

Chapter 17

Multiple Regression

Regression Analysis

- **In fact, individuals in the population are not homogeneous.**
- **Many heterogeneities among the population which can be described by interval variables.**
 - **E.g., GPA, height, weight, age, etc.**
- **Can we extract more information from these heterogeneities?**

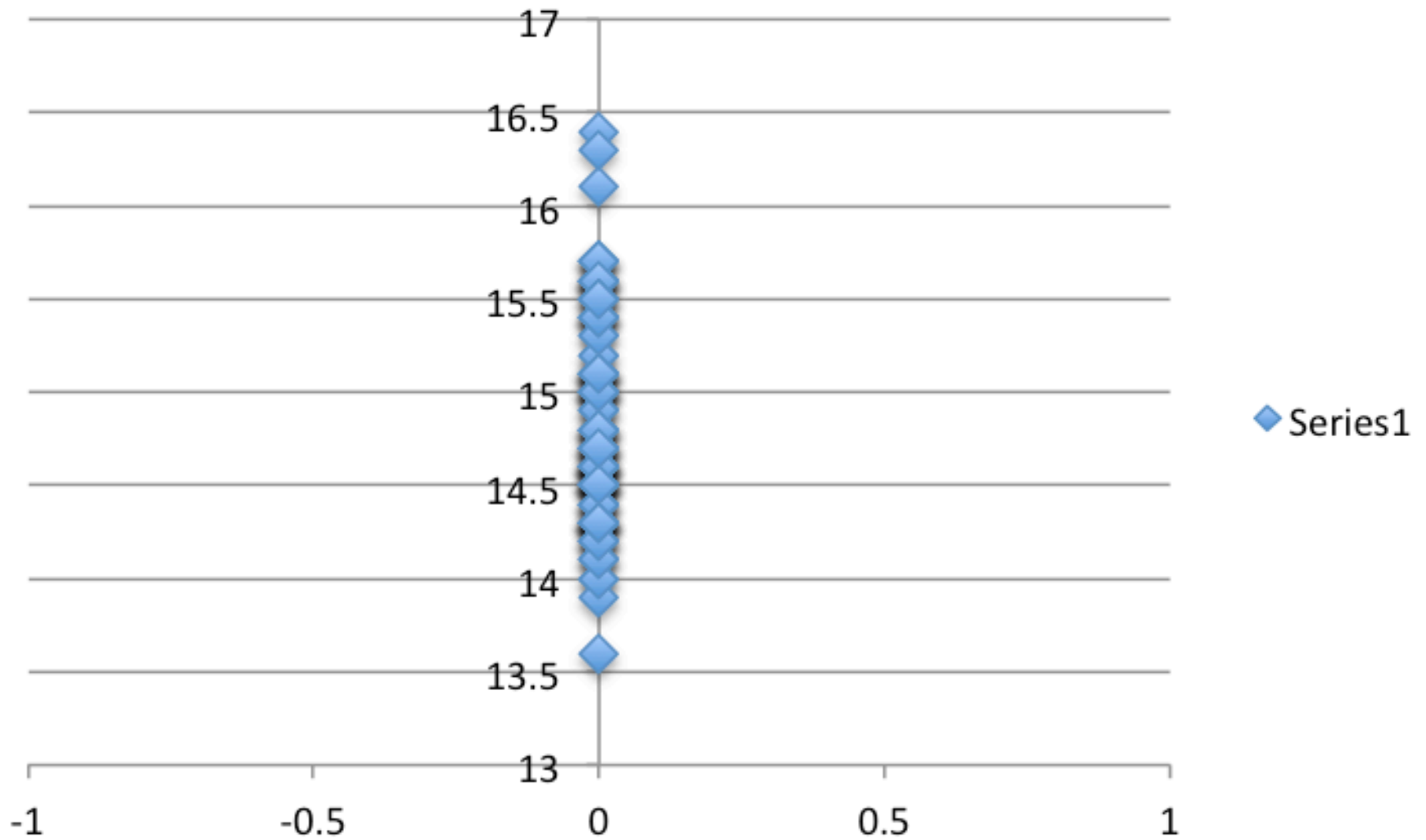
Multiple Regression...

The *simple linear regression model* was used to analyze how one interval variable (the dependent variable **y**) is related to one other interval variable (the independent variable **x**).

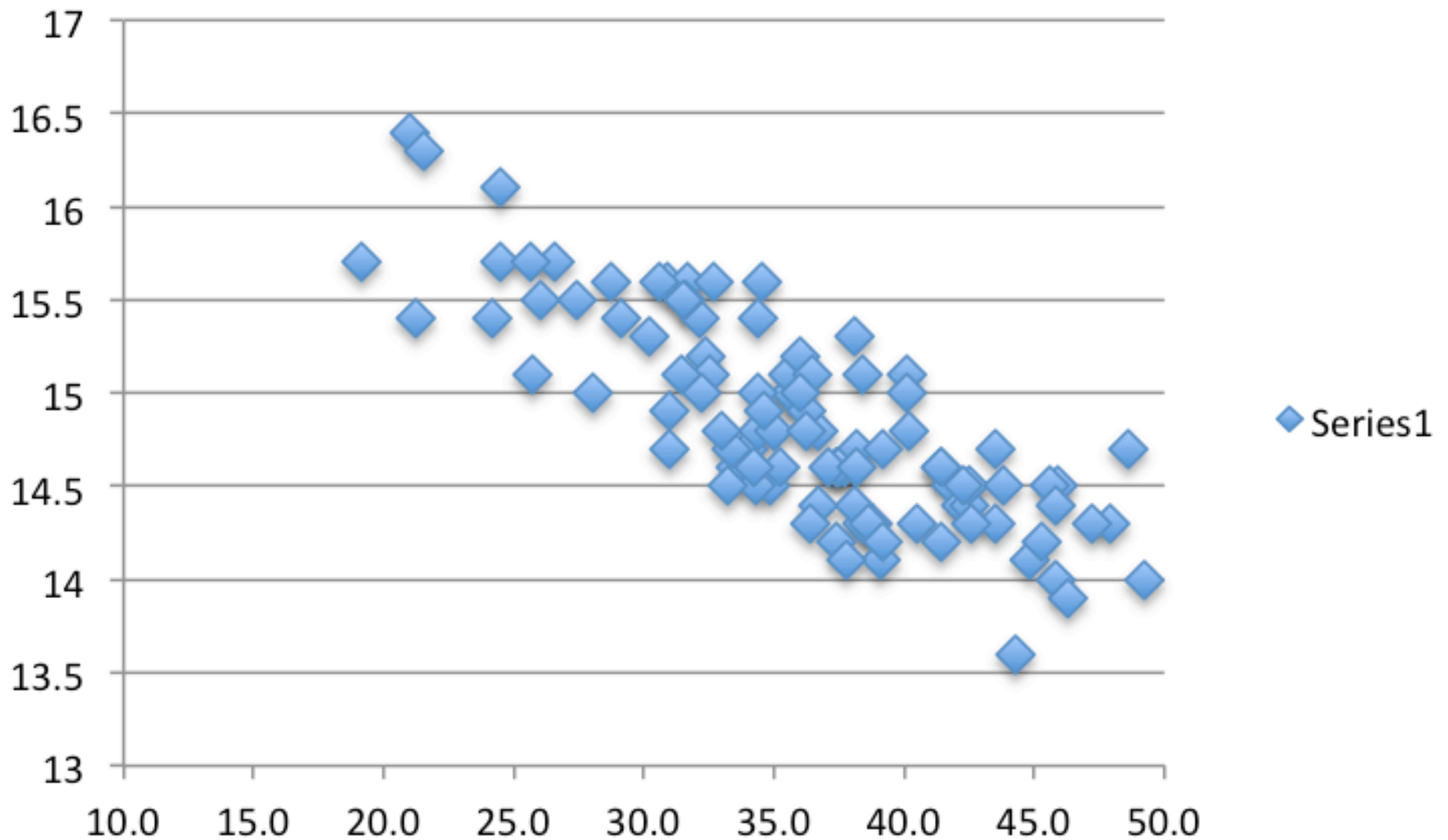
Multiple regression allows for any number of independent variables.

We expect to develop models that *fit the data better* than would a simple linear regression model.

What can we do?



Can we do this?



