

  2. Calculate the lower and upper bounds.

$$LB = \hat{\beta}_2 - t_{critical} \cdot se\left(\hat{\beta}_2\right) = 0.524 - 1.96 \cdot 0.126 = 0.277$$
$$UB = \hat{\beta}_2 + t_{critical} \cdot se\left(\hat{\beta}_2\right) = 0.524 + 1.96 \cdot 0.126 = 0.77$$

- Recall that the confidence interval shows all the hypotheses that cannot be rejected at the significance level α. Using the confidence interval just calculated, can we reject $H_0 : \beta_2 = 0.3$? No, because it is inside the confidence interval.

The F-test of a Single Coefficient

- Suppose a variable X includes the following seven values, all equally likely.

  $$X = \{-3, -2, -1, 0, 1, 2, 3\}$$

  What is $prob\,(X < -2\,or\,X > 2)$? $2/7$.
  What is $prob\left(X^2 > 2^2\right)$? $2/7$, the same as above.

- Recall that in a two-tailed test on a single coefficient, the pvalue is the area in the tails of the distribution of t_{sample}.

  $$pvalue = prob\,(t < -abs\,(t_{sample})\;or\,t > abs\,(t_{sample}))$$

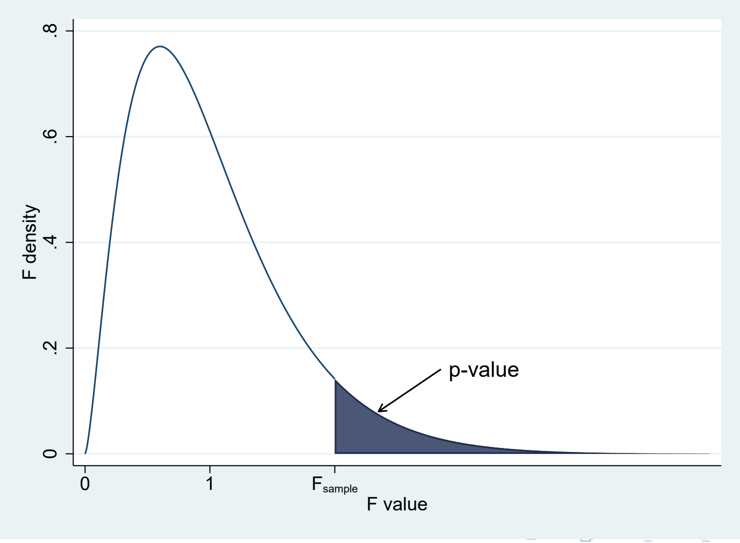
  Based on the logic above, this equals

  $$pvalue = prob\left(t^2 > t^2_{sample}\right)$$

- But how do we find this probability? It turns out that t^2 has a known probability distribution, called the F.

The F Distribution

Figure 28: The F distribution



The F Distribution

- Mathematically, the F-distribution is the ratio of two chi-squareds.

  $$F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}$$

  where n is the numerator degrees of freedom and m is the denominator degrees of freedom.
- Since $F_{n,m}$ is the ratio of two positive numbers, it is also positive.
- If H_0 is true, $F_{n,m}$ will be close to 1.
- For instance, let's find the p-value of the previous example, $H_0 : \beta_{mom\, education} = 0$. Because it is a test of one coefficient, $n = 1$, $m = 4158$, and $F_{sample} = (t_{sample})^2 = 1.69^2 = 2.8561$. Then using Excel,

  $$\begin{aligned} pvalue &= F.DIST.RT\left(F_{sample}, df_n, df_d\right) \\ &= F.DIST.RT\left(2.8561, 1, 4158\right) = 0.09 \end{aligned}$$

  the same as obtained previously using the t distribution.

Testing Hypotheses on Two or More Coefficients

- Sometimes we are interested in testing whether *multiple* coefficients are equal to specific values. For example, we may want to test whether the *neither* parent's education affects the child's wage. That is, we want to test if $\beta_2 = 0$ *and* $\beta_3 = 0$. This hypothesis is

$$H_0 : \beta_2 = 0\, and\, \beta_3 = 0$$
$$H_A : \beta_2 \neq 0\, or\, \beta_3 \neq 0$$

- Generally we reject H_0 when some test statistic is large. What test statistic would be large if *either* $\hat{\beta}_2$ *or* $\hat{\beta}_3$ is far from its null value? One possibility is the *average* of the squared t-statistics on the individual slopes. Under certain conditions, this statistic is an F.

$$F = \frac{t_2^2 + t_3^2}{2}$$

where for $i = 2, 3$

$$t_i^2 = \left(\frac{\hat{\beta}_i - \beta_{i,H_0}}{se\left(\hat{\beta}_i\right)} \right)^2$$

Testing Hypotheses on Two or More Coefficients

- Note that according to the formula above, F will be large when *either* t_1^2 is large *or* t_2^2 is large. Thus it will fail to reject only when both t_1^2 and t_2^2 are close to zero. Hence the test is a **joint hypothesis test**.
- The F formula above assumes that X_{2i} and X_{3i} are uncorrelated. If they are correlated, the general formula is

  $$F = \frac{1}{2}\frac{t_2^2 + t_3^2 - 2t_2t_3\rho_{23}}{1-\rho_{23}^2}$$

  where ρ_{23} is the correlation between $\hat{\beta}_2$ and $\hat{\beta}_3$.
- According to Stata, for the joint hypothesis above, $F = 30$. From Excel, the pvalue for this test is

  $$F.DIST.RT\,(30, 2, 4158) = 0$$

  Note that the degrees of freedom in the numerator is 2 because we are testing 2 equalities. Thus we soundly reject the null hypothesis that neither the father's nor the mother's education affects the child's wage.

Testing Restrictions on Two or More Coefficients

- Most statistical programs (including Stata) automatically test whether all the slopes equal zero.

  $$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
  $$H_A : At\, least\, one\, \beta_i \neq 0$$

  Using our wages and education example, the overall F-statistic is 70 with a p-value of 0, so at least one coefficient is not equal to zero.

- Another use of the F-test is to see if coefficients are related linearly. For example, we might want to test if mother's education has the same effect on the wage as the father's.

  $$H_0 : \beta_2 = \beta_3$$
  $$H_A : \beta_2 \neq \beta_3$$

  This test involves a single restriction, so the numerator degrees of freedom equals one. If we run this test in Stata, we get $F = 0.8$ with a p-value of 0.37. Thus we cannot reject the hypothesis that $\beta_2 = \beta_3$.

