Lecture 5

Simple Linear Regression II

Outline

- Statistical inference in regression
- Variance decomposition
- R-square
- Omitted Variable Bias

Statistical inference in regression(review)

- Why do we need statistical inference in regression?
- We will use p-value(critical-value) approach and confidence interval 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?
  - H_0: β_1=0
  - H_a:β_1≠0
- If we later reject H_0, it means β_1 is not zero. So we identify a relationship between X and Y (with certain confidence)
- If we fail to reject H_0, it means β_1 can be zero. So we fail to identify a relationship between X and Y (with certain confidence)

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: $\beta_1 \neq 0$
  - $t^*=\frac{\widehat{\beta}_1-0}{\widehat{se(\beta_1)}}=\frac{\widehat{\beta}_1}{\widehat{se(\beta_1)}}=\frac{0.9598}{0.2680}=3.5813$ (recall previously we used the formula for $t^*=\frac{\bar{X}-\mu}{se}$)
  - 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
  - $t^*=\frac{\widehat{\beta}_1}{\widehat{se}}$ ($\widehat{\beta_1}$, $\widehat{se}$ 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.

Exercise (p-value approach)

- 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

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 (confidence interval approach)

- 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

- Write out the estimated regression.
- 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?
- Additional questions to think about: What does the standard error of $\hat{\beta}_1$ mean?
- 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?

```
. reg phstat age_yrs
```

Overall fit and model comparison

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

Number of obs	=	46950
F(1, 46948)	=	476.00
Prob > F	=	0.0000
R-squared	=	0.0100
Adj R-squared	=	0.0100
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