Multiple Regression for $k = 2$,

$$b_0 n + b_1 \sum X_1 + b_2 \sum X_2 = \sum Y$$

$$b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 = \sum X_1 Y$$

$$b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 = \sum X_2 Y$$

y

The simple linear regression model allows for one independent variable, "x"

$$y = b_0 + b_1 x + \varepsilon$$

$$y = b_0 + b_1 x$$

X_1

Multiple Regression for $k = 2$,

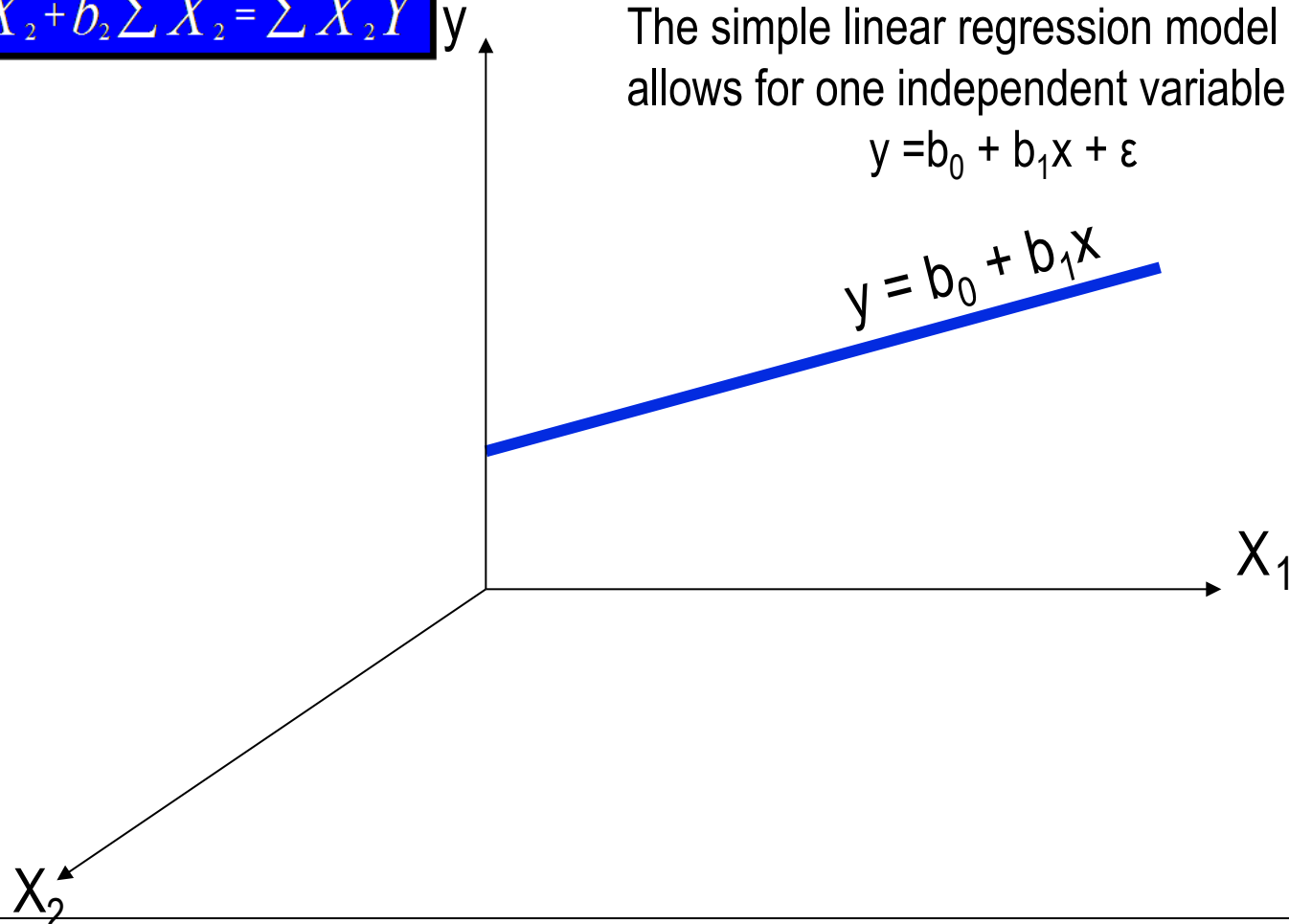
$$b_0 n + b_1 \sum X_1 + b_2 \sum X_2 = \sum Y$$

$$b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 = \sum X_1 Y$$

$$b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 = \sum X_2 Y$$

The simple linear regression model allows for one independent variable, "x"

$$y = b_0 + b_1 x + \varepsilon$$

$$y = b_0 + b_1 x$$


X_2

Multiple Regression for $k = 2$,

$$b_0 n + b_1 \sum X_1 + b_2 \sum X_2 = \sum Y$$

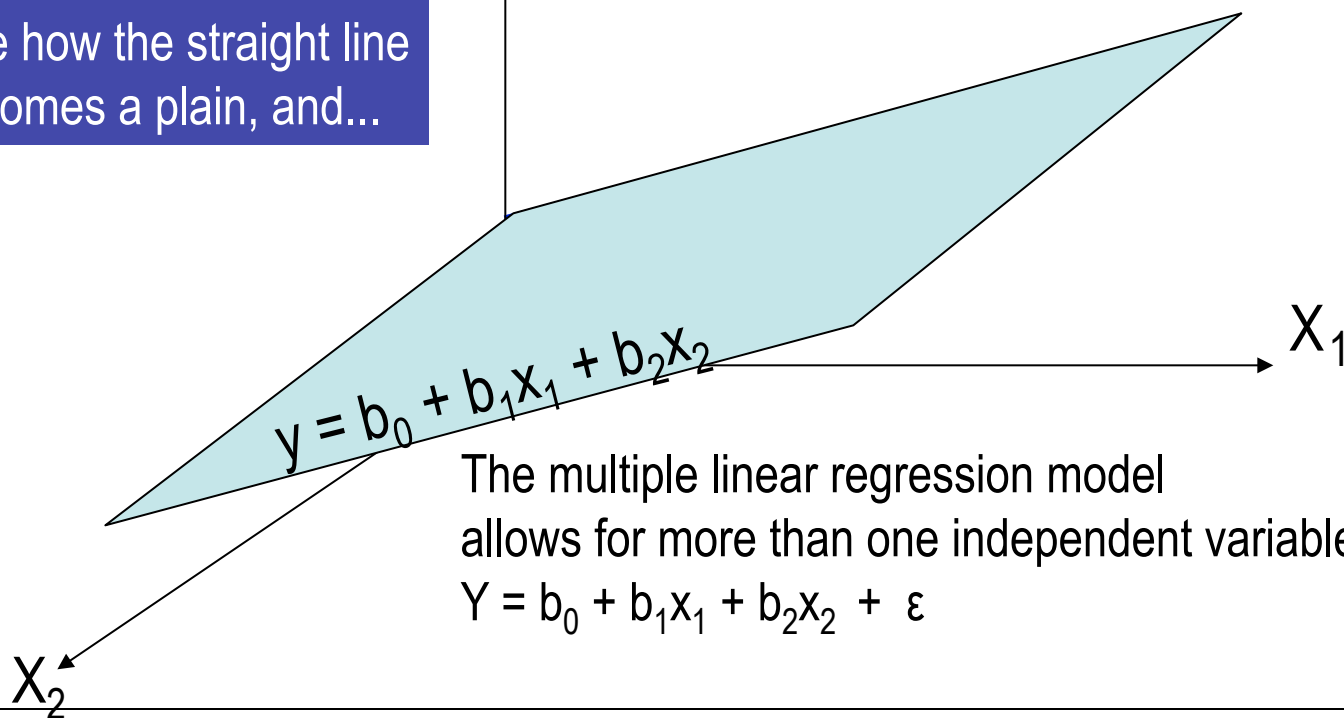
$$b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 = \sum X_1 Y$$

$$b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 = \sum X_2 Y$$

Note how the straight line becomes a plain, and...

The simple linear regression model allows for one independent variable, "x"

$$y = b_0 + b_1 x + \varepsilon$$



The multiple linear regression model allows for more than one independent variable.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon$$

The Model...

We now assume we have k independent variables *potentially* related to the one dependent variable. This relationship is represented in this first order linear equation:

The diagram shows the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$ with several annotations:

- A purple box around y is labeled "dependent variable" in purple text above it.
- Red boxes around x_1 , x_2 , and x_k are labeled "independent variables" in red text above them, with red arrows pointing to each box.
- Blue boxes around β_0 , β_1 , β_2 , and β_k are labeled "coefficients" in blue text below them, with blue arrows pointing to each box.
- A black box around ε is labeled "error variable" in black text below it.

In the one variable, two dimensional case we drew a regression line; here we imagine a *response surface*.

Estimating the Coefficients...

The sample regression equation is expressed as:

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

The Least Squares Method

- Least Squares Criterion $\min \sum (y_i - \hat{y}_i)^2$
- Computation of Coefficients' Values

The formulas for the regression coefficients $b_0, b_1, b_2, \dots, b_p$ involve the use of matrix algebra. We will rely on computer software packages to perform the calculations.

- A Note on Interpretation of Coefficients

b_i represents an estimate of the change in y corresponding to a one-unit change in x_i when all other independent variables are held constant.

Required Conditions...