Omitted Variable Bias in Multiple Regression

- Any variable that
  1. affects the dependent variable,
  2. is correlated with at least one of the independent variables, and
  3. is omitted from the model

  will bias at least one of the coefficients of the model.
- In the two-variable wage model, the variable “ability” affects the dependent variable (wage), is correlated with the independent variable (education), and was omitted from the model. Therefore it biased the coefficient on eduction.
- Sometimes including the omitted variable produces unbiased estimates of *all* the coefficients. In the population of table 17, including both X_1 and X_2 in the model makes the average error term independent of *both* variables.
- But sometimes a variable is included not because it is interesting in itself, but just to reduce the bias in the coefficient on another variable. This is called a **control variable**.

Control Variables

The true values of the coefficients are $\beta_0 = 2$, $\beta_1 = 3$, $\beta_2 = 4$. Note that $v = \beta_2 X_2 + u$

Table 19: Control Variables

Y	X_1	X_2	u	v
10	2	1	-2	2
16	2	1	4	8
16	4	1	-2	2
22	4	1	4	8
28	6	3	-4	8
40	6	3	8	20
34	8	3	-4	8
46	8	3	8	20

Control Variables

Suppose we are *only* interested in obtaining an unbaised estimate of the effect of X_1 on Y; we are *not* interested in the effect of X_2. Can we find the effect of X_1 by estimating the following model?

$$Y_i = \alpha_0 + \alpha_1 X_i + v_i$$

To find out, we need to know $E\left(v_i | X_{1i}\right)$.

$$\begin{aligned}
E\left(v_i | X_{1i} = 2\right) &= \frac{2+8}{2} = 5 \\
E\left(v_i | X_{1i} = 4\right) &= \frac{2+8}{2} = 5 \\
E\left(v_i | X_{1i} = 6\right) &= \frac{8+20}{2} = 14 \\
E\left(v_i | X_{1i} = 8\right) &= \frac{8+20}{2} = 14
\end{aligned}$$

Since $E\left(v_i | X_{1i}\right)$ is not constant, α_1 will be a biased estimate of β_1.

Control Variables

- Now let's include X_2 to estimate

$$Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 X_2 + v_i$$

- Will $\hat{\gamma}_1$ be an unbiased estimate of β_1? To answer this, we need to know $E\left(u_i | X_{1i}, X_{2i}\right)$.

$$E\left(u_i | X_{1i} = 2, X_{2i} = 1\right) = \frac{-2+4}{2} = 1$$
$$E\left(u_i | X_{1i} = 4, X_{2i} = 1\right) = \frac{-2+4}{2} = 1$$
$$E\left(u_i | X_{1i} = 6, X_{2i} = 3\right) = \frac{-4+8}{2} = 2$$
$$E\left(u_i | X_{1i} = 8, X_{2i} = 3\right) = \frac{-4+8}{2} = 2$$

- Note that when $X_{2i} = 1$, $E\left(u_i\right) = 1$ *regardless of the value of* X_1. And when $X_{2i} = 3$, $E\left(u_i\right) = 2$, *regardless of the value of* X_{1i}. Thus, *controlling for* X_{2i}, $E\left(u\right)$ does not depend on X_{1i}, so $\hat{\gamma}_1$ *is* an unbiased estimate of β_1.

Control Variables

- If we use the data above to estimate$\hat{\alpha}_1$ and $\hat{\gamma}_1$, the results are as follows.

$$\hat{Y}_i = 2.5 + 4.8 \cdot X_{1i}$$
$$\hat{Y}_i = 2.5 + 3.0 \cdot X_{1i} + 4.5 \cdot X_{2i}$$

- Note that $\hat{\alpha}_1 \neq \beta_1$, but$\hat{\gamma}_1 = \beta_1$. Thus including X_2 removed the bias in $\hat{\alpha}_1$.
- Note also that $\hat{\gamma}_2 \neq \beta_2$ *but we don't care*, because X_2 is included only as a control, to remove the bias in $\hat{\alpha}_1$. We are not interested in β_2 in this example.

General Method for Nonlinear Functions

- So far we have assumed that the relation between Y and X is a straight line. For example, we have assumed a linear relationship between the worker's age and her wage .

$$E\left(wage_i\right) = \beta_0 + \beta_1 Age_i$$

- Using the dataset acsnyc2018wages, the estimated sample regression function is

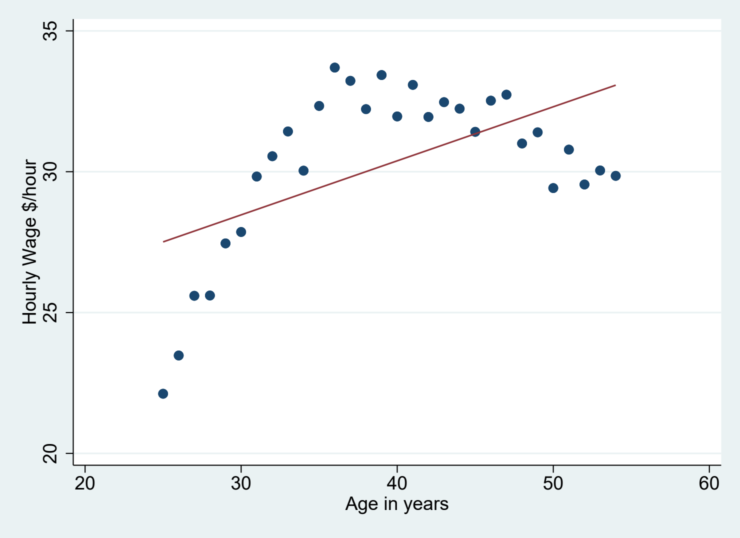
$$\widehat{wage}_i = \underset{(0.4)}{22.7} + \underset{(0.011)}{0.19}\ age_i$$

  This implies that for each additional year of age, the average wage rises by \$0.19 per hour *regardless of age*. For example, 25-year-olds make \$0.19 more per hour than 24-year-olds; 55-year-olds make \$0.19 more per hour than 54-year-olds.

- But perhaps the effect of age on wage (the slope) is not constant.
- Instead it could be **nonlinear**. Consider how well a straight line fits the data in the following graph.

Age and Wages

Figure 29: Average wage by age of worker



Age and Wages

- Obviously a straight line fits the data poorly.
- An improvement might be a **quadratic regression model**.

$$E\left(wage_i\right) = \beta_0 + \beta_1 Age_i + \beta_2 Age^2$$

  The shape of a quadratic function depends on the signs of β_1 and β_2. If $\beta_1 > 0$ and $\beta_2 < 0$, the line will first rise and then fall.

