Chapter 10: Regression Analysis

Read: Chapter 10 then read the notes and try the WEBCT assignment questions. If you need more practice, try the practice questions with answers available on the web.

Exercises:

Introduction

In Chapter 2 you were introduced to the method of least squares regression. This was a technique of fitting a line model to data. In the next several chapters the common techniques of inference were introduced and developed. Two of the most common inference methods were the test of hypothesis and confidence intervals. In this chapter the techniques of inference are applied to the line so that inferences can be made about the unknown parameters, the slope and intercept.

View the Video - Inference for Relationships

Review Chapter 2 Material

In this chapter we add tests of significance and confidence intervals to our line model introduced in Chapter 2. Thus, we begin with a brief review of that material. We now write the line as: $\mathbf{y = B_o + B_1 x}$

This is the equation of a straight line. It is in contrast to y=a+bx, the notation in Chapter 2.

B_1 and B_o are the parameters or unknowns in the line.

B_o is the y-intercept, the value of y when x=0.

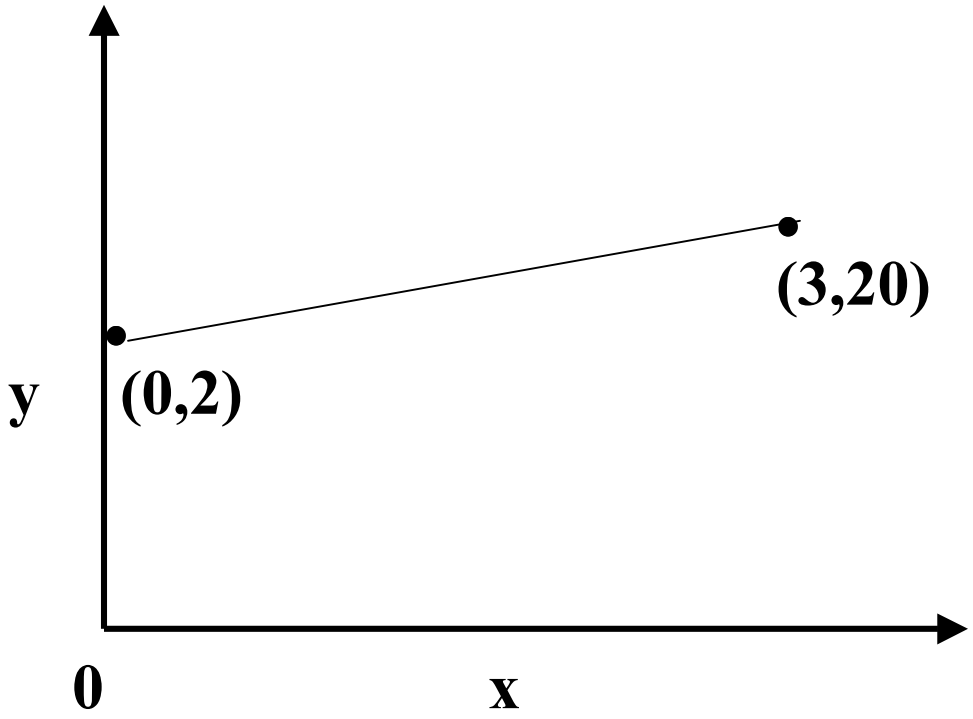
B_1 is the slope or how y changes per unit increase in x.

Example: Consider the following line

$$y = 2 + 6x$$

When x = 0, y = 2; therefore the y-intercept = 2. When x increases 1 unit, y increases by 6. In other words, the slope = 6.

In order to plot the line, pick 2 points that are on the line, say (0,2) the y-intercept and (3,20).



Suppose a researcher has a graph of x vs y for a number of points.

In chapter 2 we saw that the best fitting line was calculated using **least squares.** In other words, deviations in the vertical direction are used to select the line.