For these regression methods to be valid the following four conditions for the error variable (ε) must be met:

- The probability distribution of the error variable (ε) is normal.
- The mean of the error variable is 0.
- The standard deviation of ε is σ_{ε} , which is a constant.
- The errors are independent.

Regression Analysis Steps...

- ❶ Use a computer and software to *generate the coefficients* and the statistics used to assess the model.
- ❷ Diagnose *violations of required conditions*. If there are problems, attempt to remedy them.
- ❸ *Assess the model's fit.*
 - standard error of estimate,
 - coefficient of determination,
 - F-test of the analysis of variance.
- ❹ If ❶, ❷, and ❸ are OK, use the model to interpret the coefficients, and predict or estimate the expected value of the dependent variable.

Hint: eyeballing before and after step 1.

Example 17.1

La Quinta Motor Inns is a moderately priced chain of motor inns located across the United States. Its market is the frequent business traveler.

The chain recently launched a campaign to increase market share by building new inns. The management of the chain is aware of the difficulty in choosing locations for new motels. Moreover, making decisions without adequate information often results in poor decisions.

Consequently the chain management acquired data on 100 randomly selected inns belonging to La Quinta. The objective was to predict which sites are likely to be profitable.

Example 17.1



Example 17.1

To measure profitability La Quinta used *operating margin*, which is the ratio of the sum of profit, depreciation, and interest expenses divided by total revenue.

The higher the operating margin, the greater the success of the inn.

La Quinta defines profitable inns as those with an operating margin in excess of 50% and unprofitable inns with margins of less than 30%.

Example 17.1

After a discussion with a number of experienced managers La Quinta decided to select one or two independent variables from each of the categories:

Competition - total number of motel and hotel rooms within 3 miles of each La Quinta inn

Market awareness - number of miles to the closest competing motel

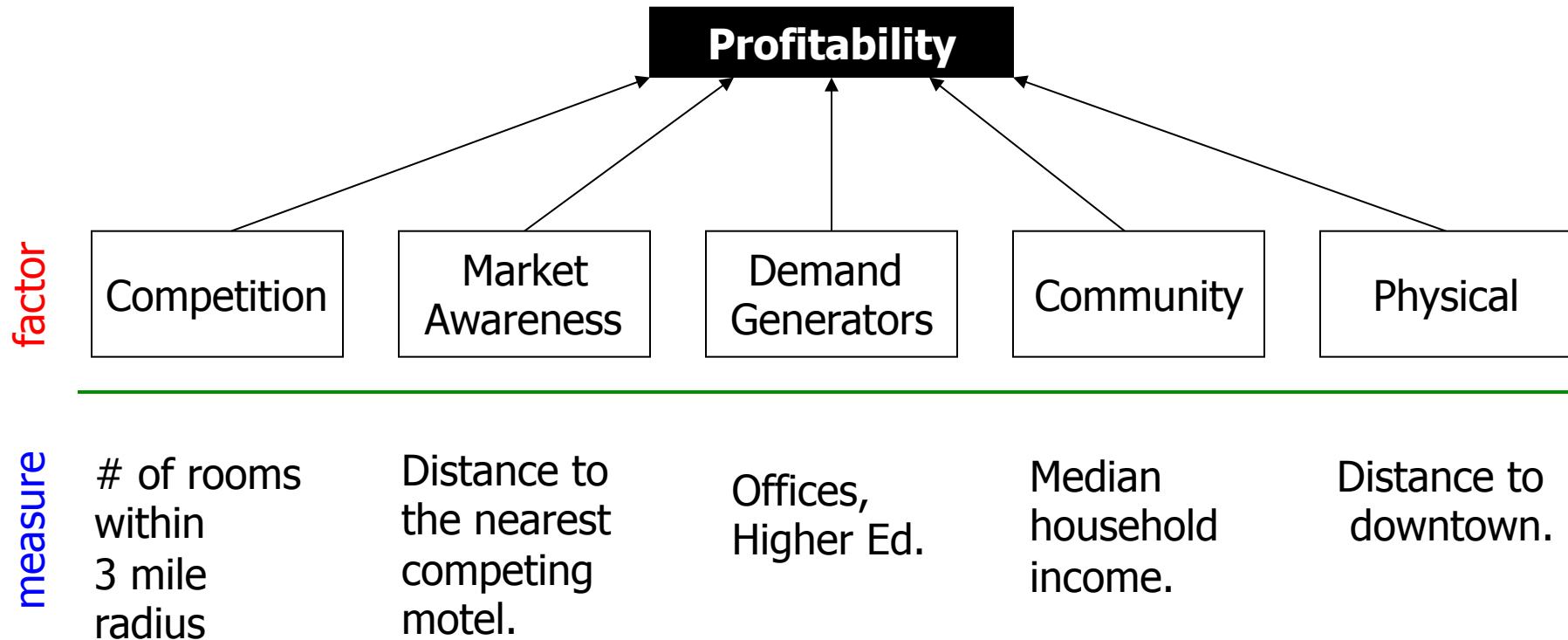
Demand generators - amount of office space and college and university enrollment in the surrounding community

Demographics - median household income

Physical - the location La Quinta chose the distance to the downtown core

Example 17.1 – La Quinta Inns...

Where should La Qunita locate a new motel? Factors influencing profitability...



*these need to be *interval* data !

Example 17.1 – La Quinta Inns...

Where should La Qunita locate a new motel?

Several possible predictors of profitability were identified, and data ([Xm17-01](#)) were collected. Its believed that operating margin (y) is dependent upon these factors:

x_1 = Total motel and hotel rooms within 3 mile radius

x_2 = Number of miles to closest competition

x_3 = Volume of office space in surrounding community

x_4 = College and university student numbers in community

x_5 = Median household income in community

x_6 = Distance (in miles) to the downtown core.

Transformation...

Can we transform this data:

	A	B	C	D	E	F	G
1	Margin	Number	Nearest	Office Space	Enrollment	Income	Distance
2	55.5	3203	4.2	549	8	37	2.7
3	33.8	2810	2.8	496	17.5	35	14.4
4	49	2890	2.4	254	20	35	2.6
5	31.9	3422	3.3	434	15.5	38	12.1
6	57.4	2687	0.9	678	15.5	42	6.9
7	49	3759	2.9	635	19	33	10.8
8	46	2341	2.3	580	23	29	7.4
9	50.2	3021	1.7	572	8.5	41	5.5
10	46	2655	1.1	666	22	34	8.1

into a mathematical model that looks like this:

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

margin

competition
(i.e. # of rooms)

awareness
(distance to
nearest alt.)

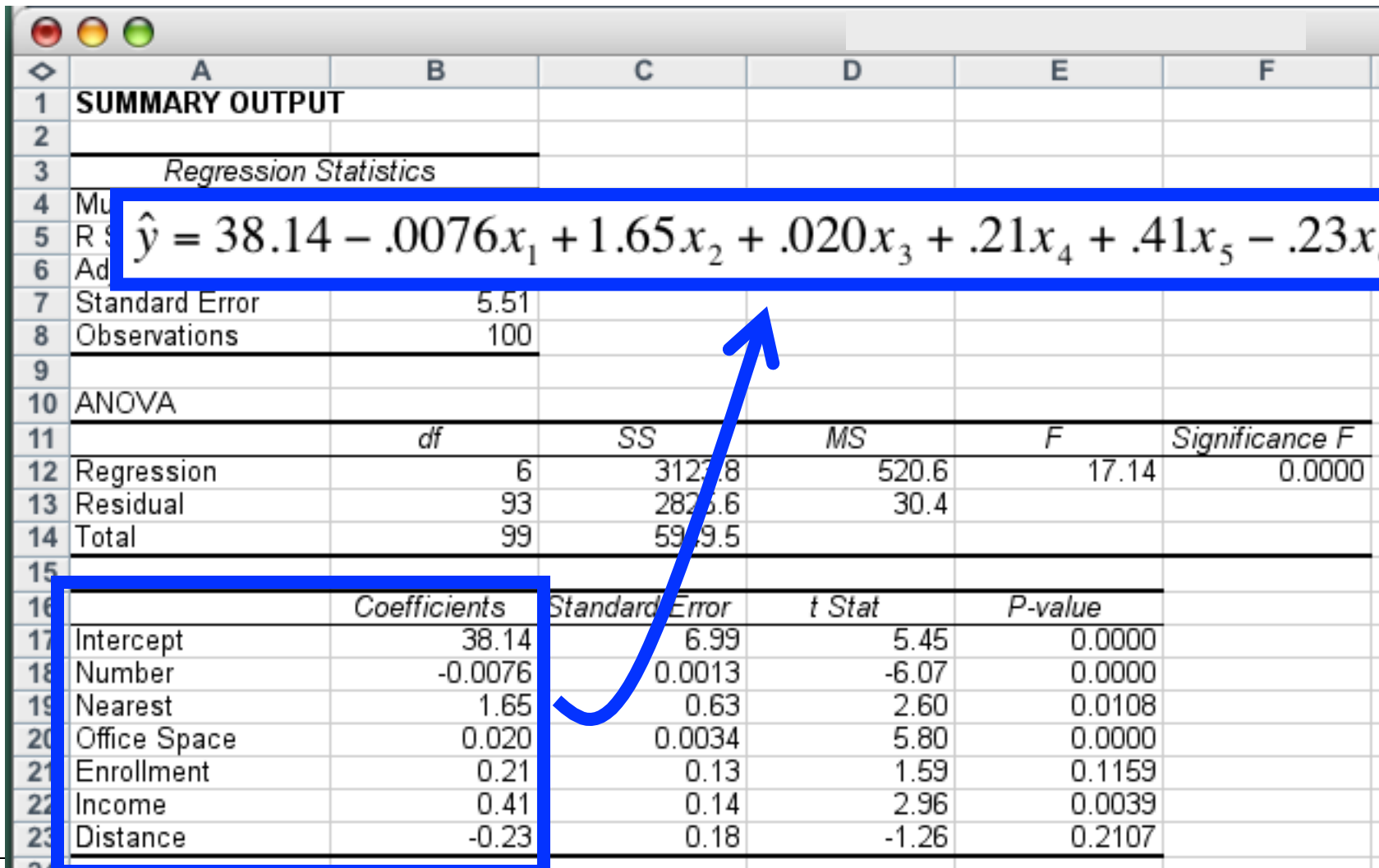
...