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

- 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

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

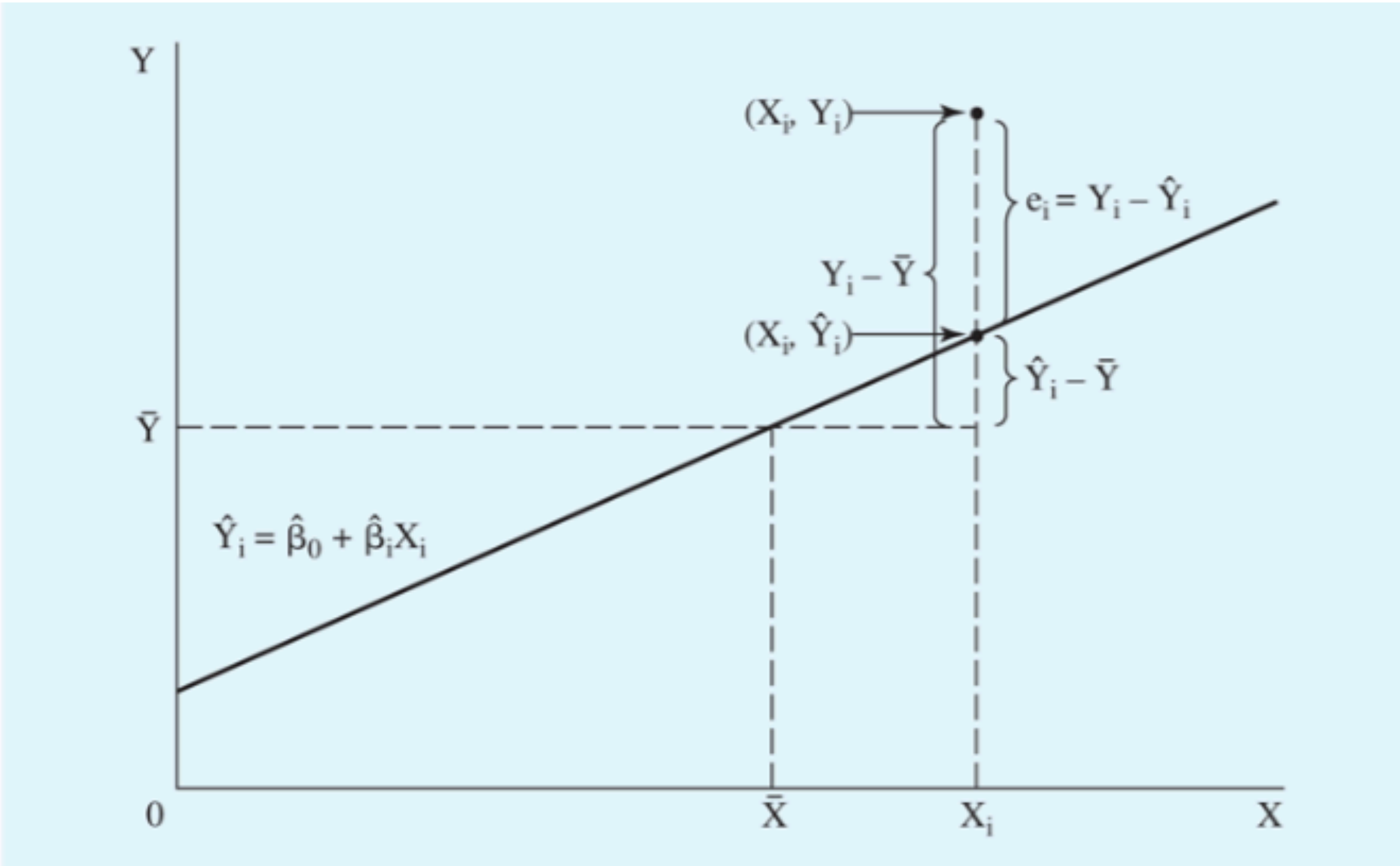


Figure 2.3 Decomposition of the Variance in Y

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

Exercise

- Run the regression of health status on education and interpret the top left panel. What is the ESS and d.f.? What is the RSS? Is the ESS bigger or smaller compared with the previous model (regress health on age)? If you want to compare models, is it a good idea to use ESS?

```
. reg phstat educ_r1

      Source |       SS           df       MS      Number of obs   =     46950
-------------+----------------------------------   F(  1, 46948)   =   3556.97
       Model |  4293.73993         1  4293.73993   Prob > F        =    0.0000
    Residual |  56672.4821    46948  1.20713304   R-squared       =    0.0704
-------------+----------------------------------   Adj R-squared   =    0.0704
       Total |   60966.222    46949  1.29856274   Root MSE        =    1.0987

------------------------------------------------------------------------------
      phstat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     educ_r1 |  -.1411819   .0023672   -59.64   0.000    -.1458217   -.1365421
       _cons |   3.212889    .010187   315.39   0.000     3.192923    3.232856
------------------------------------------------------------------------------
```

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

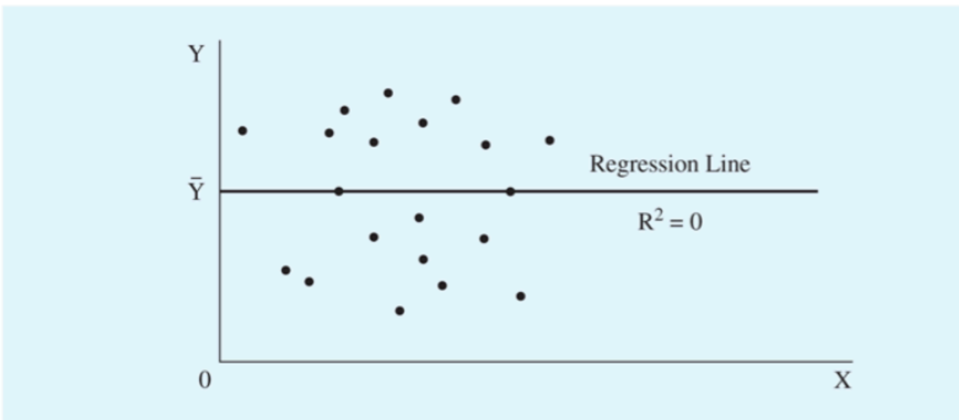


Figure 2.4
X and Y are not related; in such a case, R^2 would be 0.