In other words the line that minimizes deviations in the vertical direction is the best line:

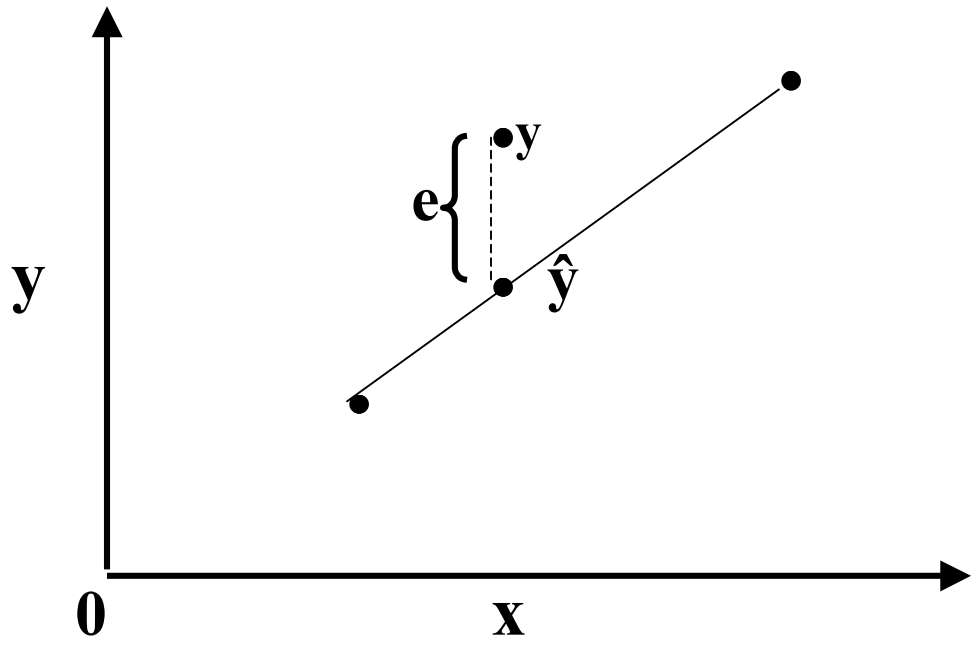
$$\hat{y}_i = b_0 + b_1 x_i$$

(read $\hat{y}_i$ as the predicted value of y by the line)

$$e_i^2 = (\text{observed y - predicted y})^2$$

residual squared

The line that minimizes e_i^2 is the best line.



We use the following formulas to find the best fitting line that minimizes the residuals about the line. Note that a computer is used to find the **least squares regression** line in practice.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

The researcher assumes that a line is an appropriate **model** for the data collected on the system under study.

The model is a line: $\mathbf{y = B_o + B_1 x}$ where B_o and B_1 are unknown **population** parameters.

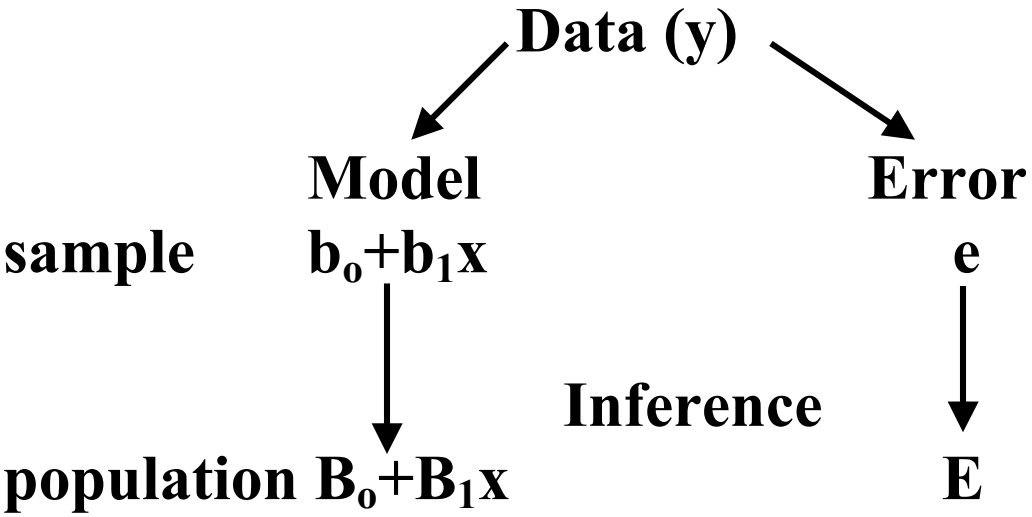
The sample data are used to make inferences about the population parameters.

