Lecture 4

Simple Linear Regression

Outline

- Simple linear regression for population and sample
- Ordinary least square
- Statistical inference in regression
- R^2
- Omitted Variable Bias (OVB)

REMEMBER THIS

Our core statistical model is

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

1. β_1, the slope, indicates how much change in Y (the dependent variable) is expected if X (the independent variable) increases by one unit.
2. β_0, the intercept, indicates where the regression line crosses the Y-axis. It is the value of Y when X is zero.
3. β_1 is almost always more interesting than β_0.

How to move from population to sample?

- We start with some theory where $Y = \beta 0 + \beta 1 X + \varepsilon$
- Through the use of data we get to $Y = \widehat{\beta_0} + \widehat{\beta_1} X + e$
- The first equation is purely theoretical while the second is the empirical counterpart
- How to move from the first equation to the second?
- In other words, how do we get $\widehat{\beta_0}, \widehat{\beta_1}$ based on our sample of X,Y?
    - Ordinary Least Squares (OLS)

Notations of simple linear regression(sample)

- Since the population is too big, we use a sample to obtain the estimated regression line:
- Sample analogue: $Y = \widehat{\beta_0} + \widehat{\beta_1} X + e$
- Defining terms:
  - Y, X are actual data
  - $\widehat{\beta_0}$is the estimate for β0
  - $\widehat{\beta_1}$is the estimate for β1
  - $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are sample statistics, they are also called regression coefficients
  - e is the estimate for ε, it is called the residual term
  - $\widehat{\beta_0} + \widehat{\beta_1} X$ is the predicted value of Y, sometimes we call it $\hat{Y}$

How to draw the regression line?

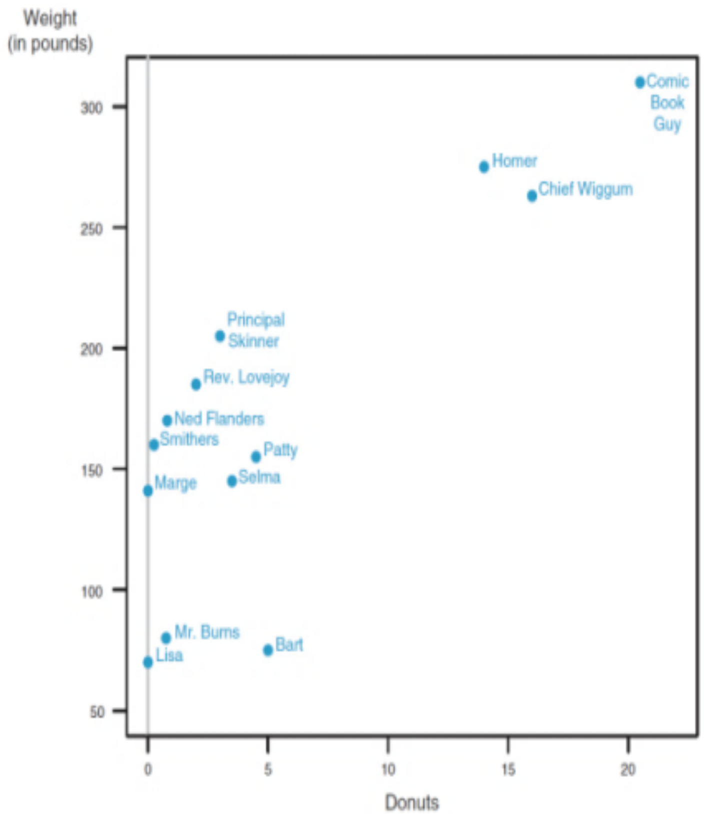
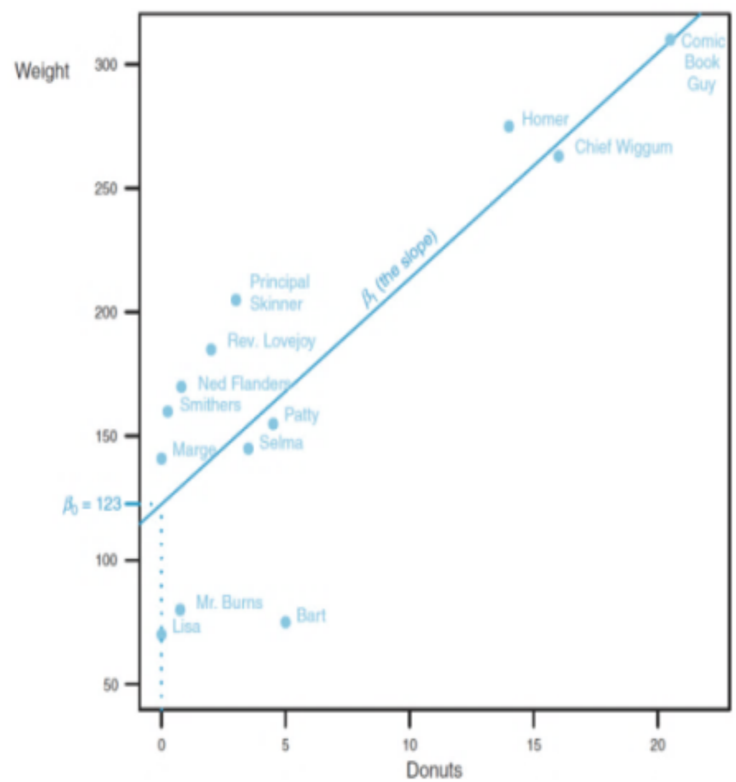