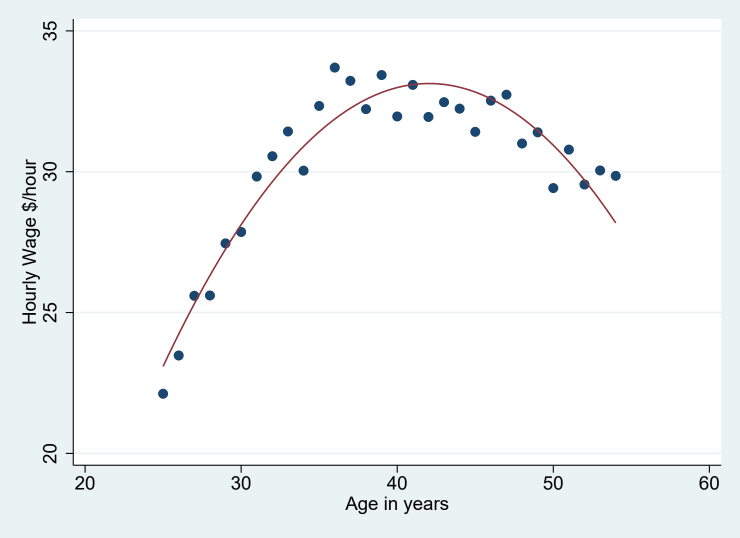
- Using the dataset acsnyc2018wages, we can estimate the model above.

$$\widehat{wage}_i = -\underset{(2.2)}{28} + \underset{(0.12)}{2.9}\, Age_i - \underset{(0.001)}{0.035} Age_i^2$$

- One way to check if the quadratic model improves the fit compared with the linear model is to plot the predicted values from the quadratic model against the actual values.

Age and Wages

Figure 30: Quadratic model of wages and age



Age and Wages

- The quadratic regression function (SRF) appears to fit the data much better than the linear function.
- We can also test statistically whether this is true. One way is to test whether $\beta_2 = 0$ against the alternative hypothesis that $\beta_2 \neq 0$. Let's assume $\alpha = 0.05$. In this example, $n = 58427$.

  1. The sample t-statistic is

  $$t_{sample} = \frac{\hat{\beta}_2 - \beta_{2,H_0}}{se\left(\hat{\beta}_2\right)} = \frac{0.035}{0.001} = 35$$

  2. Calculate the p-value.

  $$t.dist.2t\,(35, 58424) = 0$$

  3. Reject $H_0 : \beta_2 = 0$

- Thus we can safely conclude that the quadratic model fits better than the linear model.

Marginal Effects in Nonlinear Models

- With nonlinear functions, the effect of X on Y *depends on* X.
- For example, suppose we want to find the effect of an additional year of age on the wage of a 25-year-old. We need to make two predictions, one for 25-year-olds and one for 26-year-olds.

$$\begin{aligned}(\widehat{wage}|Age = 25) &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 25 + \hat{\beta}_2 \cdot 25^2 \\ &= -28 + 2.9 \cdot 25 - 0.035 \cdot 25^2 = 22.6 \\ (\widehat{wage}|Age = 26) &= -28 + 2.9 \cdot 26 - 0.035 \cdot 26^2 = 23.7\end{aligned}$$

  Thus one addititional year of age raises the wage of a 25-year-old by $23.7 - 22.6 = 1.1$ dollars per hour.

- Let's do the same calculation for a 55-year-old.

$$\begin{aligned}(\widehat{wage}|Age = 55) &= -28 + 2.9 \cdot 55 - 0.035 \cdot 55^2 = 25.6 \\ (\widehat{wage}|Age = 56) &= -28 + 2.9 \cdot 56 - 0.035 \cdot 56^2 = 24.6\end{aligned}$$

  Thus an additional year of age *reduces* the wage of a 55-year-old by \$1.

Marginal Effects using Calculus

- The marginal effect is defined as

$$\frac{\Delta E\left(Y\right)}{\Delta X}$$

  If the PRF is linear, such as

$$E\left(wage_i\right)=\beta_0+\beta_1 Age_i$$

  The marginal effect is simply β_1.

- But if the model is nonlinear, such as quadratic, the marginal effect is often more complicated, and is generally found with calculus. For example, the marginal effect in the quadratic model is

$$\frac{d\left(E\left(wage_i\right)=\beta_0+\beta_1 Age_i+\beta_2 Age^2\right)}{dAge_i}=\beta_1+2\beta_2 Age$$

Marginal Effects using Calculus

- For a 25-year-old, this is

$$ME\left(25\right) = 2.9 - 2 \cdot 0.035 \cdot 25 = 1.15$$

  which is pretty close to the estimate above of 1.1. It differs because calculus assumes the change in X is extremely small, whereas our earlier calculation assumed a one-unit change in X.
- Note that we cannot interpret coefficients in a quadratic model the same way as in other models. For example, we cannot say that an extra year of age raises the wage by $2.9 dollars per hour, holding Age^2 constant. We cannot hold Age^2 constant if we're changing Age by one year.
- We can interpret quadratic models by describing the shape of the curve as well as by calculating the marginal effect at different values of X. For example, we can say that wages rise until 40 years old and then decline. We can also note that the marginal effect of age for a 25-year-old is about $1, and for a 55-year-old, it's about -$1.

Polynomials

- A quadratic model is an example of a **polynomial function**. The general polynomial function is

$$E\left(Y_i\right) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \cdots + X_i^r$$

  where r is the **order** of the polynomial. The quadratic function is a polynomial of order 2.

- An example of a third-order or cubic polynomial is the **cost function**.

$$C\left(Q\right) = \beta_0 + \beta_1 Q + \beta_2 Q^2 + \beta_3 Q^3$$

  We know that the cost function is cubic because the **marginal cost** function, which is the first derivative of the cost function, is thought to be quadratic.

$$\frac{dC\left(Q\right)}{dQ} = \beta_1 + 2\beta_2 Q + 3\beta_3 Q^2$$

  Since at first MC declines over some range of Q before rising, we know that $\beta_2 < 0$, and $\beta_3 > 0$.

Polynomials

- Let's estimate a cubic cost function using the airlines data for the fourth quarter of 2018. The cost is "operating costs", and the quantity of output is "revenue passenger miles". The results are

$$\widehat{opexpenses}_i = \underset{(62.1)}{32.4} + \underset{(8.3)}{43.7} rpm_i - \underset{(0.166)}{0.129} rpm2_i$$
$$+ \underset{(0.0008)}{0.0019} rpm3_i$$

- Now let's test if the model is non-linear. The model is linear is $\beta_2 = 0$ and $\beta_3 = 0$. Thus

$$H_0 : \beta_2 = 0, \beta_3 = 0$$
$$H_A : \beta_2 \neq 0 \, or \, \beta_3 \neq 0$$

This test gives an $F = 36$ with a pvalue of essentially zero. Thus we can reject a linear relationship between operating costs and quantity produces.