physical
(distance to
downtown)

Example 17.1...

COMPUTE

In Excel: Data > Data Analysis > Regression



The image shows an Excel window with a regression analysis output. A blue box highlights the regression equation $\hat{y} = 38.14 - .0076x_1 + 1.65x_2 + .020x_3 + .21x_4 + .41x_5 - .23x_6$ in cell D4. Another blue box highlights the coefficients table from row 16 to 23. A blue arrow points from the coefficients table to the regression equation.

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	<i>Regression Statistics</i>					
4	Mu	$\hat{y} = 38.14 - .0076x_1 + 1.65x_2 + .020x_3 + .21x_4 + .41x_5 - .23x_6$				
5	R S					
6	Ad					
7	Standard Error	5.51				
8	Observations	100				
9						
10	ANOVA					
11		df	SS	MS	F	Significance F
12	Regression	6	3127.8	520.6	17.14	0.0000
13	Residual	93	2825.6	30.4		
14	Total	99	5953.5			
15						
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
17	Intercept	38.14	6.99	5.45	0.0000	
18	Number	-0.0076	0.0013	-6.07	0.0000	
19	Nearest	1.65	0.63	2.60	0.0108	
20	Office Space	0.020	0.0034	5.80	0.0000	
21	Enrollment	0.21	0.13	1.59	0.1159	
22	Income	0.41	0.14	2.96	0.0039	
23	Distance	-0.23	0.18	-1.26	0.2107	

The Model...

INTERPRET

Although we haven't done any assessment of the model yet, at first pass:

$$\hat{y} = 38.14 - .0076x_1 + 1.65x_2 + .020x_3 + .21x_4 + .41x_5 - .23x_6$$

it suggests that *increases* in the number of miles to closest competition, office space, student enrollment and household income will *positively impact* the operating margin.

Likewise, increases in the total number of lodging rooms within a short distance and the distance from downtown will *negatively impact* the operating margin...

Model Assessment...

We will assess the model in three ways:

Standard error of estimate,
Coefficient of determination, and
F-test of the analysis of variance.

Standard Error of Estimate...

In multiple regression, the *standard error of estimate* is defined as:

$$s_{\varepsilon} = \sqrt{\frac{SSE}{n - k - 1}}$$

n is the sample size and k is the number of independent variables in the model. We compare this value to the mean value of y:

$$s_{\varepsilon} = 5.51 \text{ compared to } \bar{y} = 45.739$$

Standard Error	5.51
----------------	------

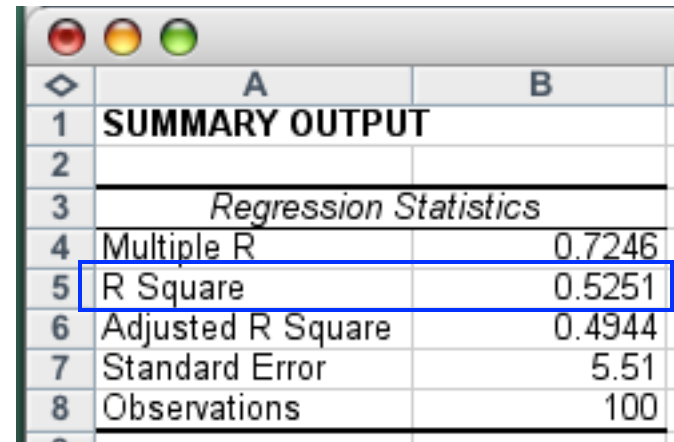
~~calculate~~

It seems the standard error of estimate is not particularly small. What can we conclude?

Coefficient of Determination...

Again, the coefficient of determination is defined as:

$$R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$



A screenshot of a software-generated regression summary output table. The table has two columns, 'A' and 'B'. Row 1 is the title 'SUMMARY OUTPUT'. Row 3 is the section header 'Regression Statistics'. Row 4 shows 'Multiple R' as 0.7246. Row 5 shows 'R Square' as 0.5251, which is highlighted with a blue border. Row 6 shows 'Adjusted R Square' as 0.4944. Row 7 shows 'Standard Error' as 5.51. Row 8 shows 'Observations' as 100.

	A	B
1	SUMMARY OUTPUT	
2		
3	<i>Regression Statistics</i>	
4	Multiple R	0.7246
5	R Square	0.5251
6	Adjusted R Square	0.4944
7	Standard Error	5.51
8	Observations	100

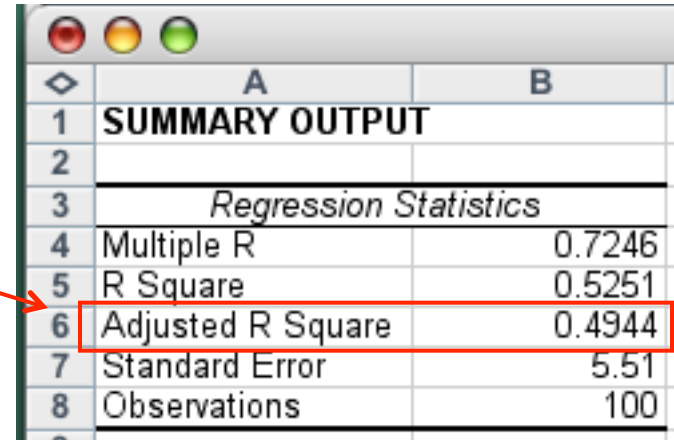
This means that 52.51% of the variation in operating margin is explained by the six independent variables, but **47.49% remains unexplained.**

Adjusted R² value...

What's this?

The “adjusted” R² is:

*the coefficient of
determination adjusted for degrees of freedom.*



A screenshot of a software window displaying a regression summary output table. The table has two columns, A and B. Row 1 is 'SUMMARY OUTPUT'. Row 3 is 'Regression Statistics'. Row 4 shows 'Multiple R' as 0.7246. Row 5 shows 'R Square' as 0.5251. Row 6 shows 'Adjusted R Square' as 0.4944, which is highlighted with a red rectangular box. Row 7 shows 'Standard Error' as 5.51. Row 8 shows 'Observations' as 100. A red arrow points from the text 'What's this?' to the 'Adjusted R Square' row.

	A	B
1	SUMMARY OUTPUT	
2		
3	Regression Statistics	
4	Multiple R	0.7246
5	R Square	0.5251
6	Adjusted R Square	0.4944
7	Standard Error	5.51
8	Observations	100

It takes into account the sample size **n**, and **k**, the number of independent variables, and is given by:

$$\text{Adjusted } R^2 = 1 - \frac{SSE / (n - k - 1)}{\sum (y_i - \bar{y})^2 / (n - 1)}$$

Testing the Validity of the Model...

In a multiple regression model (i.e. more than one independent variable), we utilize an *analysis of variance* technique to test the overall validity of the model. Here's the idea:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

H_1 : At least one β_i is not equal to zero.

If the null hypothesis is true, none of the independent variables is linearly related to y , and so the model is invalid.

If at least one β_i is not equal to 0, the model does have some validity.

Testing the Validity of the Model...

ANOVA table for regression analysis...

Source of Variation	degrees of freedom	Sums of Squares	Mean Squares	F-Statistic
Regression	k	SSR	MSR = SSR/k	F=MSR/MSE
Error	n-k-1	SSE	MSE = SSE/(n-k-1)	
Total	n-1	$\sum (y_i - \bar{y})^2$		

10	ANOVA					
11		df	SS	MS	F	Significance F
12	Regression	6	3123.8	520.6	17.14	0.0000
13	Residual	93	2825.6	30.4		
14	Total	99	5949.5			

A **large value of F** indicates that most of the variation in y is **explained** by the regression equation and that the model is valid. A **small value of F** indicates that most of the variation in y is **unexplained**.