FIGURE 1.2: Weight and Donuts in Springfield



Ordinary least square (OLS)

- **Ordinary Least Squares** (OLS) is one of the simplest methods of linear regression. The goal of OLS is to closely fit a function with the data. It does so by minimizing the sum of squared residuals from the data.

- So we noticed e is the deviation from actual Y and predicted Y. We'd like to find a way to minimize e for every data point
- How about the one minimized

$$\sum e_i = \sum (Y_i - \hat{Y}_i)$$

- How about the one minimized

$$\sum | e_i | = \sum | Y_i - \hat{Y}_i |$$

This is a reasonable candidate but it's not easy to calculate the absolute value.

Interpretation of $\widehat{\beta_0}$, $\widehat{\beta_1}$ and making predictions

- $\widehat{\beta_1}$: How **<u>mean</u>** in Y **changes** when X **changes** one unit
- $\widehat{\beta_0}$: The **<u>mean</u>** in Y when X is 0
- Question: why do I have to say **mean** every time?
  - We can also use "predicted value" interchangeably with the word "mean" here
- $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are estimates of β_0 and β_1

- Once we know $\widehat{\beta_0}$ and $\widehat{\beta_1}$, we can write out the regression model as $Y = \widehat{\beta_0} + \widehat{\beta_1}X + e$ or $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X$. Now for a given X, we will be able to make prediction about Y.

When does OLS give us true effect?

- $Y=\beta_0+\beta_1 X+\varepsilon$
- ε includes all factors that can explain Y other than X.
- In this example, what are some other factors that can explain age at first marriage other than education?
- **If we assume ε is randomly distributed, then β_1 offers true causal relationship of X on Y**. But is it likely to be true in real research projects?
- To rephrase, is it likely to be true that all other factors that affects *time to marry* are uncorrelated with *education*?
  - Note: I am not saying *education* is the only factor affecting *time to marry*. There can be other factors as long as they do NOT correlate with education (remember apples vs oranges- more on this later).

Example 1 (numerical variable):

- If X is class size, and Y is student test score. What does -0.04 and 14 mean in regression $\hat{Y}$=14-0.04X?
  - -0.04: One person increase in class size will decrease **average** test score by 0.04 unit.
  - 14: if class size is 0, average student test score is 14-It doesn't make much sense in this case.
- What is the predicted test score for a class of 10 students?
  - Plug X=10 into the equation, $\hat{Y}$=13.6.