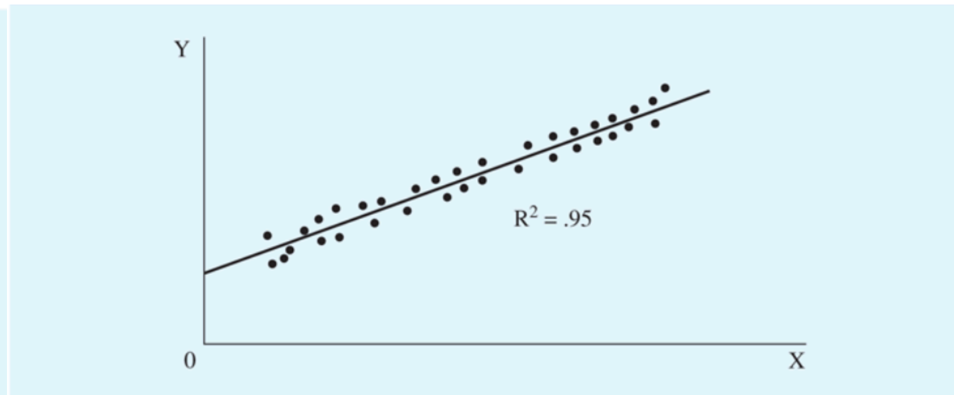


Figure 2.5
A set of data for X and Y that can be "explained" quite well with a regression line ($R^2 = .95$).

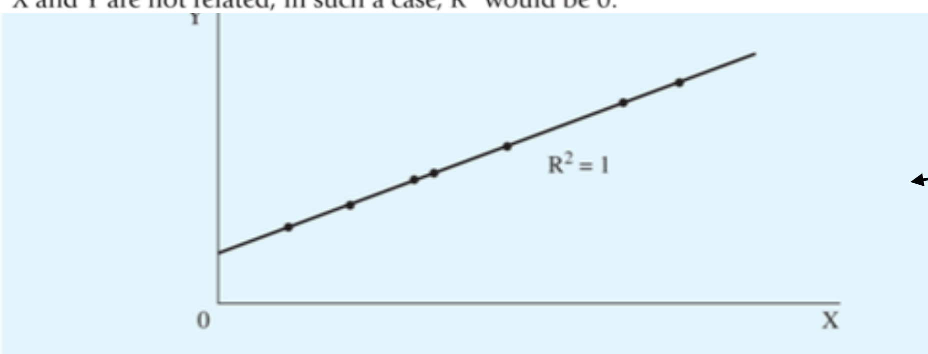


Figure 2.6
A perfect fit: all the data points are on the regression line, and the resulting R^2 is 1.

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

Exercise: compare models using R-square

- Model 1: age and health status
  - reg phstat age_yrs

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

- Model 2: education and health status
  - reg phstat educ_r1

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

- 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 in multiple regression?

- 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
------------------------------------------------------------------------------
```

Another example: Effect of weather on Shopping

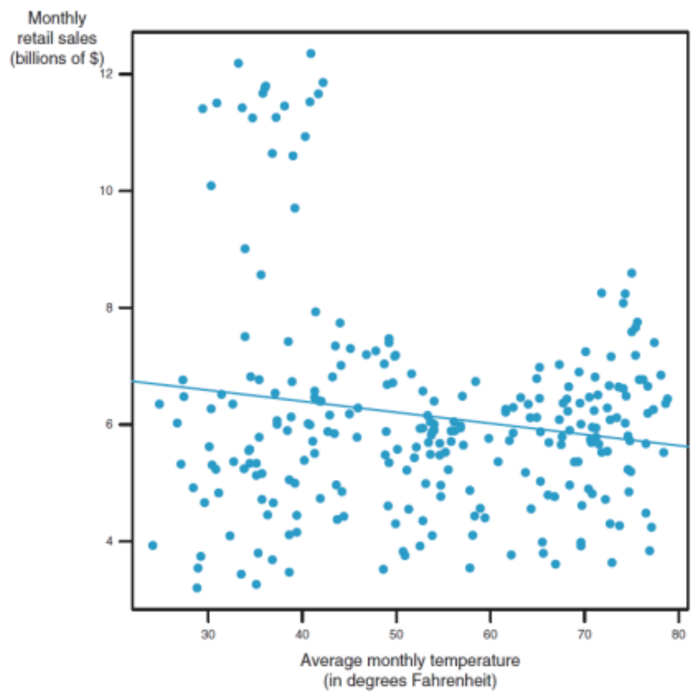


FIGURE 5.1: Monthly Retail Sales and Temperature in New Jersey from 1992 to 2013

```
reg RetailBillions temperature
Number of obs     =       256
R-squared         =    0.0256
Adj R-squared     =    0.0218
Root MSE          =    1.8155

------------------------------------------------------------------------------
RetailBill~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 temperature |  -.0189569   .0073326    -2.59   0.010    -.0333974   -.0045164
       _cons |   7.156903   .4081326    17.54   0.000     6.353148    7.960658
------------------------------------------------------------------------------
```

Re-plot data – netting out December effect