Testing the Validity of the Model...

Our rejection region is:

$$F > F_{\alpha,k,n-k-1}$$

$$F_{\alpha,k,n-k-1} = F_{.05,6,100-6-1} \approx 2.17$$

F	Significance F
17.14	0.0000

Since Excel calculated the F statistic as $F = 17.14$ and our $F_{\text{Critical}} = 2.17$, (and the **p-value** is zero) we reject H_0 in favor of H_1 , that is:

*“there is a great deal of evidence to infer
that the model is valid”*

SSE	S_{ε}	R^2	F	Assessment of Model
0	0	1	∞	Perfect
small	small	close to 1	large	Good
large	large	close to 0	small	Poor
$\sum (y_i - \bar{y})^2$	$\sqrt{\frac{\sum (y_i - \bar{y})^2}{n - k - 1}}$	0	0	Useless

Once we're satisfied that the model fits the data as well as possible, and that the required conditions are satisfied, we can interpret and test the individual coefficients and use the model to predict and estimate...

Interpreting the Coefficients*

Intercept (b_0) 38.14 • This is the average operating margin when all of the independent variables are zero. It's meaningless to try and interpret this value, particularly if 0 is outside the range of the values of the independent variables (as is the case here).

of motel and hotel rooms (b_1) $-.0076$ • Each *additional* room within three miles of the La Quinta inn, will *decrease* the operating margin. (I.e. for each additional 1000 rooms the margin decreases by 7.6%)

Distance to nearest competitor (b_2) 1.65 • For each *additional* mile that the nearest competitor is to a La Quinta inn, the average operating margin *increases* by 1.65%.

*in each case we assume all other variables are held constant...

Interpreting the Coefficients*

Office space (b_3) .020 • For each *additional* thousand square feet of office space, the margin will *increase* by .020. E.g. an extra 100,000 square feet of office space will increase margin (on average) by 2.0%.

Student enrollment (b_4) .21 • For each *additional thousand* students, the average operating margin *increases* by .21%

Median household income (b_5) .41 • For each *additional* thousand dollar increase in median household income, the average operating margin *increases* by .41%

Distance to downtown core (b_6) –.23 • For each *additional* mile to the downtown center, the operating margin *decreases* on average by .23%

*in each case we assume all other variables are held constant...

Testing the Coefficients...

For *each* independent variable, we can test to determine whether there is enough evidence of a linear relationship between it and the dependent variable for the entire population...

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

(for $i = 1, 2, \dots, k$) and using:
$$t = \frac{b_i - \beta_i}{s_{b_i}}$$

as our test statistic (with $n-k-1$ degrees of freedom).

Testing the Coefficients

INTERPRET

We can use our Excel output to quickly test each of the six coefficients in our model...

		Coefficients	Standard Error	t Stat	P-value
16					
17	Intercept	38.14	6.99	5.45	0.0000
18	Number	-0.0076	0.0013	-6.07	0.0000
19	Nearest	1.65	0.63	2.60	0.0108
20	Office Space	0.020	0.0034	5.80	0.0000
21	Enrollment	0.21	0.13	1.59	0.1159
22	Income	0.41	0.14	2.96	0.0039
23	Distance	-0.23	0.18	-1.26	0.2107

Thus, the number of hotel and motel rooms, distance to the nearest motel, amount of office space, and median household income are linearly related to the operating margin. There is no evidence to infer that college enrollment and distance to downtown center are linearly related to operating margin.

Using the Regression Equation

Much like we did with simple linear regression, we can produce a ***prediction interval*** for a particular value of y .

As well, we can produce the ***confidence interval estimate*** of the expected value of y .

Excel's tools will do the work; our role is to set up the problem, understand and interpret the results.

Using the Regression Equation

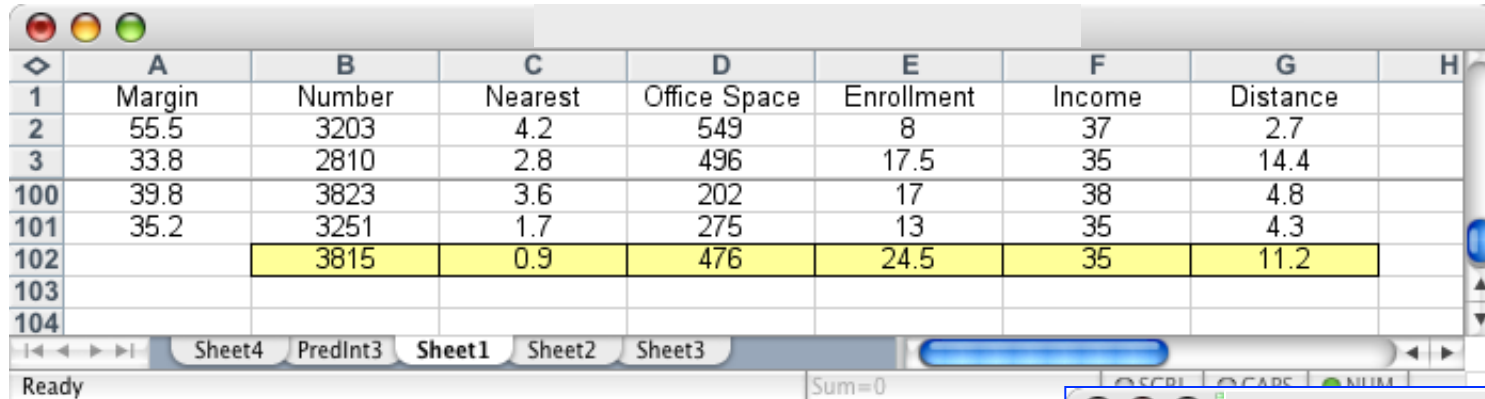
Predict the operating margin if a La Quinta Inn is built at a location where...

- ❶ There are 3815 rooms within 3 miles of the site.
- ❷ The closest other hotel or motel is .9 miles away.
- ❸ The amount of office space is 476,000 square feet.
- ❹ There is one college and one university nearby with a total enrollment of 24,500 students.
- ❺ Census data indicates the median household income in the area (rounded to the nearest thousand) is \$35,000, and,
- ❻ The distance to the downtown center is 11.2 miles.

our x_i 's...

Using the Regression Equation

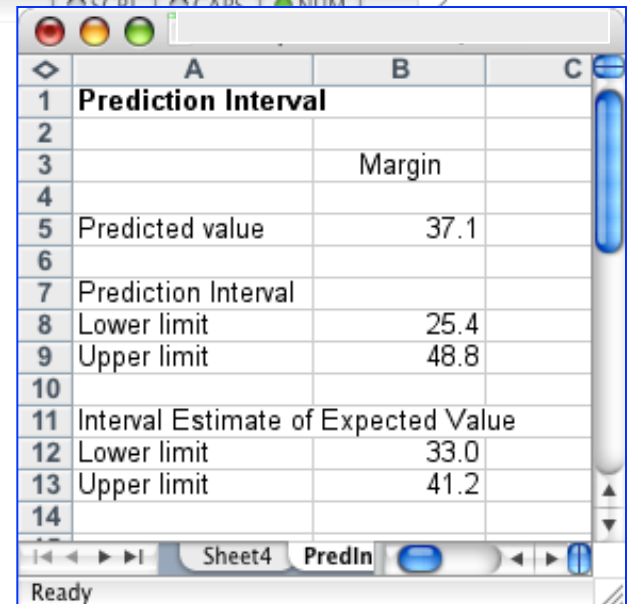
We add one row (our given values for the independent variables) to the bottom of our data set:



	A	B	C	D	E	F	G	H
1	Margin	Number	Nearest	Office Space	Enrollment	Income	Distance	
2	55.5	3203	4.2	549	8	37	2.7	
3	33.8	2810	2.8	496	17.5	35	14.4	
100	39.8	3823	3.6	202	17	38	4.8	
101	35.2	3251	1.7	275	13	35	4.3	
102		3815	0.9	476	24.5	35	11.2	
103								
104								

Then we use:

Add-Ins > Data Analysis Plus >
Prediction Interval
to crunch the numbers...



