Multiple Regression

- Multiple Regression Model
  - Identifying Regression Coefficients
  - Fit of the Model
    - Testing the Significance of the Overall Regression Model
    - Adjusted Coefficient of Determination
  - Inference About the Independent Variables

- Reading: Chapter 15
  - Skip Sections 15.4 – 15.5
  - Section 15.3 – skip confidence intervals for the coefficients

Multiple Regression Model

This chapter extends the simple regression model discussed in Chapter 14

Now consider **more than one independent variable** ($x_1, x_2, \ldots, x_k$) as a means to explain the variation in the dependent variable of interest y

Simple Regression Model (Chapter 14):

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Multiple Regression Model (Chapter 15):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Multiple Regression Model

To be more precise…

Suppose there are k independent variables $x_1, x_2, \ldots, x_k$ that affect the dependent variable y

The multiple linear regression model is defined as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

where:

y = The dependent variable (observed value of y)

$x_1, x_2, \ldots, x_k$ = The independent variables of interest

ε = The random error term (other factors that affect y)

$\beta_0, \beta_1, \ldots, \beta_k$ = The unknown parameters to be estimated

Multiple Regression Model

Estimated Multiple Regression Equation:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

where:

$\hat{y}$ = The predicted value of y given values of $x_1, x_2, \ldots, x_k$

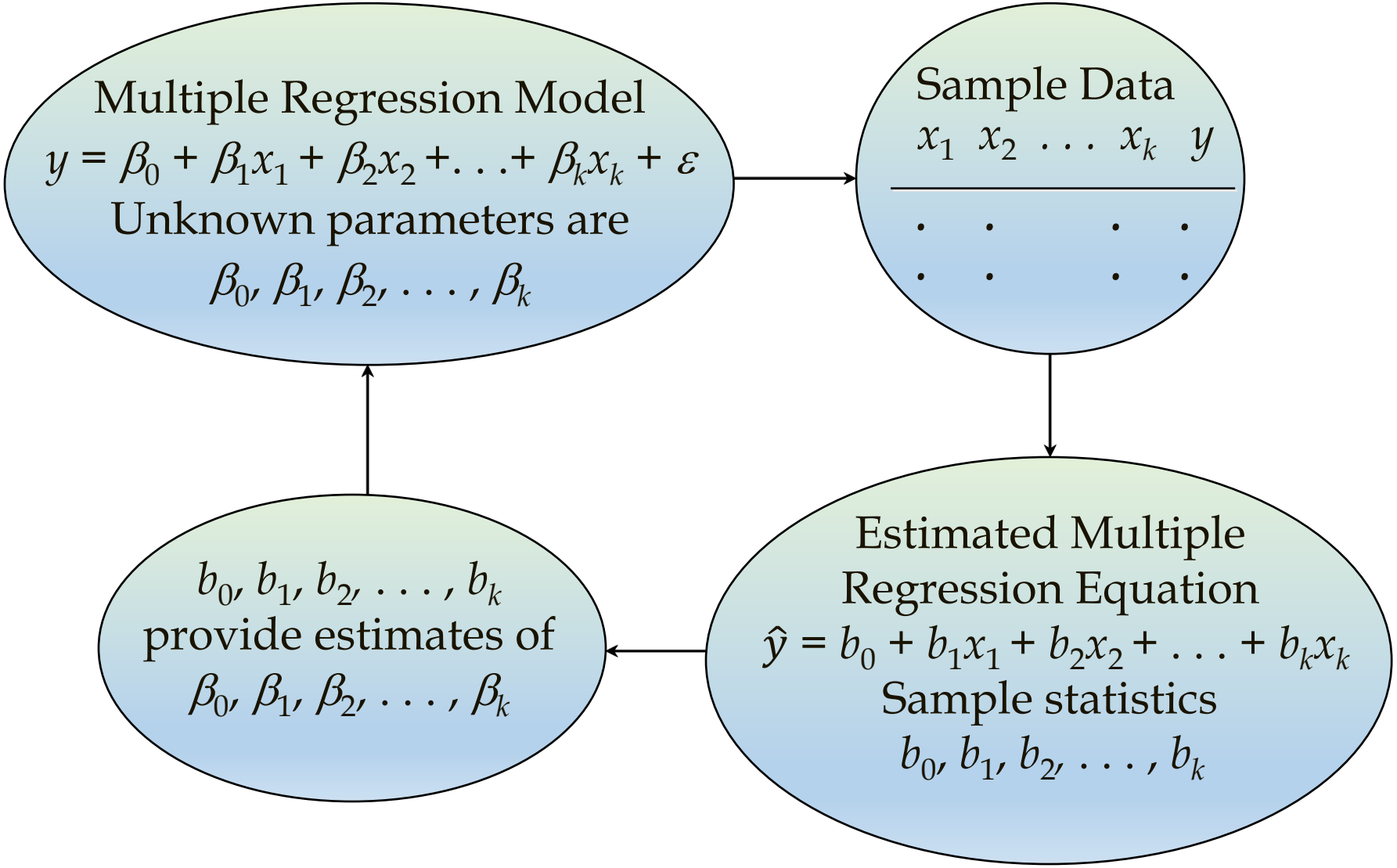
$x_1, x_2, \ldots, x_k$ = The independent variables of interest