In this problem, the data are viewed as composed of 2 parts:

Data = Model (fit) + Error (residual)

$$\mathbf{y = (b_o + b_1 x) + e}$$

From the data we make inferences about the population. The data give the researcher sample estimates of the population parameters. Using the information in the sample, inferences are made concerning the population.



NOTE: We have only one **response variable** (y) and one **predictor variable** (x) defined on a population **π.** Our response variable y is quantitative.

Based on the observed values of the predictor variable (x) we want to construct a line which we will use to predict values of y.

Example: Gebotys & Roberts (1987), CJBS, examined two variables, age of a person in years, and how that person views the seriousness of a crime as recorded on a scale where a higher number indicates more seriousness. The sample data for 10 randomly sampled people are given below. The population of interest is all people in Ontario over the age of 18.

age (x)	serious (y)
20	21
25	28
26	27
25	26
30	33
34	36
40	31
40	35
40	41
80	95

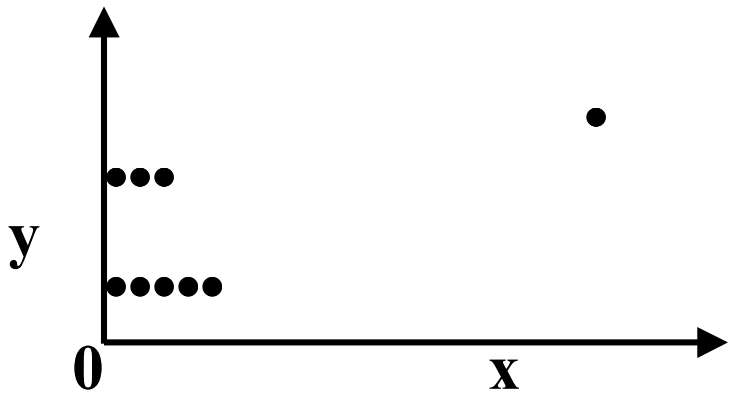
A line is assumed an appropriate model for this problem.

The means are: $\bar{x} = \mathbf{36}$ $\bar{y} = \mathbf{37.3}$

The least squares estimates for the slope and intercept are:

$$\mathbf{b_1 = 1.197}$$
$$\mathbf{b_0 = \bar{y} - b_1\bar{x} = -5.792}$$

Our equation is then $\mathbf{y = 5.792 + 1.197x}$. A plot of the data confirms the researcher's assumption that a line is a reasonable model to consider for this data.



Use table 9.1 below to verify the slope and intercept. In practice, these calculations are done using a computer. The hand calculations are given here for your information and better understanding of the method.

Table 9.1

	x_i	y_i	$(x- \bar{x})$	$(y- \bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x - \bar{x})(y- \bar{y})$
1	20	21	-16	-16.3	256	265.69	260.8
2	25	28	-11	-9.3	121	86.49	102.3
3	26	27	-10	-10.3	100	106.09	103.0
4	25	26	-11	-11.3	121	172-69	124.3
5	30	33	-6	-4.3	36	18.49	25.8
6	34	36	-2	-1.3	4	1.69	2.6
7	40	31	4	-6.3	16	39.69	.25.2
8	40	35	4	-2.3	16	5.29	-9.2
9	40	41	4	3.7	16	13.69	14.8
10	80.	95	44	57.7	1936	3329.29	2538.8
Totals	360	373	0	0	2622	3994.1	3138.0

$$\bar{x} = 36 \qquad \bar{y} = 37.3$$

The least squares estimate of B_1 the slope of the line is:

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{3138}{2622} = 1.197$$

A one unit increase in age (x) gives a 1.197 increase in the rating of crime seriousness (y).