Euler's constant

- Suppose you have $1 in the bank and it earns 100% interest ($i = 1$). Call the $1 the **present value** (**PV**) of your wealth. After one year, the **future value** (**FV**) of your wealth will be the original value plus the interest.

$$FV_{1\,year} = \$1 + \$1 = \$1 \cdot (1+1) = PV \cdot (1+i)$$

- Now suppose that instead of paying you %100 at the end of the year, the bank pays you 50% after the first six months, and 50% after the next six months. After 6 months you have $1.5.

$$FV_{6\,months} = \$1 + \$0.5 = \$1 \cdot \left(1 + \frac{1}{2}\right) = PV \cdot \left(1 + \frac{i}{2}\right)$$

- Over the next six months you earn 50% *on your $150*. This is called **compound interest**.

$$FV_{1\,year} = 100 \cdot 1.5 \cdot 1.5 = PV \cdot \left(1 + \frac{i}{2}\right)^2 = \$2.25$$

Euler's constant

- Now suppose the bank pays you interest every month. Then

$$FV_{1\,year} = PV \cdot \left(1 + \frac{i}{12}\right)^2 = \$1 \cdot \left(1 + \frac{1}{12}\right)^{12} = \$2.61$$

- As the number of periods rises, the future value approaches a specific number, called e, known as Euler's (pronounced "oy-lah") constant.

$$\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = e = 2.718\ldots$$

- The **exponential function** of some number, x, raises e to the power x.

$$exp\,(x) = e^x$$

For example,

$$exp\,(2) = e^2 = 7.39$$

The logarithmic function

- Suppose two numbers, for example 10 and 100, are related as follows

  $$10^2 = 100$$

  Then 2 is the **logarithm** of 100 to the **base** 10. This is written

  $$log_{10}(100) = 2$$

  Thus the $log_b(y)$ is the power to which b must be raised to equal y.
- In general,

  $$b^x = y \iff log_b(y) = x$$
- Thus $log_b(x)$ and b^x are inverse functions. The functions $f(x) = x^2$ and $g(x) = \sqrt{x}$ are inverse functions of each other because $f(g(x)) = (\sqrt{x})^2 = x$. Similarly,

  $$b^{log_b(y)} = y$$

The Natural Logarithm

- The most commonly used base in economics is Euler's constant, e. The logarithm of x to the base e is called the **natural logarithm**, written $ln\left(x\right)$.
- As with any logarithm, raising the base–in this case, e–to the logarithm of x gives you x.

$$e^{ln(x)} = x$$

  That is, $exp\left(x\right)$ and $ln\left(x\right)$ are inverse functions; they undo each other. For example, to what power do you need to raise e to get e^x? This is the question answered by $ln\left(x\right)$.

$$ln\left(e^x\right) = x$$

- The natural log has a number of useful properties.

$$ln\left(a \cdot b\right) = ln\left(a\right) + ln\left(b\right)$$
$$ln\left(a^b\right) = b \cdot ln\left(a\right)$$

The Natural Logarithm and percent change

- The natural logarithm is especially popular in economics because the *change* in the natural logarithm of a variable is approximately equal to the percent change of the variable itself.
- For example, suppose $X_1 = 100$ and $X_2 = 110$. What is the percent change from X_1 to X_2?

$$\%\triangle X = \frac{X_2 - X_1}{X_1} \cdot 100 = \frac{110 - 100}{100} \cdot 100 = 10\%$$

- We get approximately the same answer using the difference in natural logs.

$$(ln\,(110) - ln\,(100)) \cdot 100 = (4.7 - 4.6) \cdot 100 = 10\%$$

The Natural Logarithm and percent change

- The approximation is worse when the percent difference is greater. Suppose $X_1 = 100$ and $X_2 = 150$. Then

$$\%\triangle X = \frac{150 - 100}{100} \cdot 100 = 50\%$$
$$log\,difference = (ln\,(150) - ln\,(100)) \cdot 100 = 40\%$$

- Nonetheless, economists generally prefer the log-difference because the traditional calculation gives a different percent change depending on whether the change is from 100 to 150 or 150 to 100.

$$\frac{100 - 150}{150} = 33\%$$

- The log-difference is almost the same as a different formula for the percent change in which the denominator is the *average* of X_1 and X_2

$$\frac{X_2 - X_1}{\frac{X_1 + X_2}{2}} = \frac{150 - 100}{125} = 0.4$$

Logarithmic regression models

- The Natural logarithm is used in three regression functions. In one model the independent variable is a logarithm but the dependent variable is not.

  $$Y_i = \beta_0 + \beta_1 ln\left(X_i\right) + u_i$$

  This is the **linear-logarithmic** or **lin-log** regression.

- In a second model the dependent variable is a logarithm but the independent variable is not.

  $$ln\left(Y_i\right) = \beta_0 + \beta_1 X_i + u_i$$

  This is the **logarithmic-linear** or **log-lin** regression.

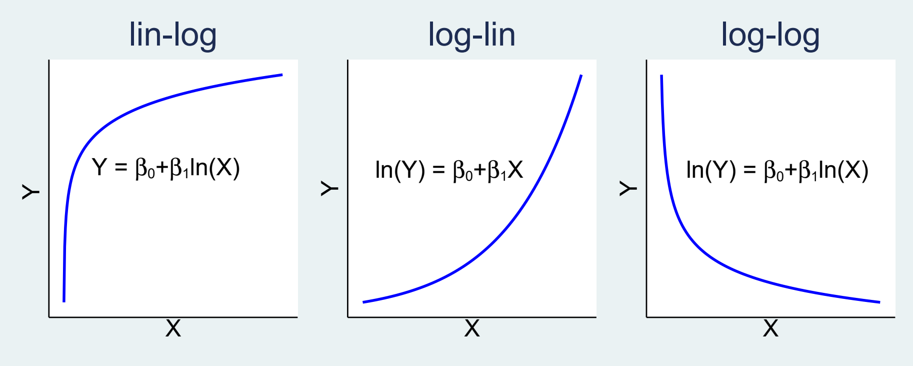
- In the final model both the dependent and independent variables are in logarithms.

  $$ln\left(Y_i\right) = \beta_0 + \beta_1 ln\left(X_i\right) + u_i$$

  This is the **logarithm-logarithm** or **log-log** regression.

Logarithmic regression models

Figure 31: Logarithmic models



The lin-log model

- The lin-log model is useful when we want a function that rises at a declining rate.
- One example is a production function. Suppose output depends on capital and labor.

  $$Q = f\left(K, L\right)$$

  Holding the amount of capital constant, each additional worker increases production, but increases it by less and less. This property is known as "diminishing marginal product".
- Another example is a utility function. Suppose that happiness increases with income, but that happiness increases less and less as income increases.