k = The number of independent variables

b_0 = The y-intercept of the regression line

$b_1, b_2, \ldots, b_k$ = **Estimated** slope coefficients

Estimation Process



Estimating Regression Coefficients

Least Squares Method: identify $b_0, b_1, b_2, \ldots, b_k$ such that

$$\min \sum (y_i - \hat{y}_i)^2$$

Computation of Coefficient Values

- Even though the same techniques are used to identify $b_0, b_1, b_2, \ldots, b_k$, the formulas for the regression coefficients are quite complicated and may involve the use of matrix algebra
- Manual calculations are quite involved
- We will rely on Excel to perform calculations and will focus on the interpretation of the results

Identifying Regression Coefficients

Example: A dairy cooperative wants to examine household milk consumption (y, in quarts per week) based on annual household income (x_1, measured in \$1,000 per year) and the size of the family (x_2)

Data for 15 randomly selected families is obtained ($n = 15$)

Note: it is hard to represent this data using visual tools like scatter plot ⇒ after defining dependent and independent variables, we will jump into the estimation right away

Family	Milk Consumption (Quarts)	Income ($1,000)	Family Size
1	21	46	5
2	10	55	2
3	16	37	3
4	38	60	5
5	9	35	1
6	26	55	3
7	22	41	3
8	28	50	4
9	25	52	3
10	18	49	2
11	12	34	3
12	20	39	3
13	28	44	4
14	30	56	4
15	35	49	6

Identifying Regression Coefficients

Same estimation technique in Excel using *Data Analysis*

but

In the **Input X Range**, we select array of data with all values of ALL independent variables at once

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	***Regression Statistics***						
4	Multiple R	0.8816					
5	R Square	0.7771					
6	Adjusted R Square	0.7400					
7	Standard Error	4.4196					
8	Observations	15					
9							
10	**ANOVA**						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	817.3375	408.6687	20.9220	0.0001	
13	Residual	12	234.3958	19.5330			
14	Total	14	1051.7333				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-11.1378	6.8686	-1.6216	0.1309	-26.1032	3.8276
18	Income	0.3913	0.1545	2.5324	0.0263	0.0546	0.7281
19	Family Size	4.5165	0.9765	4.6250	0.0006	2.3888	6.6442

The estimated regression equation is:

$$\hat{y} = -11.1378 + 0.3913\, x_1 + 4.5165\, x_2$$

Interpretation of Regression Coefficients

In multiple regression, there is a slight modification of the interpretation of the slopes

Regression coefficient b_j shows the average *change* in the dependent variable y due to a one-unit *increase* in the associated independent variable x_j **while other variables are held constant**

For example,

- b_1 = The average change in y due to a one-unit *increase* in x_1 holding $x_2, \ldots, x_k$ constant
- b_2 = The average change in y due to a one-unit *increase* in x_2 holding $x_1, x_3, \ldots, x_k$ constant

Note: do not forget about units in the interpretation!

Interpreting Regression Coefficients, I

The estimated regression equation is:

$$\hat{y} = -11.1378 + 0.3913\, x_1 + 4.5165\, x_2$$

where:

x_1 = Annual family income (in \$1,000)

x_2 = Family size (# of people)

Interpreting the regression coefficients:

- The income coefficient ($b_1 = 0.3913$) tells us that an additional \$1,000 of annual income, *holding family size constant*, will **increase** a family's milk consumption by 0.3913 quarts per week, on average

We are holding other variables constant

Because the coefficient is positive

Interpreting Regression Coefficients, II

The estimated regression equation is:

$$\hat{y} = -11.1378 + 0.3913\, x_1 + 4.5165\, x_2$$

where:

x_1 = Annual family income (in \$1,000)

x_2 = Family size (# of people)