Exercise 1 (Do on your own)

- Assume you are interested in knowing whether higher educated parents usually have higher educated children. You run a regression of children's highest grades completed on their father's highest grades completed. A regression result shows $\hat{Y}$=5.1+0.2X. How would you interpret the coefficients? What is the average highest grades completed for a child whose father's highest grade completed is 5?

Example 2 (categorical variable)

- What if X is categorical variable such as gender?
- We will have to define X=0 if male; X=1 if female. Assume the research questions is how gender may affect wage
- What does 2800 and -300 mean in linear regression $\hat{Y}$=2800-300X?

- $\hat{Y}_{male}$=2800-300*0=2800
  - Average (or predicted) wage for male is 2800
- $\hat{Y}_{female}$ =2800-300*1=2500
- -300=2500-2800=$\hat{Y}_{female} - \hat{Y}_{male}$:The difference in **average** wage between female and male. Negative sign means female on average has a lower wage

Exercise 2 (categorical variable: Do yourself)

- After estimating whether having health insurance will increase number of doctor visit, you get a regression results: $\hat{Y}$=4.2+0.09X. Interpret $\widehat{\beta_0}\ and\ \widehat{\beta_1}$

Statistical inference in regression

- So far we've been focused on the economic significance-β. Now let's talk about statistical significance
- Why do we need statistical inference in regression?
  - We know the mean of one sample may not be representative of the population mean-because of standard error
  - Same idea here: the $\widehat{\beta_0}$ and $\widehat{\beta_1}$ from one sample may not be representative of true population parameters $\beta_0 and\ \beta_1$
  - With different samples we will have different $\widehat{\beta_0}$ and $\widehat{\beta_1}$. So $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are random variables and follow sampling distributions.
  - We need to incorporate the idea of standard error of $\hat{\beta}$ (sampling variability) in statistical inference

- We can use p-value and confidence interval approach to perform hypothesis test of β

Statistical inference in regression

Years of schooling(X)	Age at first marriage(Y)
16	25
14	30
18	26
10	22
12	29
12	33
16	32
8	24
7	27
18	35
12	19
8	18
16	29
9	21
10	18
16	35
8	20
9	23
16	27
12	23

- Y=$\beta_0 + \beta_1$X+ε
- We are interested in finding a relationship between X and Y
- What is my null hypothesis?
  - H0: β_1=0
  - Ha:β_1≠0
- If we later reject H0, β_1 is not zero. It means we identify a relationship between X and Y (with certain confidence)
- If we fail to reject H0, β_1 can be zero. It means we fail to identify a relationship between X and Y (with certain confidence)
- Note: 1) we usually care less about β_0

  2) STATA will provide the estimate and statistical inference output, but we need to know what's happening behind.

Example (p-value approach)

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Years of schooling (X)	$\boxed{0.9598}$	$\boxed{0.2680}$	3.5812	0.0021	0.3967	1.5229
Intercept	13.9463	3.4443	4.0491	0.0008	6.7102	21.1825

$\hat{\beta}_1$ $\widehat{se(\beta_1)}$

- $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X \Rightarrow \hat{Y}=13.9463+0.9598X$
- Hypothesis test:
  - H_0: $\beta_1=0$
  - H_a: $\beta_1\neq 0$
  - $t^*=\frac{\hat{\beta}_1-0}{\widehat{se(\beta_1)}}=\frac{\hat{\beta}_1}{\widehat{se(\beta_1)}}=\frac{0.9598}{0.2680}=3.5813$ (recall previously we used the formula for $t^*=\frac{\bar{X}-\mu}{se}$)
  - degree of freedom d.f. =n-k-1, where k is number of independent variables.
  - So d.f.=20-1-1=18, p(t>t*) for d.f=18 is 0.0011 P-value=2P(t>t*)= 0.0022<0.05 (α)
  - Reject $H_0 \Rightarrow$ There is a relationship between education and age at first marriage
  - Compare our calculation with software output

Steps for performing test of significance on regression coefficients

- Step1: set up H_0, H_a
  - H_0: β_1=0
  - H_a:β_1≠0
- Step2: calculate t-stat
  - t*=$\frac{\hat{\beta}_1}{\widehat{se}}$ ($\widehat{\beta_1}$, $\widehat{se}$ are information estimated from our sample)
- Step3: translate t* into p-value using t-table
- Step4: compare p-value with α. If p< α, reject H_0, β_1≠0. There is a relationship between X and Y.