The lin-log model

- A supplement to the Current Population Survey asked respondents to rate their happiness, sadness, and other moods on a scale from 0 to 6. This dataset is on Blackboard under the name “atusmoods”.
- In this survey happiness appears unrelated to income, but sadness is strongly negatively related. A negative relation shows a downward sloping line. Merely to match the upward-sloping lin-log graph above, define “contentment” as 6-sadness. Thus when sadness = 6 (very sad) contentment = 0 (very discontented), etc. Also, define faminc000 as nfaminc/1000, so that income is measured in $1000 rather than $1.
- Then estimate

$$Contentment_i = \beta_0 + \beta_1 ln\left(income000_i\right) + u_i$$

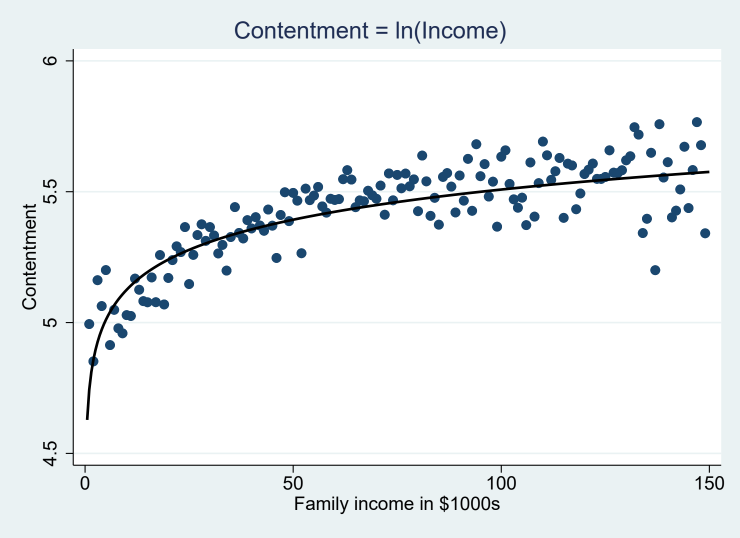
The results are

$$\widehat{contentment} = \underset{(0.3)}{4.7} + \underset{(0.007)}{0.17} \cdot ln\left(income000\right)$$

The actual and predicted values are shown on the next slide.

The lin-log model

Figure 32: Contentment and Income



The lin-log model

- To interpret the results, predict $\hat{Y}$ at two different values of X, then take the difference in the predictions.

$$
\begin{aligned}
\widehat{contentment}_1 &= \hat{\beta}_0 + \hat{\beta}_1 \cdot ln\left(income000_1\right) \\
\widehat{contentment}_2 &= \hat{\beta}_0 + \hat{\beta}_1 \cdot ln\left(income000_2\right) \\
\Delta\widehat{contentment} &= \hat{\beta}_1 \cdot ln\left(income000_2\right) - \hat{\beta}_1 \cdot ln\left(income000_1\right) \\
&= \hat{\beta}_1 \cdot \triangle ln\left(income000\right) \\
&= \frac{\hat{\beta}_1}{100}\left(100 \cdot \triangle ln\left(income000\right)\right) \\
\triangle\widehat{contentment} &= \frac{\hat{\beta}_1}{100} \cdot \%\triangle income
\end{aligned}
$$

- Recall that $100 \cdot \triangle ln\left(x\right)$ is the percent change in X. Thus the interpretation of $\hat{\beta}_1$ is as follows. “A one percent change in income is associated with a 0.0017 point increase in contentment.”

The log-lin model

- The shape of the scatter diagram in Figure 32 suggests that perhaps a quadratic model might work as well as or better than the lin-log model. The results of the quadratic model are

$$\widehat{contentment} = \underset{(0.015)}{5.07} + \underset{(0.0003)}{.007}\, income000 - \underset{(0.000001)}{0.00002}\, income000^2$$

- All the coefficients are statistically significantly different from zero at the 1% level. How do we choose which model to use, the lin-log or the quadratic?
- One way is to judge the plausibility of the results. The lin-log model implies that additional income always increases contentment, even if by only a small amount, whereas the quadratic model implies that contentment actually goes down after a certain income.
- Another way is compare the adjusted R^2s. $\overline{R}^2$s of different models can be compared *as long as the dependent variables are the same,* which is true in this case. $\overline{R}^2 = 0.0171$ in the lin-log model and $\overline{R}^2 = 0.017$ in the quadratic model.

The log-lin model

- the log-linear model is often chosen because the researcher wants to know the *percent change* in $\hat{Y}$ resulting from a unit change in X. For example, using the dataset acsnyc2018wages, the regression

$$\widehat{wage}_i = \underset{(0.39)}{-6.3} + \underset{(0.026)}{2.5}\ highest\,grade$$

  tells us that each year of education raises the predicted wage by $2.5. Is that a lot? A little? It's hard to say without more information.

- The log-lin model tells us the *percent change* in the predicted wage resulting from a one-year increase in highest grade completed (hgc). To see this, calculate $ln\,\widehat{(wage)}$ at two different values of education, hgc_1 and hgc_2, then take the difference.

$$\begin{aligned} ln\,\widehat{(wage_1)} &= \hat{\beta}_0 + \hat{\beta}_1 hgc_1 \\ ln\,\widehat{(wage_2)} &= \hat{\beta}_0 + \hat{\beta}_1 hgc_2 \\ \triangle ln\,\widehat{(wage)} &= \hat{\beta}_1 hgc_2 - \hat{\beta}_2 hgc_1 = \hat{\beta}_1 \triangle hgc \end{aligned}$$

The log-lin model

- Multiply both sides by 100.

$$100 \cdot \triangle ln\,\widehat{(wage)} = 100 \cdot \hat{\beta}_1 \triangle hgc$$

  Recall that $100 \cdot ln\,(Y)$ is the percent change in Y.

- Suppose $\triangle hgc = 1$. Then a one-year increase in grade is associated with a change in the predicted wage of $100 \cdot \hat{\beta}_1\%$.
- Using the dataset acsnyc2018wages gives

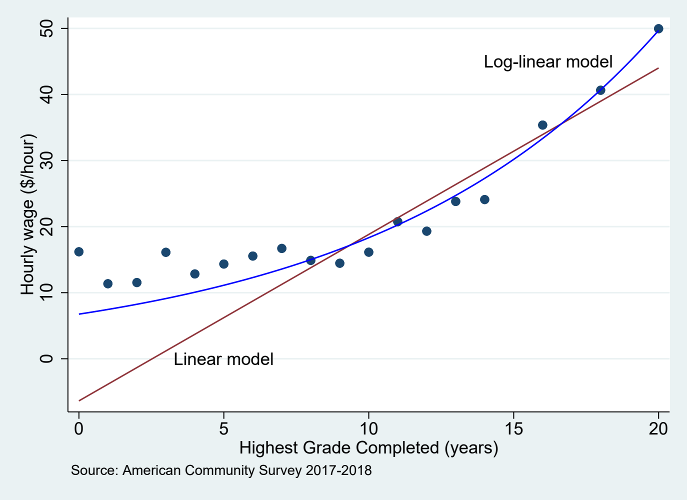
$$ln\,\widehat{(wage_i)} = \underset{(0.01)}{1.9} + \underset{(0.0007)}{0.09}\; highest\, grade_i$$

  Thus for each extra year of education, the predicted wage rises by $0.09 * 100\%$ or 9%.