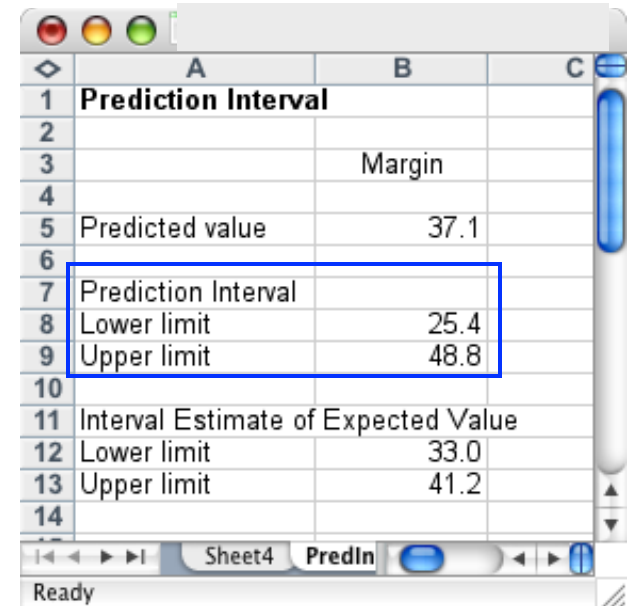
	A	B	C
1	Prediction Interval		
2			
3		Margin	
4			
5	Predicted value	37.1	
6			
7	Prediction Interval		
8	Lower limit	25.4	
9	Upper limit	48.8	
10			
11	Interval Estimate of Expected Value		
12	Lower limit	33.0	
13	Upper limit	41.2	
14			

Prediction Interval...

INTERPRET

We predict that the operating margin will fall between 25.4 and 48.8.

If management defines a profitable inn as one with an operating margin greater than 50% and an unprofitable inn as one with an operating margin below 30%, they will pass on this site, since *the entire prediction interval* is below 50%.



The screenshot shows a spreadsheet with the following data:

	A	B	C
1	Prediction Interval		
2			
3		Margin	
4			
5	Predicted value	37.1	
6			
7	Prediction Interval		
8	Lower limit	25.4	
9	Upper limit	48.8	
10			
11	Interval Estimate of Expected Value		
12	Lower limit	33.0	
13	Upper limit	41.2	
14			

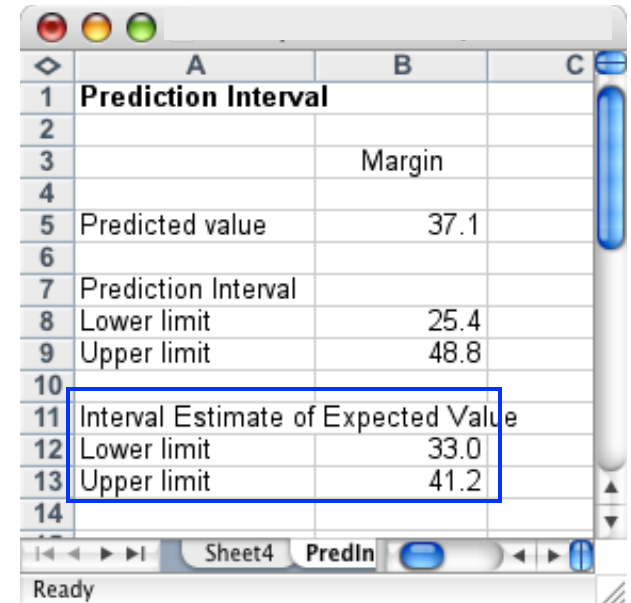
The spreadsheet interface includes a status bar at the bottom showing "Ready" and a tab labeled "Sheet4 PredIn". A blue box highlights the "Prediction Interval" section (rows 7-9).

Confidence Interval

INTERPRET

The expected operating margin of **all** sites that fit this category is estimated to be between 33.0 and 41.2.

We interpret this to mean that if we built inns on an infinite number of sites that fit the category described, the *mean operating margin* would fall between 33.0 and 41.2. In other words, the average inn *would not be profitable* either...



A screenshot of a spreadsheet window titled 'Sheet4 PredIn'. The spreadsheet displays data for a prediction interval and a confidence interval. The 'Interval Estimate of Expected Value' section is highlighted with a blue box.

	A	B	C
1	Prediction Interval		
2			
3		Margin	
4			
5	Predicted value	37.1	
6			
7	Prediction Interval		
8	Lower limit	25.4	
9	Upper limit	48.8	
10			
11	Interval Estimate of Expected Value		
12	Lower limit	33.0	
13	Upper limit	41.2	
14			

Regression Diagnostics

- Calculate the residuals and check the following:
- *Is the error variable nonnormal?*
- Draw the histogram of the residuals
- *Is the error variance constant?*
- Plot the residuals versus the predicted values of y .
- *Are the errors independent (time-series data)?*
- Plot the residuals versus the time periods.
- *Are there observations that are inaccurate or do not belong to the target population?*
- Double-check the accuracy of outliers and influential observations.

IQ and Physical Characteristics

Are a person's brain size and body size predictive of his or her intelligence? (Willerman, *et al*, 1991)

- $n=38$
- Response (y): Performance IQ scores (**PIQ**) from the revised Wechsler Adult Intelligence Scale. This variable served as the investigator's measure of the individual's intelligence.
- Potential predictor (x_1): Brain size based on the count obtained from **MRI** scans (given as count/10,000).
- Potential predictor (x_2): **Height** in inches.
- Potential predictor (x_3): **Weight** in pounds.

IQ and Physical Characteristics

As always, the first thing we should want to do when presented with a set of data is to **plot it**.

Plotting the data is a little more challenging in the multiple regression setting, as there is one scatter plot for each pair of variables.

Not only do we have to consider the relationship between the response and each of the predictors, we also have to consider how the predictors are related among each other.

IQ and Physical Characteristics

Result:

- The R^2 value is 29.49%. This tells us that 29.49% of the variation in intelligence, as quantified by PIQ, is reduced by taking into account brain size, height and weight.
- The P -values for the t -tests appearing in the table of estimates suggest that the slope parameters for Brain ($P = 0.001$) and Height ($P = 0.033$) are significantly different from 0, while the slope parameter for Weight ($P = 0.998$) is not.
- The P -value for the analysis of variance F -test ($P = 0.007$) suggests that the model containing Brain, Height and Weight is more useful in predicting intelligence than not taking into account the three predictors. (Note that this does not tell us that the model with the three predictors is the *best* model!)

Underground Air Quality

What are the breathing habits of baby birds that live in underground burrows?

Some mammals burrow into the ground to live. Scientists have found that the quality of the air in these burrows is not as good as the air aboveground. In fact, some mammals change the way that they breathe in order to accommodate living in the poor air quality conditions underground.

Some researchers (Colby, *et al*, 1987) wanted to find out if nestling bank swallows, which live in underground burrows, also alter how they breathe.



Underground Air Quality

What are the breathing habits of baby birds that live in underground burrows?

The researchers conducted a randomized experiment on $n = 120$ nestling bank swallows. In an underground burrow, they varied the percentage of oxygen at four different levels (13%, 15%, 17%, and 19%) and the percentage of carbon dioxide at five different levels (0%, 3%, 4.5%, 6%, and 9%). Under each of the resulting $5 \times 4 = 20$ experimental conditions, the researchers observed the total volume of air breathed per minute for each of 6 nestling bank swallows.



Underground Air Quality

What are the breathing habits of baby birds that live in underground burrows? (Colby, *et al*, 1987)

- $n=120$
- Response (y): percentage increase in "minute ventilation," (**Vent**), *i.e.*, total volume of air breathed per minute.
- Potential predictor (x_1): percentage of oxygen (**O2**) in the air the baby birds breathe. (13%, 15%, 17%, and 19%)
- Potential predictor (x_2): percentage of carbon dioxide (**CO2**) in the air the baby birds breathe. (0%, 3%, 4.5%, 6%, and 9%)

Underground Air Quality

The plot between percentage of oxygen (**O₂**) and percentage of carbon dioxide (**CO₂**) is the classical appearance of a scatter plot for the experimental conditions.

The plot suggests that there is no correlation at all between the two variables. You should be able to observe from the plot the 4 levels of **O₂** and the 5 levels of **CO₂** that make up the $5 \times 4 = 20$ experimental conditions.

Underground Air Quality

Result:

- Only 26.82% of the variation in minute ventilation is reduced by taking into account the percentages of oxygen and carbon dioxide.
- The P -values for the t -tests suggest that the slope parameter for carbon dioxide level ($P < 0.001$) is significantly different from 0, while the slope parameter for oxygen level ($P = 0.408$) is not. Does this conclusion appear consistent with the above scatter plot matrix? Yes!
- The P -value for the analysis of variance F -test ($P < 0.001$) suggests that the model containing oxygen and carbon dioxide levels is more useful in predicting minute ventilation than not taking into account the two predictors. (Again, the F -test does not tell us that the model with the two predictors is the *best* model!)

Pastry Sweetness Data

How moisture content and sweetness of a pastry product affect a taster's rating of the product?

In a designed experiment, the eight possible combinations of four moisture levels and two sweetness levels are studied. Two pastries are prepared and rated for each of the eight combinations.

(Data source: *Applied Regression Models*, (4th edition), Kutner, Neter, and Nachtsheim).

Pastry Sweetness Data

How moisture content and sweetness of a pastry product affect a taster's rating of the product?