```
** Show that sales are higher in December
reg RetailBillions dec
Number of obs    =        256
R-squared        =     0.6484
Root MSE         =     1.0906
RetailBill~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dec |   5.376067   .2483885    21.64   0.000     4.886904     5.86523
       _cons |   5.702362   .0711412    80.16   0.000      5.56226    5.842463
```

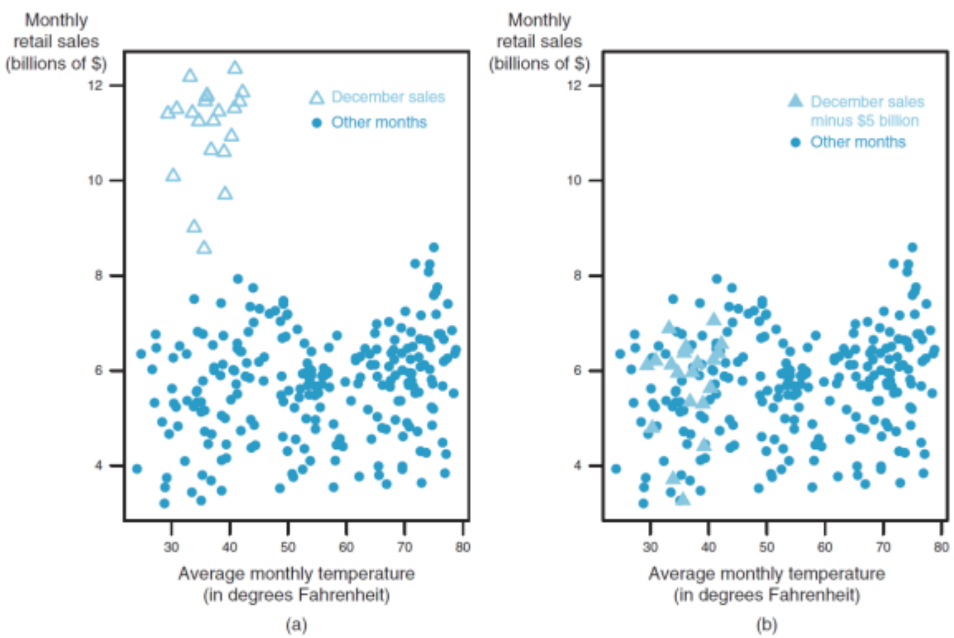
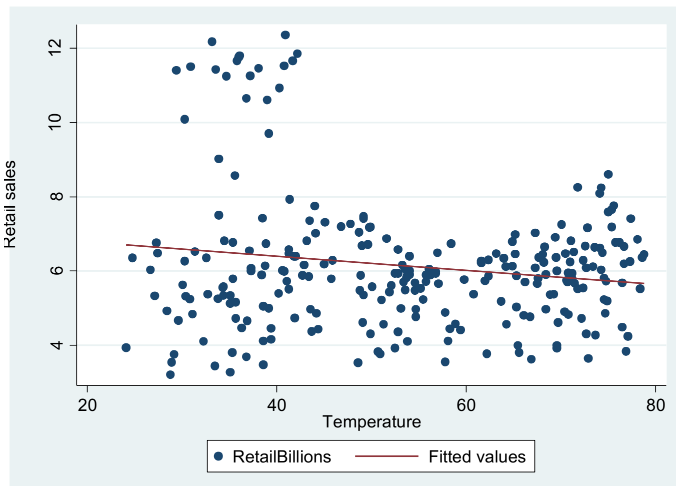
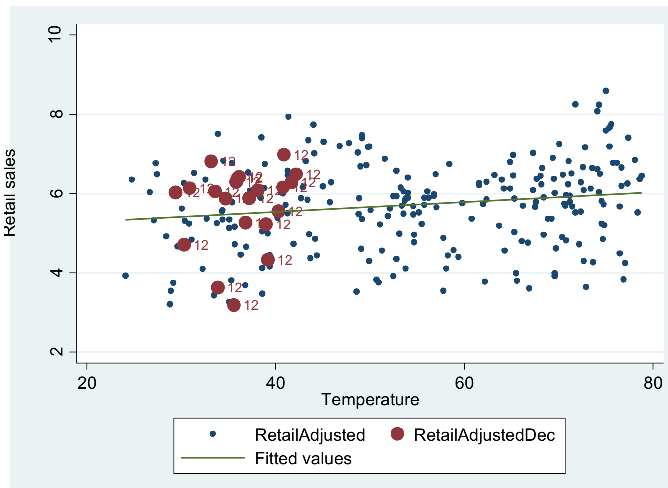


FIGURE 5.2: Monthly Retail Sales and Temperature in New Jersey with December Indicated

Heuristic description of multivariate OLS



Original bivariate model