Example (confidence interval approach)

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Years of schooling(X)	0.9598	$\boxed{0.2680}$	3.5812	0.0021	0.3967	1.5229
Intercept	13.9463	3.4443	4.0491	0.0008	6.7102	21.1825

- Confidence interval:
  - Formula for the CI=$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}} *$se (recall previously CI=$\bar{X} \pm t_{\frac{\alpha}{2}} * se$)
  - Find $t_{\frac{\alpha}{2}}$ for α=0.05 and d.f.=18. $t_{\frac{\alpha}{2}}$=2.101
  - Margin of error: m=$t_{\frac{\alpha}{2}} * se$= $2.101 * 0.2680$=0.5631
  - CI=$\hat{\beta}_1 \pm \mathrm{m}$=0.9598$\pm$ $0.5631 = (0.3967{,}1.5229)$
  - 95% of time the true relationship between X and Y will fall within the interval $(0.3967{,}1.5229)$. Since 0 does not fall within this interval, we reject the null hypothesis and conclude there is a statistically significant relationship between X and Y

Exercise (p-value approach: Do yourself)

- Suppose we are interested in whether higher GDP can predict higher life expectancy, we ran a regression of life expectancy on GDP. The regression output is listed below:

	Coefficients	standard error	t Stat	P-value	Lower 95%	Upper 95%
GDP	0.0005	0.0006	0.8333	0.4242	-0.0008	0.0018
Intercept	63.3158	0.8629	73.3786	0.0000	61.6104	65.0212

- Write out the estimated regression.

- Manually perform test of significance on the coefficient for GDP using p-value approach (use d.f.=12, $\alpha = 0.05$). Compare your t-stat and P-value with the STATA output

Exercise (Do yourself and ask questions if needed)

- Suppose we are interested in whether higher GDP can predict higher life expectancy. So I ran a regression of life expectancy on GDP. The regression output is listed below:

	Coefficients	standard error	t Stat	P-value	Lower 95%	Upper 95%
GDP	0.0005	0.0006	0.8333	0.4242	-0.0008	0.0018
Intercept	63.3158	0.8629	73.3786	0.0000	61.6104	65.0212

- Manually perform test of significance on coefficient for GDP using confidence interval approach (use d.f.=12, $\alpha = 0.05$). Compare your CI with the STATA output. Is the conclusion consistent with the previous exercise?
- Do you think there is strong evidence that there is positive correlation between GDP and life expectancy? Why or why not?

Read STATA output

- Model: Health=$\beta_0+\beta_1$Age+ε
- codes: reg phstat age_yrs
- (reg: command of regression; phstat: physical status-dependent variable; age_yrs: age measured in years-independent variable)
- note: phstat is an index ranging from 1-5. 1 means excellent health, and 5 means worst health

STATA Output

What does coefficient of age_yrs mean?

Overall fit and model comparison

```
. reg phstat age_yrs

      Source |       SS           df       MS      Number of obs   =     46950
-------------+----------------------------------   F(1, 46948)     =    476.00
       Model |  611.927847         1  611.927847   Prob > F        =    0.0000
    Residual |  60354.2942    46,948  1.28555624   R-squared       =    0.0100
-------------+----------------------------------   Adj R-squared   =    0.0100
       Total |   60966.222    46,949  1.29856274   Root MSE        =    1.1338

------------------------------------------------------------------------------
      phstat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_yrs |   .0198173   .0009083    21.82   0.000      .018037    .0215977
       _cons |   1.426175   .0579778    24.60   0.000     1.312538    1.539813
------------------------------------------------------------------------------
```