- Another reason to prefer the log-lin model over the lin-lin model is that it may fit the data better. Consider the following graph of both models against the data.

The log-lin versus lin-lin models

Figure 33: The log-lin versus lin-lin model of wages and education



Source: American Community Survey 2017-2018

The log-lin model when X is a dummy

- The lin-log model can be estimated just as easily when X is a dummy variable as when it is continuous. Suppose the model is

$$ln\left(wage_i\right) = \beta_0 + \beta_1 \cdot female_i + u_i$$

Using the acsnyc2018wages data we get

$$ln\widehat{\left(wage_i\right)} = \underset{(0.004)}{3.2} - \underset{(0.02)}{0.15}\, female_i$$

- Now let's make two predictions, one when $female = 0$ and another when $female = 1$, then take the difference.

$$ln\widehat{\left(wage_0\right)} = 3.2 - 0.15 \cdot 0$$
$$ln\widehat{\left(wage_1\right)} = 3.2 - 0.15 \cdot 1$$
$$\triangle ln\widehat{\left(wage\right)} = -0.15$$

The log-lin model when X is a dummy

- Multiply both sides by 100.

$$100 \cdot \triangle ln\,\widehat{(wage)} = -0.15 \cdot 100$$

- This says that on average women make 15% less than men.
- Note that we cannot say, "For each additional unit of women..." Womenhood is not measured in units. Dummy variables must be interpreted as $\hat{Y}$ when the dummy equals 1 compared with $\hat{Y}$ when the dummy equals zero.
- Suppose we estimate the effect of marital status on wage. The omitted catagory is "married".

$$\begin{aligned} ln\,\widehat{(wage_i)} &= \beta_0 + \beta_1 separated_i + \beta_2 divorced_i \\ &\quad + \beta_3 widowed_i + \beta_4 never\,married_i + u_i \end{aligned}$$

The log-lin model when X is a dummy

- Using the acsnyc2018wages dataset, the results are as follows.

$$\widehat{lnwage}_i = \underset{(0.004)}{3.23} - \underset{(0.018)}{0.29}\, separated_i - \underset{(0.012)}{0.11}\, divorced_i$$
$$- \underset{(0.037)}{0.128} widowed_i - \underset{(0.006)}{0.117} nevermarried_i$$

- Since the omitted category is “married”, all coefficients are the percent difference from the predicted wage of married persons. Thus the predicted wage of people who are separated is 29% lower than married persons.
- The predicted wage of divorced persons is 11% percent lower than that of married persons.
- The predicted wage of widowed persons is 13% less than that of married persons.
- The predicted wage of never-married persons is 12% less than that of married persons.
- Note that in each case, the comparison is with the omitted category, married persons.

The log-log model.

- To interpret the log-log model, make two predictions, one at $X = X_1$ and another at $X = X_2$, then take the difference and multiply both sides by 100.

$$\begin{aligned}
\widehat{ln\left(Y_1\right)} &= \hat{\beta}_0 + \hat{\beta}_1 ln\left(X_1\right) \\
\widehat{ln\left(Y_2\right)} &= \hat{\beta}_0 + \hat{\beta}_1 ln\left(X_2\right) \\
\triangle ln\left(Y\right) &= \hat{\beta}_1 \triangle ln\left(X\right) \\
100 \cdot \triangle ln\left(Y\right) &= \hat{\beta}_1 \cdot 100 \cdot \triangle ln\left(X\right) \\
\% \triangle Y &= \hat{\beta}_1 \% \triangle X
\end{aligned}$$

- Thus

$$\hat{\beta}_1 = \frac{\% \triangle Y}{\% \triangle X}$$

which is the *elasticity* of Y with respect to X.

The log-log model

- As an example of a log-log model, consider the effect on a person's current wage of her parent's income when she was in high school.

$$ln\left(wage_i\right) = \beta_0 + \beta_1 ln\left(parental\, income_i\right) + u_i$$

- Using the nlsy97wages2015 dataset gives

$$\widehat{ln\left(wage\right)} = \underset{(0.013)}{2.8} + \underset{(0.01)}{0.17} ln\left(family\, income\right)$$

  Thus a one percent increase in parental income is associated with an increase in current wage of 0.17%.

- In economic terms, the elasticity of current wage with respect to parental income is 0.17. This elasticity is often used to measure inter-generational income mobility. A high elasticity means there is less income mobility, because it implies that parental and child incomes are highly correlated.

Summary of log models

1. Lin-log model

$$Y_i = \beta_0 + \beta_1 ln\left(X_i\right) + u_i$$

   A one-percent change in X is associated with a change in $E\left(Y\right)$ of $\beta_1/100$ units-of-Y.

2. Log-lin model

$$ln\left(Y_i\right) = \beta_0 + \beta_1 X_i + u_i$$

   A one-unit change in X is associated with a change in $E\left(Y\right)$ of $100 \cdot \beta_1$%.

3. Log-log model

$$ln\left(Y_i\right) = \beta_0 + \beta_1 ln\left(X_i\right) + u_i$$

   A one-percent change in X is associated with a change in $E\left(Y\right)$ of β_1 percent.

Interactions between two binary variables

- The log-lin model of wages and gender showed that *on average* women make 8% less per hour than men. But that is an *average* over all women. Perhaps the wage difference between *married* women and men contrasts with the wage difference between *unmarried* women and men. To test this hypothesis we can **interact** a dummy for married with a dummy for female.
- An **interaction** term is simply a product of two variable. For example,

$$wage_i = \beta_0 + \beta_1 female_i + \beta_2 married_i + \beta_3 female_i \cdot married_i + u_i$$

- To understand this model, we need to examine $E\left(wage\right)$ for each combination of $female$ and $married$.

$$E\left(wage|female = 1, married = 1\right) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 1 + \beta_3 \cdot 1 \cdot 1$$
$$E\left(wage|female = 1, married = 0\right) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 + \beta_3 \cdot 1 \cdot 0$$
$$E\left(wage|female = 0, married = 1\right) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 1 + \beta_3 \cdot 0 \cdot 1$$
$$E\left(wage|female = 0, married = 0\right) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0 \cdot 0$$

Interactions between two binary variables

- Thus

$$
\begin{aligned}
E\left(wage|female=1,married=1\right) &= \beta_0+\beta_1+\beta_2+\beta_3 \\
E\left(wage|female=1,married=0\right) &= \beta_0+\beta_1 \\
E\left(wage|female=0,married=1\right) &= \beta_0+\beta_2 \\
E\left(wage|female=0,married=0\right) &= \beta_0
\end{aligned}
$$

- To interpret a coefficient, take the difference between two $E\left(wage\right)$s that isolate that coefficient.
- For example, subtract the fourth equation from the second.

$$
\begin{aligned}
&E\left(wage|female=1,married=0\right) \\
-&E\left(wage|female=0,married=0\right)=\beta_1
\end{aligned}
$$

Thus β_1 is the difference between $E\left(wage\right)$ of an unmarried woman and an unmarried man.