Model in which “December effect” has been controlled for

gen RetailAdjusted = retail_ns – 5.376*dec

* reduces Dec sales by $5.376

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

- 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 direction

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

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    46948  1.28555624   R-squared       =    0.0100
-------------+----------------------------------   Adj R-squared   =    0.0100
       Total |   60966.222    46949  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

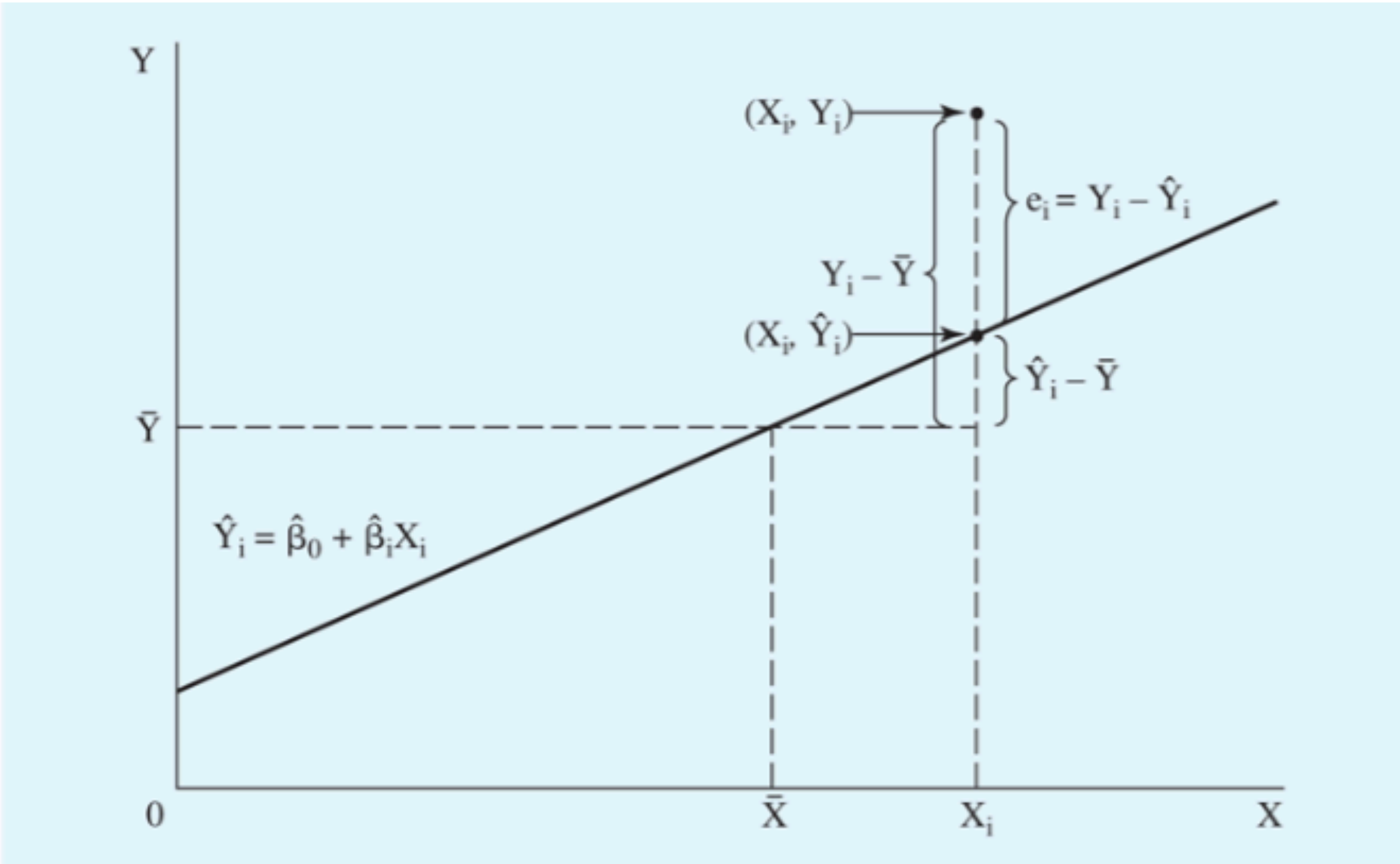


Figure 2.3 Decomposition of the Variance in Y