Coefficients and statistical inference

Variance decomposition

- It's the top left panel of STATA output
- Variance decomposition:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i e_i{}^2$$

TSS = ESS + RSS

- TSS is Total Sum of Squares
  - Measures how spread out Y_i is in the sample
- ESS is Explained Sum of Squares
  - Stata calls this Model Sum of Squares
- RSS is the Residual Sum of Squares
  - It's the part that cannot be explained by our model

MS is mean sum of squares-
the SS divided by degree of freedom

Source	SS	df	MS
Model	611.927847	1	611.927847
Residual	60354.2942	46948	1.28555624
Total	60966.222	46949	1.29856274

- Total degree of freedom=N-1
- Model degree of freedom=K, where K=number of coefficient excluding the constant term
- Residual degree of freedom= Total-Model

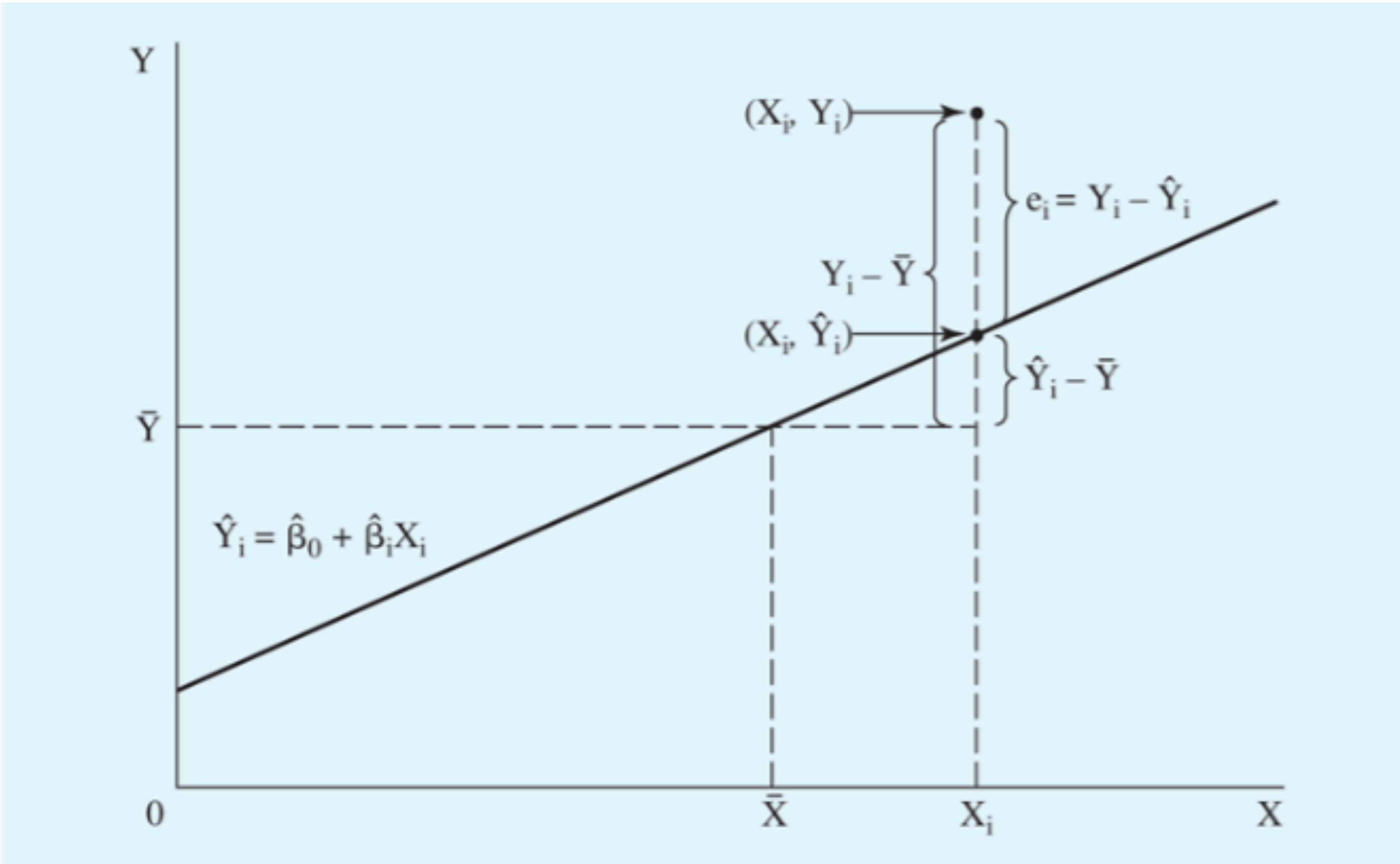


Figure 2.3 Decomposition of the Variance in Y