Interactions between two binary variables

- Similarly, subtract the fourth equation from the third.

$$\begin{aligned} &E\left(wage|female=0,married=1\right)\\ -&E\left(wage|female=0,married=0\right)=\beta_2 \end{aligned}$$

  Thus β_2 is the difference in $E\left(wage\right)$ between a married man and an unmarried man.

- Now subtract the second from the first equation.

$$\begin{aligned} &E\left(wage|female=1,married=1\right)\\ -&E\left(wage|female=1,married=0\right)=\beta_2+\beta_3 \end{aligned}$$

  Thus $\beta_2+\beta_3$ is the difference in $E\left(wage\right)$ between a married woman and an unmarried woman.

Interactions between two binary variables

- Now subtract the difference between married and unmarried men (β_2) from the difference between married and unmarried women ($\beta_2 + \beta_3$). Thus β_3 shows the *difference* in the effect of being married for women compared with men.
- Let's estimate this model using nlsy97wages2015.

$$\widehat{wage}_i = \underset{(0.5)}{20.9} - \underset{(0.7)}{2.3}\, female + \underset{(0.8)}{6.8}\, married - \underset{(1.1)}{3}\, female \cdot married$$

- Thus unmarried women make \$2.3 per hour less than married women; married men make \$6.8 per hour more than unmarried men; and the wage benefit of being married is \$3 less per hour for women than for men.

Interactions between a binary and a continuous variable

- So far we have assumed that the effect of education on wage is the same for women as for men. This need not be true. To allow for this, we interact years of education with a dummy for female.

$$E\left(wage_i\right) = \beta_0 + \beta_1 Educ_i + \beta_2 Female_i + \beta_3 Educ_i \cdot Female_i$$

- If the person is male, $female = 0$, and

$$E\left(wage_i\right) = \beta_0 + \beta_1 Educ_i$$

  Thus for each additional year of education, a man's expected wage changes by β_1; with no education, his expected wage is β_0.
- If the person is female, $female = 1$, and

$$\begin{aligned} E\left(wage_i\right) &= \beta_0 + \beta_1 Educ_i + \beta_2 + \beta_3 Educ_i \\ &= \beta_0 + \beta_2 + \left(\beta_1 + \beta_3\right) Educ_i \end{aligned}$$

  Thus for each additional year of education, a woman's expected wage changes by $\beta_1 + \beta_3$; with no education, her expected wage is $\beta_0 + \beta_2$.

Interactions Between a Binary and a Continuous variable

- Let's estimate the model above using nlsy97wages2015. Instead of "wage" let's use the natural log of the wage.

$$\widehat{ln\left(wage\right)} = \underset{(0.06)}{1.97} + \underset{(0.004)}{0.073} Educ_i - \underset{(0.09)}{0.43} Female_i + \underset{(0.006)}{0.015} Educ_i \cdot Female_i$$

- Thus for a male, an extra year of education is associated with an increase in the predicted wage of 7.3%.
- For a female, however, an extra year of education is associated with an increase in the predicted wage of

$$100 \cdot \left(\hat{\beta}_2 + \hat{\beta}_3\right) = 100 \cdot (0.073 + 0.015) = 8.8\%$$

The return to education is higher for women than for men, which may be one reason women go to college at higher rates than men.

Interactions Between Two Continuous Variables

- Suppose we suspect that the effect of education depends on years of work experience.

$$E\left(wage_i\right) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + \beta_3 Educ_i \cdot Exper_i$$

- First we find the marginal effect of education, holding experience constant. We need two expectations, one where $educ = educ_1$ and the other where $educ = educ_1$. For both, we keep $exper = exper_1$.

$$E\left(wage_1\right) = \beta_0 + \beta_1 Educ_1 + \beta_2 Exper_1 + \beta_3 Educ_i \cdot Exper_1$$
$$E\left(wage_2\right) = \beta_0 + \beta_1 Educ_2 + \beta_2 Exper_1 + \beta_3 Educ_2 \cdot Exper_1$$

Take the difference.

$$\begin{aligned} \triangle E\left(wage\right) &= \beta_1 \triangle Educ + \beta_3 \triangle Educ \cdot Exper_1 \\ &= \left(\beta_1 + \beta_3 Exper_1\right) \triangle Educ \end{aligned}$$

Interactions Between Two Continuous Variables

- Hence a one-year increase in education is associated with change in $E\left(wage\right)$ of $\left(\beta_1 + \beta_3 Exper_1\right)$.
- Let's estimate this model using nlsy97wages2015 and $ln\left(wage\right)$ as the dependent variable.

$$\widehat{lnwage}_i = \underset{(0.11)}{1.1} + \underset{(0.008)}{0.09}\ Educ_i + \underset{(0.004)}{0.03}\ Exper_i - \underset{(0.0003)}{0.0008} Educ_i \cdot Exper_i$$

- Thus the marginal effect of education is

$$ME\left(Educ_i\right) = \left(0.09 - 0.0008 \cdot Exper_i\right)$$

The marginal effect of education appears to decline with experience.

Binary Dependent Variables

- All the dependent variables we have looked at so far have been continuous, such as the hourly wage or minutes spent watching TV.
- But binary variables can also be dependent variables. For example, suppose a person applies for a loan from a bank to buy a house. We can define the dependent variable, call it $Approved$, as equal to one if the bank approves her loan and zero otherwise.
- A statistical model cannot predict exactly whether the outcome is a one or a zero, but only the *probability* that it is one or zero.
- For example, suppose the probability that a bank will approve a loan to buy a house depends on the borrower's income (Inc) and the value of the house (V). Then

$$P\left(Approve_i = 1\right) = F\left(\beta_0 + \beta_1 Inc_i + \beta_2 V_i\right)$$

where $F\left(\beta_0 + \beta_1 Inc_i + \beta_2 V_i\right)$ is some cumulative probability function.

Linear Probability Model

- The simplest cumulative probability function is linear.

$$P\left(Approve_i = 1\right) = \beta_0 + \beta_1 Inc_i + \beta_2 V_i$$

This is called the **linear probability model** or **LPM**.