- $n=16$
- Response (y): Rating of the pastry.
- Potential predictor (x_1): Moisture. (4, 6, 8, and 10)
- Potential predictor (x_2): Sweetness. (2 and 4)

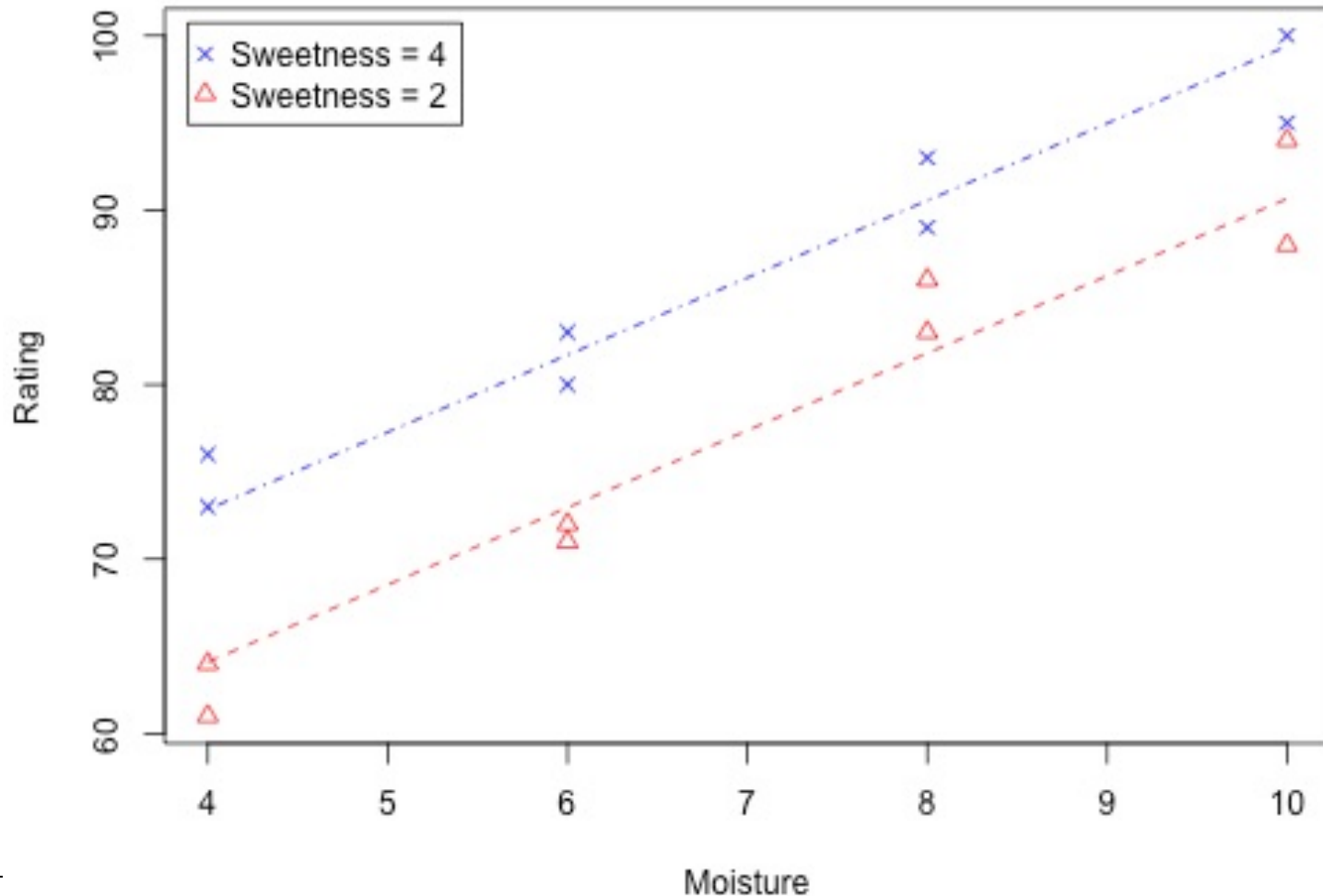
Pastry Sweetness Data

The plot between moisture and sweetness is the classical appearance of a scatter plot for the experimental conditions.

Notice that the points are on a rectangular grid so the correlation between the two variables is 0.

Pastry Sweetness Data

The following figure shows how the two x-variables affect the pastry rating.



Pastry Sweetness Data

Result:

- The sample coefficient that multiplies **Moisture** is 4.425 in both the simple and the multiple regression. The sample coefficient that multiplies **Sweetness** is 4.375 in both the simple and the multiple regression. This result does not generally occur; the only reason that it does in this case is that **Moisture** and **Sweetness** are not correlated, so the estimated slopes are independent of each other. For most observational studies, predictors are typically correlated and estimated slopes in a multiple linear regression model do not match the corresponding slope estimates in simple linear regression models.

Pastry Sweetness Data

Result:

- The R^2 for the multiple regression, 95.21%, is the sum of the R^2 values for the simple regressions (79.64% and 15.57%). Again, this will only happen when we have uncorrelated x -variables.
- The variable **Sweetness** is not statistically significant in the simple regression ($p = 0.130$), but it is in the multiple regression. **This is a benefit of doing a multiple regression.** By putting both variables into the equation, we have greatly reduced the standard deviation of the residuals. This in turn reduces the standard errors of the coefficients, leading to greater t -values and smaller p -values.

Female Stat Students' height

- $n=214$
- Response (y): student's self-reported height
- Potential predictor (x_1): student's guess at her mother's height
- Potential predictor (x_2): student's guess at her father's height.
- All heights are in inches

Female Stat Students' height

Result

- The p -values given for the two x -variables tell us that student height is significantly related to each.
- The value of $R^2 = 43.35\%$ means that the model (the two x -variables) explains 43.35% of the observed variation in student heights.

Hospital Example

Data from $n = 113$ hospitals in the United States are used to assess factors related to the likelihood that a hospital patient acquires an infection while hospitalized.

- $n=113$
- Response (y): infection risk
- Potential predictor (x_1): average length of patient stay
- Potential predictor (x_2): average patient age
- Potential predictor (x_3): measure of how many x-rays are given in the hospital

(Data source: *Applied Regression Models*, (4th edition), Kutner, Neter, and Nachtsheim).

Hospital Example

Result

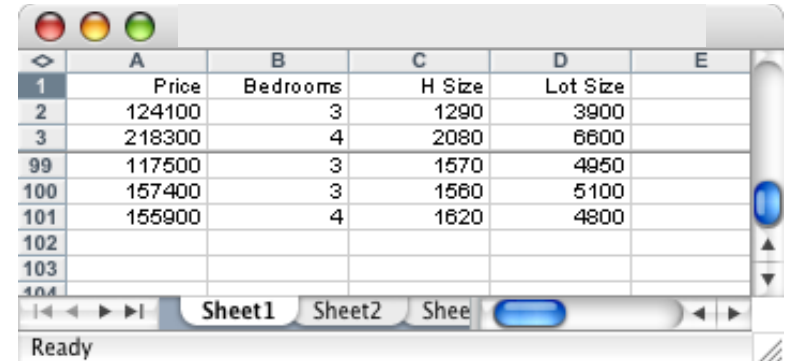
- The p -value for testing the coefficient of **Age** is 0.330. Thus we cannot reject the null hypothesis $H_0: \beta_2 = 0$. The variable **Age** is not a useful predictor within this model that includes **Stay** and **Xrays**.
- For the variables **Stay** and **X-rays**, the p -values for testing their coefficients are at a statistically significant level so both are useful predictors of infection risk (within the context of this model!).
- We usually don't worry about the p -value for Constant. It has to do with the “intercept” of the model and seldom has any practical meaning. It also doesn't give information about how changing an x -variable might change y -values

Regression Diagnostics

- Multiple regression models have a problem that simple regressions do not, namely *multicollinearity*.
- It happens when the *independent variables* are highly *correlated*.
- We'll explore this concept through the following example...

Example 17.2

- A real estate agent wanted to develop a model to predict the selling price of a home. The agent believed that the most important variables in determining the price of a house are its:
 - ① size,
 - ② number of bedrooms,
 - ③ and lot size.
- The proposed model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$
- Housing market data has been gathered and Excel is the analysis tool of choice



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E
1	Price	Bedrooms	H Size	Lot Size	
2	124100	3	1290	3900	
3	218300	4	2080	6600	
99	117500	3	1570	4950	
100	157400	3	1560	5100	
101	155900	4	1620	4800	
102					
103					
104					

The spreadsheet has tabs for 'Sheet1', 'Sheet2', and 'Sheet3'. The status bar at the bottom indicates 'Ready'.