The variation of Y around its mean $(Y - \overline{Y})$ can be decomposed into two parts: (1) $(\hat{Y}_i - \overline{Y})$, the difference between the estimated value of $Y(\hat{Y})$ and the mean value of Y $(\overline{Y})$; and (2) $(Y_i - \hat{Y}_i)$, the difference between the actual value of Y and the estimated value of Y.

R-square: how well did our line fit the data?

- The most widely used measure of fit is the goodness of fit, R^2
- $\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i {e_i}^2$
- TSS = ESS + RSS
- $R^2 = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i {e_i}^2}{\sum_i (Y_i - \bar{Y})^2}$
- The R^2 is the ratio of explained variation to total variation
  - The proportion of total variation in Y that our model has captured (with independent variables)
  - We can use R^2 to compare models

Examples of R-square

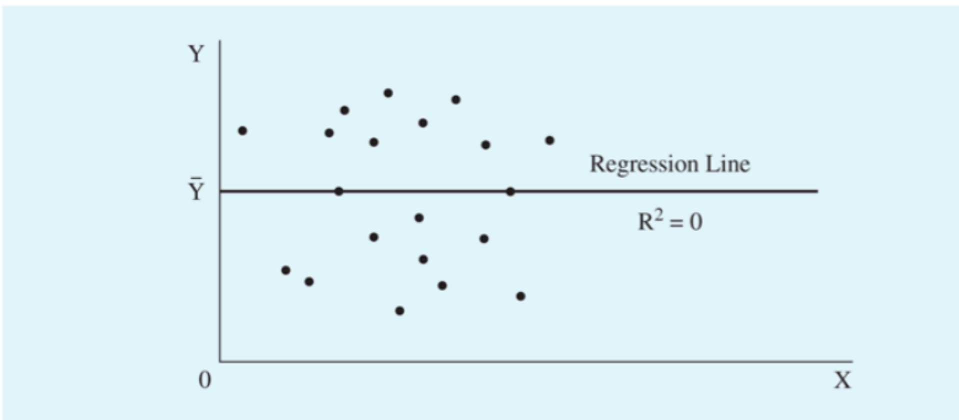


Figure 2.4
X and Y are not related; in such a case, R^2 would be 0.