The least squares estimate of B_0, the y-intercept of the line, is:

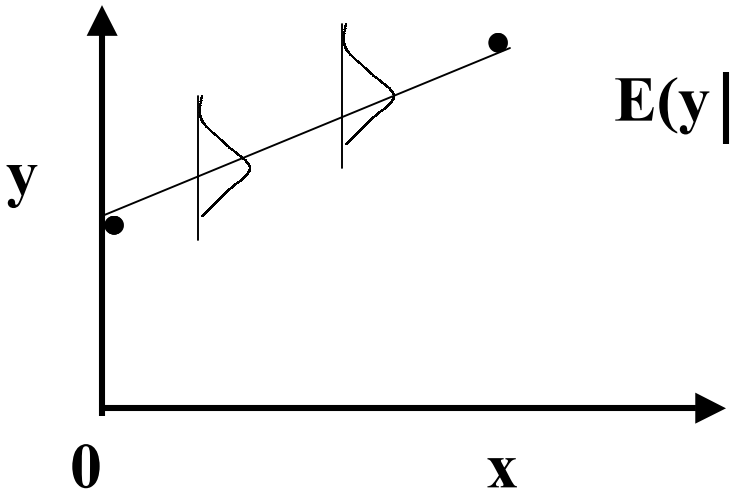
$$\mathbf{b_0 = \bar{y} - b_1\bar{x}}$$
$$\mathbf{= 37.3 - 1.197(36)}$$
$$\mathbf{= -5.792}$$

This completes a brief review of Chapter 2. We now introduce inference for the line.

Inference for the Line

For the model $y=B_o + B_1x$ the population of y values is assumed **normally distributed** about a mean that depends on x.

We write $E(y \mid x)$ to read average value of y given x to represent these means. We assume the means lie on a line given by the equation of a line. See the plots below.



$$E(y \mid x) = B_o + B_1 x$$

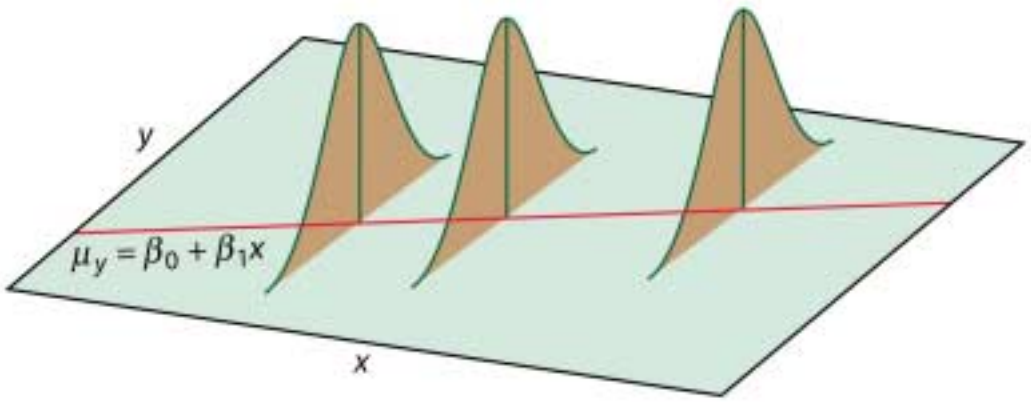


Figure 10.2

Properties of Residuals

The residuals (e_i) are assumed to:

1) Be independent
2) Follow a normal distribution with mean = 0
3) Have unknown variance equal to σ^2

We check 2) with a probability plot and simulation.We estimate σ^2 the variance in the population using s^2 our sample estimate.
The calculation of s depends on the residuals. If residuals are large (poor fit of the data by a line) then s will be large. Conversely, if they are small (line a good model for the data), "s" will be small. Table 9.2 is used to calculate s below:

Table 9.2

	x_i	y_i	$\hat{y}=b_0+b_1$	$e_i = y_i - \hat{y}_i$	e_i^2
1	20	21	18.15	2.85	8.12
2	25	28	24.13	3.87	14.98
3	26	27	25.33	1.67	2.79
4	25	26	24.13	1.87	3.50
5	30	33	30.12	2.88	8.29
6	34	36	34.91	1.09	1.19
7	40	31	42.09	-11.09	122.99
8	40	35	42.09	-7.09	50.27
9	40	41	42.09	-1.09	1.19
10	80	95	89.97	5.03	25.30

Totals	360	373	373.01	-0.01	238.62

$$s^2 = \frac{\sum e_i^2}{n - 2} = \frac{238.62}{10 - 2} = 29.83$$

In other words, the spread, or variance, of the residuals about the line is 29.83.