Interpreting the regression coefficients (notice units!):

- The family size coefficient ($b_2 = 4.5165$) tells us that for each additional family member, *holding family income constant*, a family's milk consumption will **increase** by 4.5165 quarts per week, on average

Interpreting Regression Coefficients

The estimated regression equation is:

$$\hat{y} = -11.1378 + 0.3913\, x_1 + 4.5165\, x_2$$

- It is tempting to interpret the *y*-intercept (-11.1378) as the milk consumption for a family with *zero* income and *zero* family members, but…
  - A negative value of milk consumption does not make sense
  - Zero family members? Does not make sense
- We do not have enough observations in our sample with values of x_1 and x_2 close to zero $\Rightarrow$ our regression may produce unreliable predictions of milk consumption if values of x_1 and x_2 close to zero

Making Predictions

- Plug in the desired values of $x_1, x_2, \ldots, x_k$ into the estimated regression equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

to obtain the predicted value of the dependent variable, $\hat{y}$

- In our example, the predicted milk consumption for a family with annual income of \$50,000 and a size of 4 people ($x_1 = 50$, $x_2 = 4$):

$$\hat{y} = -11.1378 + 0.3913(50) + 4.5165(4)$$

$$= -11.1378 + 19.565 + 18.066$$

$$= 26.493 \text{ (quarts per week)}$$

Goodness-of-Fit Measures

Three measures to judge how well the estimated regression fits the data:

- **The Standard Error of the Estimate, s_e**
  - Similar to the simple linear regression (same meaning and same formula)
  - We will not look at it in the context of multiple regression
- **The Coefficient of Determination, R^2**
  - Again, similar to the simple linear regression but we will talk about it
- **The Adjusted R^2**

Multiple Coefficient of Determination

Goal: Determine the amount of the variation in y that is due to variation in ALL independent variables

Note: same idea as in the simple linear regression

$$SST = SSR + SSE$$

Total sum of Squares

$$SST = \sum (y_i - \bar{y})^2$$

Sum of Squares Regression

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

Explained by the regression model

Sum of Squares Error

$$SSE = \sum (y_i - \hat{y}_i)^2$$

Not explained by the regression model

The Multiple Coefficient of Determination

The **multiple coefficient of determination, R^2**, is the percentage of variation in the dependent variable that is explained by ***ALL of the independent variables***

$$R^2 = \frac{SSR}{SST}$$

In our example,

$$R^2 = \frac{SSR}{SST} = \frac{817.3375}{1051.7333} = 0.7771$$

- 77.71% of the variation in milk consumption (y) is explained by **family income and family size** (all x variables)

The Multiple Coefficient of Determination

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	***Regression Statistics***						
4	Multiple R	0.8816					
5	R Square	0.7771					
6	Adjusted R Square	0.7400					
7	Standard Error	4.4196					
8	Observations	15					
9							
10	**ANOVA**						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	817.3375	408.6687	20.9220	0.0001	
13	Residual	12	234.3958	19.5330			
14	Total	14	1051.7333				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-11.1378	6.8686	-1.6216	0.1309	-26.1032	3.8276
18	Income ($1000s)	0.3913	0.1545	2.5324	0.0263	0.0546	0.7281
19	Family Size	4.5165	0.9765	4.6250	0.0006	2.3888	6.6442

$$R^2 = \frac{SSR}{SST}$$

SSR

SSE

SST

Testing the Significance of the Overall Regression Model

Goal: decide if the relationship between the dependent and **ALL independent variables** is statistically significant

- It's pretty much the test to determine the significance of the coefficient of determination as we know it from the simple linear regression
- But it is often stated differently in the context of multiple regression

General idea:

- If the population coefficient of determination = 0, then ALL variables x explain 0% of variation in y
- I.e., none of the x variables affect y which implies that ALL regression slopes are equal to 0

$$H_0: \text{All population slopes } \beta_j = 0$$

$$H_1: \text{At least one } \beta_j \neq 0$$

Testing the Significance of the Overall Regression Model

In our example, to test the overall significance of the regression model:

- Values of b_1 and b_2 are obtained based on the sample and serve as estimates of the population coefficients β_1 and β_2
- If the population slopes β_1 and β_2 are both zero, then x_1 and x_2 have no effect on y, and we would conclude that there is no relationship between y and ALL independent variables
- So, we examine the following hypotheses:

$H_0 : \beta_1 = \beta_2 = 0$ (no relationship exists between the y and ALL independent variables)