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

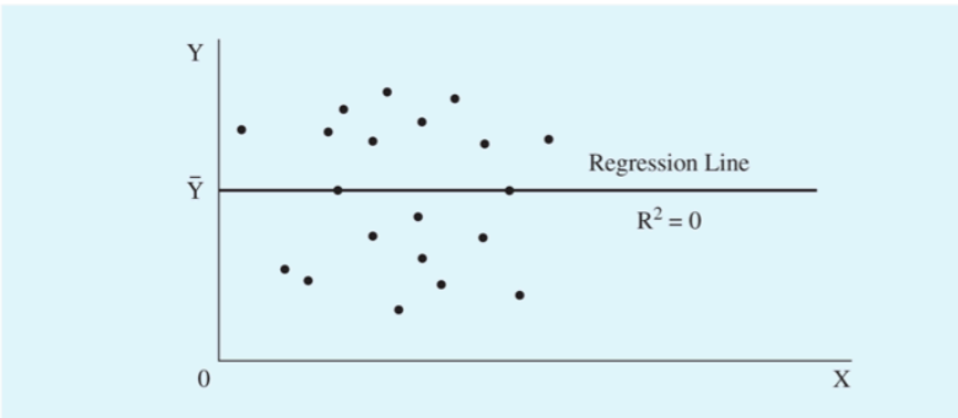


Figure 2.4
X and Y are not related; in such a case, R^2 would be 0.

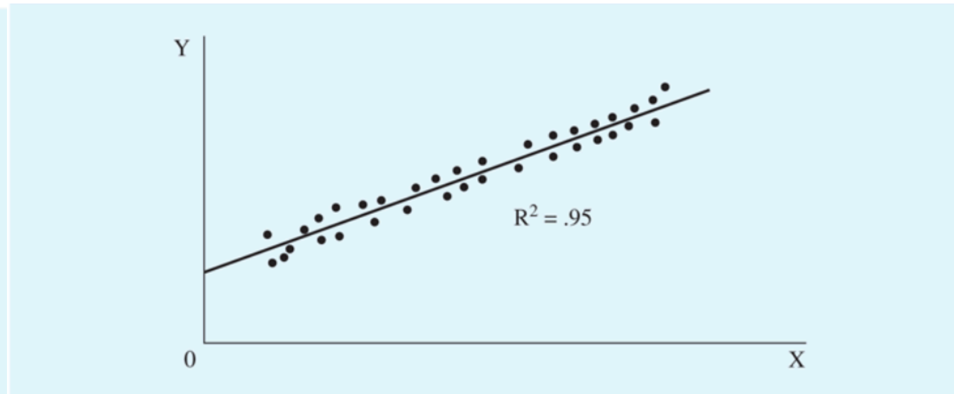


Figure 2.5
A set of data for X and Y that can be "explained" quite well with a regression line (R^2 = .95).

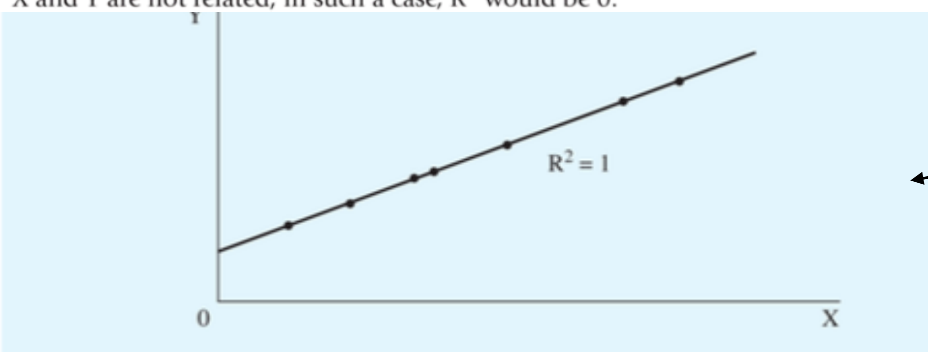


Figure 2.6
A perfect fit: all the data points are on the regression line, and the resulting R^2 is 1.

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?

Review for quizzes

- How to do hypothesis test in regression using P-value and CI approach
- 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