Example 17.2

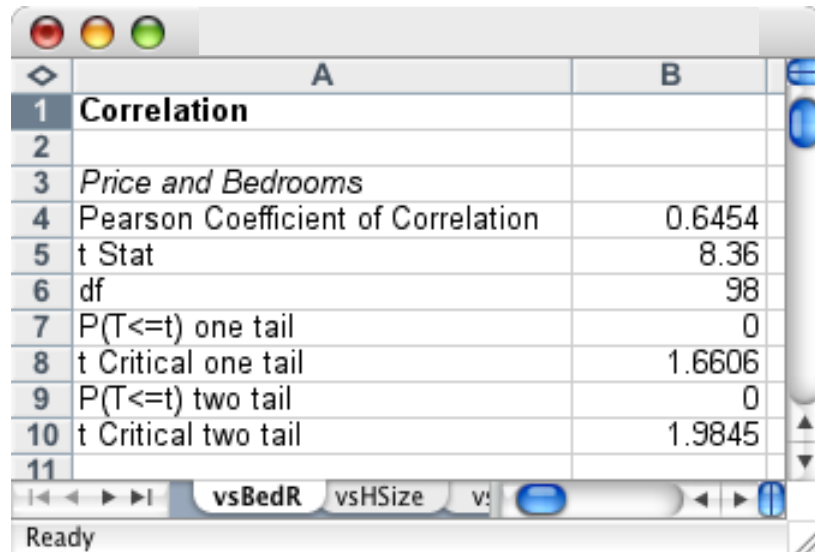
- Data > Data Analysis > Regression

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	<i>Regression Statistics</i>					
4	Multiple R	0.7483				
5	R Square	0.5600				
6	Adjusted R Square	0.5462				
7	Standard Error	25023				
8	Observations	100				
9						
10	ANOVA					
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
12	Regression	3	76501718347	25500572782	40.73	0.0000
13	Residual	96	60109046053	626135896		
14	Total	99	136610764400			
15						
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
17	Intercept	37718	14177	2.66	0.0091	
18	Bedrooms	2306	6994	0.33	0.7423	
19	H Size	74.30	52.98	1.40	0.1640	
20	Lot Size	-4.36	17.02	-0.26	0.7982	

The F-test indicates the model is valid...

...but these t-stats suggest none of the variables are related to the selling price.

Example 17.2 (Data Analysis Plus)

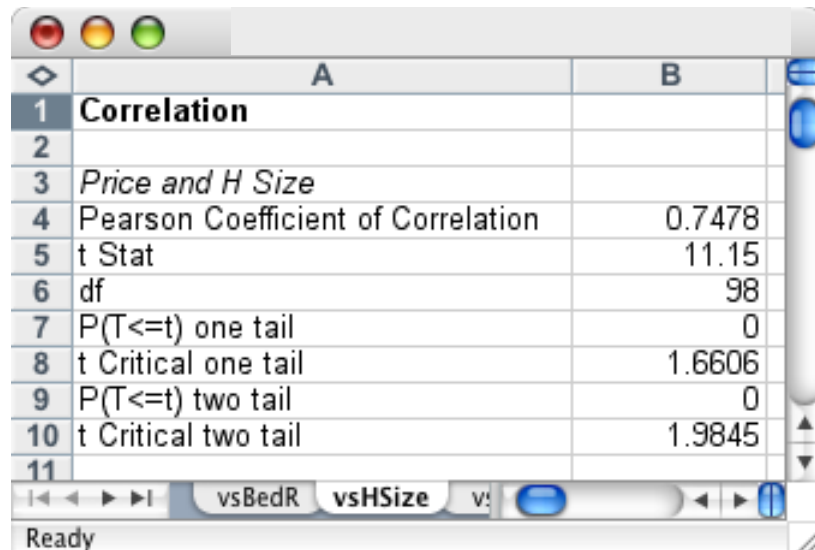


Ready

	A	B
1	Correlation	
2		
3	Price and Bedrooms	
4	Pearson Coefficient of Correlation	0.6454
5	t Stat	8.36
6	df	98
7	P(T<=t) one tail	0
8	t Critical one tail	1.6606
9	P(T<=t) two tail	0
10	t Critical two tail	1.9845
11		

vsBedR vsHSize vs

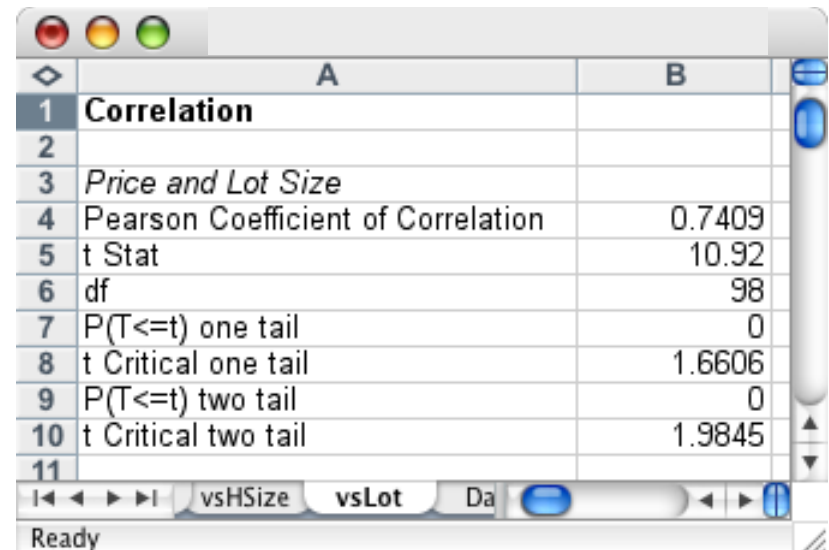
Unlike the t-tests in the multiple regression model, these three t-tests tell us that the number of bedrooms, the house size, and the lot size ***are all linearly related*** to the price...



Ready

	A	B
1	Correlation	
2		
3	Price and H Size	
4	Pearson Coefficient of Correlation	0.7478
5	t Stat	11.15
6	df	98
7	P(T<=t) one tail	0
8	t Critical one tail	1.6606
9	P(T<=t) two tail	0
10	t Critical two tail	1.9845
11		

vsBedR vsHSize vs



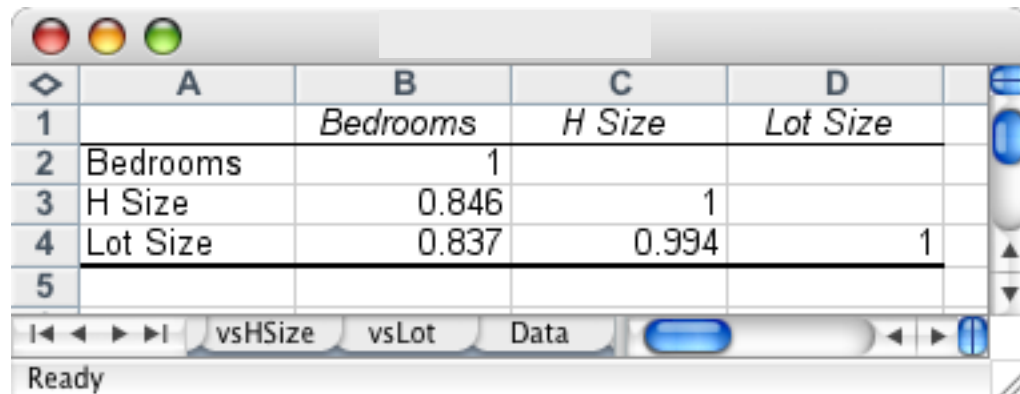
Ready

	A	B
1	Correlation	
2		
3	Price and Lot Size	
4	Pearson Coefficient of Correlation	0.7409
5	t Stat	10.92
6	df	98
7	P(T<=t) one tail	0
8	t Critical one tail	1.6606
9	P(T<=t) two tail	0
10	t Critical two tail	1.9845
11		

vsHSize vsLot Da

Example 17.2

- How to account for this apparent contradiction?
- The answer is that the *three independent variables are correlated with each other* !



	A	B	C	D
1		Bedrooms	H Size	Lot Size
2	Bedrooms	1		
3	H Size	0.846	1	
4	Lot Size	0.837	0.994	1
5				

Ready

(i.e. this is reasonable: larger houses have more bedrooms and are situated on larger lots, and smaller houses have fewer bedrooms and are located on smaller lots.)

multicollinearity affected the t-tests so that they implied that none of the independent variables is linearly related to price when, in fact, all are

Physiological Measurements

- $n=20$
- Response (y): body fat
- Potential predictor (x_1): triceps skinfold thickness
- Potential predictor (x_2): thigh circumference
- Potential predictor (x_3): midarm circumference

(Data source: *Applied Regression Models*, (4th edition), Kutner, Neter, and Nachtsheim).

Physiological Measurements

Result

- F-test: Significant
- R^2 : 0.8
- p-values for the slopes are all high!?
- A high correlation between the Triceps and Thigh variables
- It is difficult to separate the individual effects of these two variables.