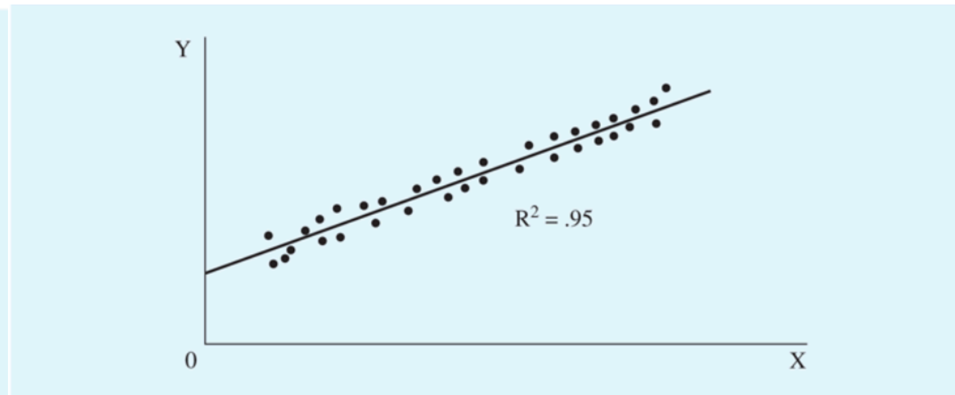


Figure 2.5
A set of data for X and Y that can be "explained" quite well with a regression line ($R^2 = .95$).

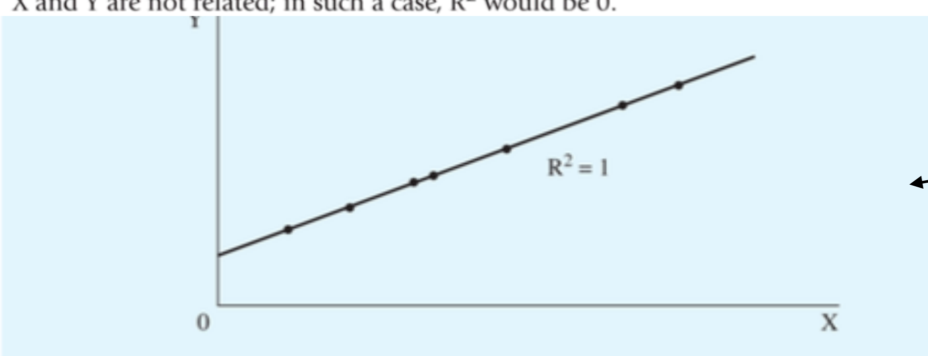


Figure 2.6
A perfect fit: all the data points are on the regression line, and the resulting R^2 is 1.

e.g. $WAGE_i = \beta_0 + \beta_1 WAGE_i + \varepsilon_i$

Exercise: compare models using R-square

Source	SS	df	MS
Model	611.927847	1	611.927847
Residual	60354.2942	46948	1.28555624
Total	60966.222	46949	1.29856274

Source	SS	df	MS
Model	4293.73993	1	4293.73993
Residual	56672.4821	46948	1.20713304
Total	60966.222	46949	1.29856274

- Model 1: age and health status
  - reg phstat age_yrs
- Model 2: education and health status
  - reg phstat educ_r1
- Manually calculate the R-squares from ESS and RSS
- Interpret what R-square means in each model in words
- Which model is better? Why?
- Does big R-squares necessarily means better models?
- Does big R-square necessarily means big β?
- Note the total SS are the same in both models, why?

Omitted variable bias

- The error ε arises because of factors that influence Y but are not included in the regression function; so, there are always omitted variables.

- Sometimes, the omission of those variables can lead to bias in the OLS estimator.

Omitted Variable Bias

- Suppose we are interested in studying whether getting health insurance makes people healthier?

Model in mind: $Health = \alpha + \beta \ * \ Insured + \ \varepsilon$

- Caveat: people with health insurance might be different from people without health insurance (ex: income). In this case, the selection bias is also called **Omitted Variable Bias**
- It is one of the most important issues in micro-econometrics. Most techniques we will learn later in this course focus on how to overcome this problem
- Recall (when we learned selection bias): naïve comparison is not “apple to apple” comparison-e.g. compare the health status of a person insured with a person not insured.

One potential way to fix Omitted Variable Bias (OVB): add controls in a Multivariate regression

- How can we make apple to apple comparison? In regression analysis, we add in controls.

- Let’s compare the health status of insured/uninsured people of the same/ similar ages with similar income.

- $Health = \alpha + \beta Insured + \gamma_1 Age + \gamma_2 Income + \varepsilon$

- Failing to include proper controls results in omitted variable bias.

How to interpret β?