H_1 : At least one $\beta_j \neq 0$ (A relationship exists and at least one variable x affects y)

Testing the Significance of the Overall Regression Model

F-test Statistic for the Overall Regression Model

$$F = \frac{MSR}{MSE}$$

With degrees of freedom
$D_1 = k$
$D_2 = n - k - 1$

where:

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n - k - 1}$$

MSR = The mean square regression
MSE = The mean square error
SSR = The sum of squares regression
SSE = The sum of squares error
n = The number of observations in the sample
k = The number of independent variables

Pay attention to k because k ≠ 1 in multiple regression

Testing the Significance of the Overall Regression Model

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	***Regression Statistics***						
4	Multiple R	0.8816					
5	R Square	0.7771					
6	Adjusted R Square	0.7400					
7	Standard Error	4.4196					
8	Observations	15					
9							
10	**ANOVA**						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	817.3375	408.6687	20.9220	0.0001	
13	Residual	12	234.3958	19.5330			
14	Total	14	1051.7333				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-11.1378	6.8686	-1.6216	0.1309	-26.1032	3.8276
18	Income ($1000s)	0.3913	0.1545	2.5324	0.0263	0.0546	0.7281
19	Family Size	4.5165	0.9765	4.6250	0.0006	2.3888	6.6442

$$F = \frac{MSR}{MSE}$$

MSR

MSE

Testing the Significance of the Overall Regression Model

In our example, the critical value and the *p*-value are computed from the *F* distribution with *df*

- $D_1 = k = 2$
- $D_2 = n - k - 1 = 15 - 2 - 1 = 12$

Conclusion:

1. The *p*-value = 0.0001 is provided in the Excel output (*Significance F* column on the previous slide)
2. Suppose $\alpha = 0.05$
   - $\Rightarrow$ Since *p-value* = 0.0001 < α, we **reject H_0** that all regression coefficients are equal to zero
   - $\Rightarrow$ We have enough evidence for H_1 and conclude that *at least one* of the population coefficients is not equal 0

The Adjusted Multiple Coefficient of Determination: Intuition

Usually, R^2 is NOT used for model comparison when the competing models do not include the same number of x variables

- R^2 never decreases as we add more x variables to the model
  - For example, we start with Model 1 which includes only x_1
  - Next, we run Model 2 which includes x_1 and x_2
  - It may be that additional variable x_2 is useless to explain variation in y… but R^2 in Model 2 is still going to be higher
- This makes is difficult to use R^2 to compare the goodness-of-fit of different models

Even though the $R^2 = 0.8$ for Model 2 is greater than that of Model 1 ($R^2 = 0.6$), we do NOT conclude that Model 2 provides a better fit

	Model 1	Model 2
R^2	0.6	0.8

The Adjusted Multiple Coefficient of Determination, R_A^2

- More x variables ⇒ higher R^2
- But some of these variables may be unimportant and should not be included in the model
- The **adjusted coefficient of determination**, R_A^2, modifies, or adjusts, the coefficient of determination R^2 by accounting for the number of independent variables (k) used to develop the regression
  - R_A^2 penalizes for adding more x variables in the model

$$R_A^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

The Adjusted Multiple Coefficient of Determination, R_A^2

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	***Regression Statistics***						
4	Multiple R	0.8816					
5	R Square	0.7771					
6	Adjusted R Square	0.7400					
7	Standard Error	4.4196					
8	Observations	15					
9							
10	**ANOVA**						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	817.3375	408.6687	20.9220	0.0001	
13	Residual	12	234.3958	19.5330			
14	Total	14	1051.7333				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-11.1378	6.8686	-1.6216	0.1309	-26.1032	3.8276
18	Income ($1000s)	0.3913	0.1545	2.5324	0.0263	0.0546	0.7281
19	Family Size	4.5165	0.9765	4.6250	0.0006	2.3888	6.6442

The Adjusted Multiple Coefficient of Determination, R_A^2

Notes about R_A^2:

1. R_A^2 is NOT interpreted!!!
2. R_A^2 is used to compare regressions with different number of x variables
   - The higher the R_A^2, the better the model
3. R_A^2 is used to decide whether the variable needs to be included in the model
   - Adding new x variable to the model will always increase R^2
   - But… If an additional x variable causes a reduction in R_A^2, it's an evidence that the new x variable might not be worth keeping in the model

Example: R_A^2

- We are interested in studying factors responsible for debt disparities across U.S. cities
- Main suspect for the disparity – income level
- We collect the data on average debt payment and average income in 35 U.S. cities and run