- To understand the LPM, recall the standard OLS population regression function (PRF).

$$E\left(Y_i\right) = \beta_0 + \beta_1 X_i$$

In a linear probability model, $E\left(Y_i\right)$ has a particular meaning.

- Suppose we have a population of four applicants. Also, suppose outcomes of the four applications are $Approved = (1, 0, 0, 1)$. What is the numerical value of $E\left(Approved\right)$?

$$E\left(Approved\right) = \frac{1+0+0+1}{4} = 0.5$$

Linear Probability Model

- The key point is that $E\left(Approved\right) = 0.5 = proportion\left(Approved = 1\right)$. The average value of a binary variable equals the proportion of times the variable equals 1.
- Recall the relationship between proportion and probability.

$$Probability\left(Y = 1\right) = 100 \cdot Proportion\left(Y = 1\right)$$

- Thus

$$Probability\left(Y = 1\right) = 100 \cdot E\left(Y = 1\right)$$

- Suppose the model is

$$\begin{aligned} E\left(Approved_i\right) &= \beta_0 + \beta_1 Inc_i + \beta_2 V_i \\ 100 \cdot E\left(Approved_i\right) &= 100 \cdot \left(\beta_0 + \beta_1 Inc_i + \beta_2 V_i\right) \\ Prob\left(Approved_i = 1\right) &= 100 \cdot \left(\beta_0 + \beta_1 Inc_i + \beta_2 V_i\right) \end{aligned}$$

Linear Probability Model

- Let's make two predictions of the probability of approval, one when $Income = Income_1$ and another when $Income = Income_2$.

$$P\left(Approved_i = 1 | Inc = Inc_1\right) = 100 \cdot \left(\beta_0 + \beta_1 Inc_1 + \beta_2 V_1\right)$$
$$P\left(Approved_i = 1 | Inc = Inc_2\right) = 100 \cdot \left(\beta_0 + \beta_1 Inc_2 + \beta_2 V_1\right)$$

- Now let's take the difference. Note that both β_0 and $\beta_2 V_1$ cancel out.

$$\triangle P\left(Approved_i = 1\right) = 100 \cdot \beta_1 \triangle Inc$$

  Thus a one-dollar increase in $Income$ is associated with a change in the probability that $Approve = 1$ of $\beta_1 \cdot 100$ *percentage points*.

- The distinction between *percent* and *percentage points* is crucial. Suppose the unemployment rate rises from 5% to 10%. That's an increase of 100% but of 5 *percentage points*.

Linear Probability Model, example

- Let's estimate the model above using data on loan applications in New York City in 2018. The dependent variable is $Approved$, which equals 1 if the loan was approved, and zero otherwise. Income is measured in thousands of dollars and property value in millions of dollars.

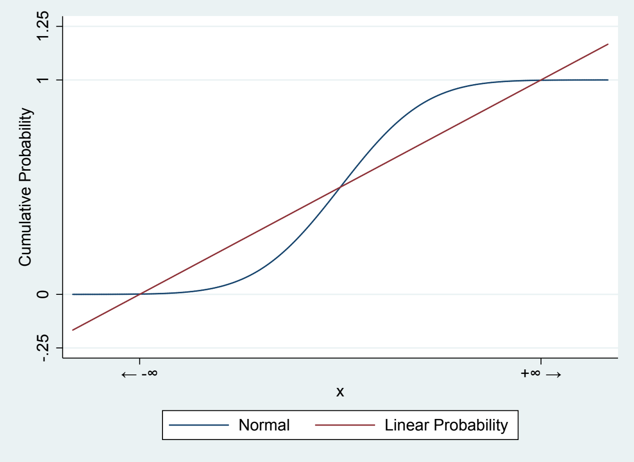
$$\widehat{Approved}_i = \underset{(0.003)}{0.8} + \underset{(0.005)}{.047}\, Income_i - \underset{(0.002)}{0.009} Value_i$$

- Thus an increase in borrower income of \$1000 is associated with an increase in the probability that the bank approves the loan of 4.7pp.
- Similarly, an increase in the value of the house by \$1 million reduces the probability that the bank approves the loan by 0.9pp.
- If the borrower has no income and the house is worth nothing, the probability of approval is 80%.

Drawbacks of LPM

The linear probability model (LPM) has several drawbacks, which can be seen in the following graph.

Figure 34: Linear versus Normal Probabilities



Drawbacks of LPM

- Recall that with the cumulative Normal function, the maximum probability is one, and the minumum probability is zero. For example, using normsdist(z) in Excel, no matter what number you use for z, negative one billion or positive one billion, the result will always be a number between zero and one.
- But this is not true of the linear probability function. Using our estimated LPM above, the probability of approval rises with income. If income is high enough, there is nothing stopping the predicted probability from exceeding one.
- This can be seen in Figure 34. The LPM is the line in red. Since the LPM is *linear*, the predicted probability exceeds one towards the right-side of the diagram, and falls below zero towards the left-side.
- The cumulative Normal function, shown in blue, does not have this problem. As the predicted probability approaches either one or zero, it flattens out so that it stays within those bounds. A model based on the cumulative Normal function is called a **probit**.

Probit

- The probit model is defined as follows.

$$P\left(Approve_i = 1\right) = \Phi\left(\beta_0 + \beta_1 Inc_i + \beta_2 V_i\right)$$

  where $\Phi\left(z\right)$ is the standard cumulative Normal function. In Excel, it is normsdist(z).

- We will not go into the details of estimating the probit, instead relying on Stata, in which the command is simply `probit y x`.
- Let's estimate a probit model of loan approval using the same data as above.

$$prob\left(\widehat{Approved_i} = 1\right) = \Phi\left(\underset{(0.01)}{0.82} + \underset{(0.044)}{0.64}\, Income_i - \underset{(0.088)}{0.087} Value_i\right)$$

- Let's calculate the probability of approval for someone with $1,000 trying to buy a $1 million property.

Probit

- The predicted probability is

$$prob\left(\widehat{Approved}_i = 1\right) = \Phi\left(\underset{(0.01)}{0.82} + \underset{(0.044)}{0.64} \cdot 1 - \underset{(0.088)}{0.087} \cdot 1\right)$$
$$= \Phi\left(1.373\right) = normsdist\left(1.373\right) = 0.92$$

- Now let's predict the probability of approval for someone with \$2000 and a property of the same value as above.

$$prob\left(\widehat{Approved}_i = 1\right) = \Phi\left(\underset{(0.01)}{0.82} + \underset{(0.044)}{0.64} \cdot 2 - \underset{(0.088)}{0.087} \cdot 1\right)$$
$$= \Phi\left(2.013\right) = normsdist\left(2.013\right) = 0.98$$

- Thus the **marginal effect** of an extra \$1000 in income on the probability of approval is 0.98-0.92 = 0.06. This is reasonably close the coefficient on income in the linear probability model.