- A big difference between multiple and single regression model is in the interpretation of the slope coefficients

- Now a slope coefficient indicates the change in the average of the dependent variable associated with a one-unit increase in the explanatory variable *holding the other explanatory variables constant or fixed*

How to interpret β?

- Example: $Health_i = \beta_0 + \beta_1 Educ_i + \beta_2 Age_i + \varepsilon_i$

- $\widehat{\beta}_1$=-0.15: Holding age constant, one unit increase in education level is associated with 0.15 unit decrease in average health index. (note here 1=excellent, so it actually increase health)

- Exercise: $\hat{\beta}_2$=0.07, explain what it means in the context.

Illustration about controls - DV

- We may be interested in discrimination
  - Are women discriminated against and paid less than men?
- We could estimate the following regression
  - $WAGE_i = \beta_0 + \beta_1 FEMALE_i + \varepsilon_i$

Our Regression

```
. reg wage female

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =   68.54
       Model |  828.220467     1  828.220467           Prob > F      =  0.0000
    Residual |  6332.19382   524  12.0843394           R-squared     =  0.1157
-------------+------------------------------           Adj R-squared =  0.1140
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.4763

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -2.51183   .3034092    -8.28   0.000    -3.107878   -1.915782
       _cons |   7.099489   .2100082    33.81   0.000     6.686928     7.51205
------------------------------------------------------------------------------
```

Control for Education

- When we control for other variables, we hold them constant in the regression
- For example, if we want to compare women and men who have the same education, we must control for education
- Our regression becomes
  - $WAGE_i = \beta_0 + \beta_1 FEMALE_i + \beta_2 EDUC_i + \varepsilon_i$

Comparing wages for women and men with same education

```
. reg wage female educ

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.626518           Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359           R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.273362   .2790444    -8.15   0.000    -2.821547   -1.725176
        educ |   .5064521   .0503906    10.05   0.000     .4074592     .605445
       _cons |   .6228168   .6725334     0.93   0.355    -.698382     1.944016
------------------------------------------------------------------------------
```

Same education and experience

```
. reg wage female educ exper

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   77.92
       Model |  2214.74206     3  738.247353           Prob > F      =  0.0000
    Residual |  4945.67223   522  9.47446788           R-squared     =  0.3093
-------------+------------------------------           Adj R-squared =  0.3053
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.0781

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.155517   .2703055    -7.97   0.000    -2.686537   -1.624497
        educ |   .6025802   .0511174    11.79   0.000     .5021591    .7030012
       exper |   .0642417   .0104003     6.18   0.000     .0438101    .0846734
       _cons |  -1.734481   .7536203    -2.30   0.022    -3.214982   -.2539797
------------------------------------------------------------------------------
```

Omitted variable bias

- The bias in the OLS estimator that occurs as a result of an omitted factor is called ***omitted variable*** bias. For omitted variable bias to occur, the omitted factor "*Z*" must be:

  1. A determinant of Y (i.e. Z is part of ε); **and**

  2. Correlated with the regressor X (*i.e.* $\text{corr}(Z,X) \neq 0$)

- ***Both*** *conditions must hold for the omission of Z to result in omitted variable bias*.

Omitted variable bias example

- Consider a model that regresses grade on attendance
  - Study time affects grades: *Z* is a determinant of *Y*.
  - Students who attend the class tend to study more: *Z* is correlated with *X*.

- Accordingly, our $\hat{\beta}_1$ is biased. What is the direction of this bias?
  - What does the common sense suggest?

Omitted variable bias direction

		Corr(omitted variable,x)	
		positive	negative
Corr(omitted variable,y)	positive	upward bias	downward bias
	negative	downward bias	upward bias

Review for quizzes/exam

- Understand simple linear regression model (difference between population and sample)
- Able to interpret $\widehat{\beta_0}$ and $\widehat{\beta_1}$ and know how to perform test of significance for β_1
- Able to interpret the STATA output of 1) coefficients and statistical inference (bottom panel), 2) variance decomposition (top left panel), and 3) R-square/adjusted R-square (top right panel)
- Understand Omitted variable bias