  $\Longrightarrow$ ***Model 1:*** Debt Payments (y) on Income (x_1)
- We are criticized that Model 1 ignores effect of unemployment

  $\Longrightarrow$ ***Model 2:*** Debt Payments (y) on Income (x_1) and Unemployment (x_2)

	Model 1	Model 2
R^2	0.7526	0.7529
R_A^2	0.7451	0.7375

- Even though R^2 is lower for Model 1, R_A^2 is higher for this model
- Therefore, we'd choose Model 1 to predict debt payments

Inference about Independent Variables

The ***individual*** regression coefficients can be examined for significance

For example,

$H_0 : \beta_1 = 0$ (No relationship exists between milk consumption and family income (x_1))

$H_1 : \beta_1 \neq 0$ (A relationship *does exist* between milk consumption and family income (x_1))

$H_0 : \beta_2 = 0$ (No relationship exists between milk consumption and family size (x_2))

$H_1 : \beta_2 \neq 0$ (A relationship *does exist* between milk consumption and family size (x_2))

A Significance Test for the Regression Coefficients

t test Statistic for the Regression Slope on x_j

$$t = \frac{b_j - \beta_j}{s_{b_j}}$$

where:

b_j = The estimated regression slope on x_j
β_j = The population regression slope for x_j (from H_0)
s_{b_j} = The standard error of the slope for x_j

To find the critical value and p-value:

- Use a t distribution with $df = n - k - 1$

A Significance Test for the Regression Coefficients

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	***Regression Statistics***						
4	Multiple R	0.8816					
5	R Square	0.7771					
6	Adjusted R Square	0.7400					
7	Standard Error	4.4196					
8	Observations	15					
9							
10	**ANOVA**						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	817.3375	408.6687	20.9220	0.0001	
13	Residual	12	234.3958	19.5330			
14	Total	14	1051.7333				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-11.1378	6.8686	-1.6216	0.1309	-26.1032	3.8276
18	Income ($1000s)	0.3913	0.1545	2.5324	0.0263	0.0546	0.7281
19	Family Size	4.5165	0.9765	4.6250	0.0006	2.3888	6.6442

A Significance Test for the Regression Coefficients

15					
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17	Intercept	-11.1378	6.8686	-1.6216	0.1309
18	Income ($1000s)	0.3913	0.1545	2.5324	0.0263
19	Family Size	4.5165	0.9765	4.6250	0.0006

To state conclusions, use the p-values from Excel output

1. In our example, p-values are both less than $\alpha = 0.05$
2. For each variable, we reject the null hypothesis that the coefficient is equal to zero
   - $\Rightarrow$ the relationship between family income and milk consumption is *statistically significant* (income is statistically significant in explaining milk consumption)
   - $\Rightarrow$ the relationship between family size and milk consumption is *statistically significant*

A Significance Test for the Regression Coefficients

		Coefficients	*Standard Error*	*t Stat*	*P-value*
15					
16					
17	Intercept	-11.1378	6.8686	-1.6216	0.1309
18	Income ($1000s)	0.3913	0.1545	2.5324	0.0263
19	Family Size	4.5165	0.9765	4.6250	0.0006

Let's see if conclusion are the same when $\boldsymbol{\alpha = 0.01}$

1. Use the *p*-values from the regression output
2. For *Family Size*, we reject H_0 (*p*-value < α)

   $\Rightarrow$ (same conclusion) the relationship between family size and milk consumption is *statistically significant*
3. For *Income*, we CANNOT reject H_0 (*p*-value > α)

   $\Rightarrow$ (different conclusion) we CANNOT conclude that the relationship between family income and milk consumption is *statistically significant*

Exercise 15.29 (Modified)

City Hospital would like to develop a regression model to predict the total hospital bill for a patient based on his or her length of stay, number of days in the hospital's intensive care unit (ICU), and the age of the patient. Data for these variables can be found in the Excel file **City Hospital.xlsx** (*Excel Files → Ch 15*).

1. Define independent and dependent variables.
2. Estimate the regression.
3. Write down the estimated regression equation and interpret the values of the regression coefficients.

Exercise 15.29 (Continued)

4. What is the value of the coefficient of determination? How would you interpret it?
5. Using $\alpha = 0.01$, test the overall significance of the model.
6. Using $\alpha = 0.01$, test the significance of the regression coefficients?
7. Using your estimated regression equation, predict a hospital bill for 36-years old individual who stayed in the hospital for 4 days with 2 days in the hospital's intensive care unit (ICU).