Researchers often report the square root of s^2 ,s the standard deviation since given the assumption of normality some quick calculations can be made about the residuals and their distribution.

We know that ± 1 standard deviation for the normal gives approximately 68% of the data. In our example:

$$s = 5.46$$

so, plus or minus 5.46 will give approximately 68% of the data. Sometimes the residuals are standardized, Z scores are created so that the

rules learned in Chapter 1 concerning the normal can easily be applied.

Standardized residuals greater than 3 or less than -3 are called outliers.

Let us examine our sample variance formula which we use to estimate our population variance given below:

$$S^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y-\hat{y})^2}{n-2}$$

n-2 is called the degrees of freedom

We subtract 2 from our sample size since we want to estimate 2 parameters (B_o,B_1) or unknowns.

σ^2 is the spread of the observations about the line in the population.

Tests of Significance & Confidence Intervals:

We previously spoke of significance tests and confidence intervals for the t distribution in Chapter 7.

In general we saw that the test of hypothesis that a parameter equals zero has the following form:

$$T = \frac{\text{estimate}}{\text{standard error of the estimate}}$$

If the sample size is large or the variance known we could use a z test to perform our test of significance. Usually, however, the variance is unknown and the sample size small (n<50).

For the slope of our line we have the following test of hypothesis:

H_o: $B_1 = 0$ (x is not related to y)

H_a: $B_1 \neq 0$ (x is related to y)

The test statistic is T, the sample slope over the corresponding standard error, is given below:

$$T = \frac{b_1}{s(b_1)}$$

The statistic is compared to the t-distribution with n–2 degrees of freedom. If T is an unreasonable value from the t distribution (i.e., $p < .05$) we have evidence against the null.

A 1 - α confidence interval for the slope is given by:

$$b_1 \pm t_{\alpha/2}\, s(b_1)$$

With the standard error is given by:

$$s(b_1) = s/\sqrt{\sum(x - \bar{x})^2}$$

For example (continued)

We want to see if age and crime seriousness are related, so a line model was fit to the data. To confirm a relationship we want to test if the slope of the line for the crime seriousness data is zero. Our test of hypothesis is written as:

$$\mathbf{H_o: B_1 = 0}$$

(age & crime seriousness are NOT related)

$$\mathbf{H_a: B_1 \neq 0}$$

(age and crime seriousness ARE related)

We saw **s = 5.46** from our previous calculation.

The standard error is:

$$\mathbf{s(b_1) = 5.46/51.21 = .11}$$

The t statistic is:

$$\mathbf{T = b_1/s(b_1) = 1.197/.11 = 10.88}$$

with **degrees of freedom** n-2 = 10-2 = **8 df**

The p-value is less than .0001 indicating we have very strong evidence against the null hypothesis. Therefore, we reject H_o: $B_1 = 0$, and conclude that age is useful in predicting a judgement of crime seriousness.

A 95% CI for the slope is given by $b_1 \pm t\ s(b_1)$ which is:

$1.197 \pm 2.306\ (.11)$
(.943, 1.451)

The slope of line in the population is between .943 and 1.451 with 95% confidence**.**

Analysis of Variance for the Line

The total variation in "y" is divided into two components (as before):

Data = Model + Error

= 1 + 2

1) **amount of variance in the data accounted for by the model (line)**
2) **amount of variance in the data NOT accounted for by the model (line) or error**

If the model is good (i.e., there are small residuals) then the model component should be large and the error small.
We can write this data partitioning as a **sums of squares (SS):**

SS Total = SS Model + SS Error
SST = SSM + SSE

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

The **degrees of freedom** associated with each SS is:

1 for SSM since we are testing H_o: B_1=0

(1 parameter)

n-2 for SSE since there are 2 parameters in model (B_o, B_1)

n-1 for SST since we do not test the y intercept **(i.e., H_o: $B_o = 0$ (one parameter))**

Each component of 1) and 2) has a mean square:

$$\mathbf{MS = \frac{sum\ of\ squares\ (SS)}{degrees\ of\ freedom\ (df)}}$$

For example:

$$\mathbf{MSE = SSE/\ DFE}$$

$$\mathbf{= \sum (y - \hat{y})^2 / (n - 2)}$$

$$\mathbf{= s^2 \quad (an\ estimate\ of\ \sigma^2)}$$

To test:

$$\mathbf{H_0 : B_1 = 0}$$

$$H_a : B_1 \neq 0$$

Use an F-statistic:

$$F = MSM/MSE$$

Which will have an F (1, n-2) distribution under the null hypothesis. If the model is not a good one, our estimate of error is about the same size as the effect of the model and we would expect an F ratio equal to 1.0. On the other hand, if the effect of the model is much larger than error, we would expect the F-ratio to be > 1.0.

We report our results in an **ANOVA** table since our estimates of model and error are based on the amount of variance accounted for by each of the two components:

Source	Sums of Squares	Degrees of Freedom	Mean Square	F
Model	$\sum (\hat{y} - \bar{y})^2$	**1**	SSM/DFM	MSM/MSE
Error	$\sum (y - \hat{y})^2$	**n – 2**	SSE/DFE	
Total	$\sum (y - \bar{y})^2$	**n – 1**		

In the case of the line, the square of the correlation coefficient (r) is:

$$\mathbf{R^2 = SSM/SST}$$

This gives the proportion of variation in y accounted for by x assuming a line model. What is 1- $\mathbf{R^2}$? (hint write 1 as **SST/SST.)** $\mathbf{R^2}$ was first introduced in Chapter 2.

Example (cont'd)

For the Gebotys and Roberts (1987) data construct an ANOVA table. Indicate how well the model accounts for the data.

Refer to table 9.1 for SST:

$$\mathbf{SST = \sum (y - \bar{y})^2 = 3994.1}$$

See table 9.2 for SSE and SSM:

$$\mathbf{SSE} = \sum e_i^2 = \mathbf{238.62}$$

$$\mathbf{SSM = SST - SSE}$$

$$= 3994.1 - 238.62$$

$$= 3755.48$$

since n = 10

$$DFM = 1$$
$$DFE = n-2 = 8$$
$$DFT = n-1 = 9$$

$$MSM = \frac{SSM}{DFM} = \frac{3755.48}{1} = 3755.48$$

$$MSE = \frac{SSE}{DFE} = \frac{238.62}{8} = 29.83$$

$$F = \frac{MSM}{MSE} = \frac{3755.48}{29.83} = 125.90$$

The results are summarized in an ANOVA table below:

Source	SS	DF	MS	F
Model	**3755.48**	**1**	**3755.48**	**125.90**
Error	**238.62**	**8**	**29.83**	
Total	**3994.1**	**9**		

Note that the F statistic is equal to 125.90. Since the OLS or p-value is <.0001 with 1, 8 degrees of freedom we have very strong evidence against:

$$\mathbf{H_0 = B_1 = 0}$$

and conclude there is a significant linear relationship between age and seriousness.

An estimate of σ^2, the spread about the line, can also be easily calculated from the table since s^2=MSE = 29.83.

$$\mathbf{R^2 = \frac{SSM}{SST} = \frac{3755.48}{3994.1} = .94}$$

In other words 94% of the variance in crime seriousness (y) is accounted for by the model (i.e., age). The model is adequate from an ANOVA as well as percent variance accounted for (R^2) point of view. An examination of the residuals is next.

In terms of proportion of variance in y accounted for by x we have:

- $R^2 = 0$ if the model accounts for 0 variance.
- $R^2 = 1$ if the model accounts for all the variance or perfect fit.

Researchers want models that account for a high and significant amount of variance. Click SPSS Example one to see the computer calculation of the above statistics.

SPSS EXAMPLE ONE

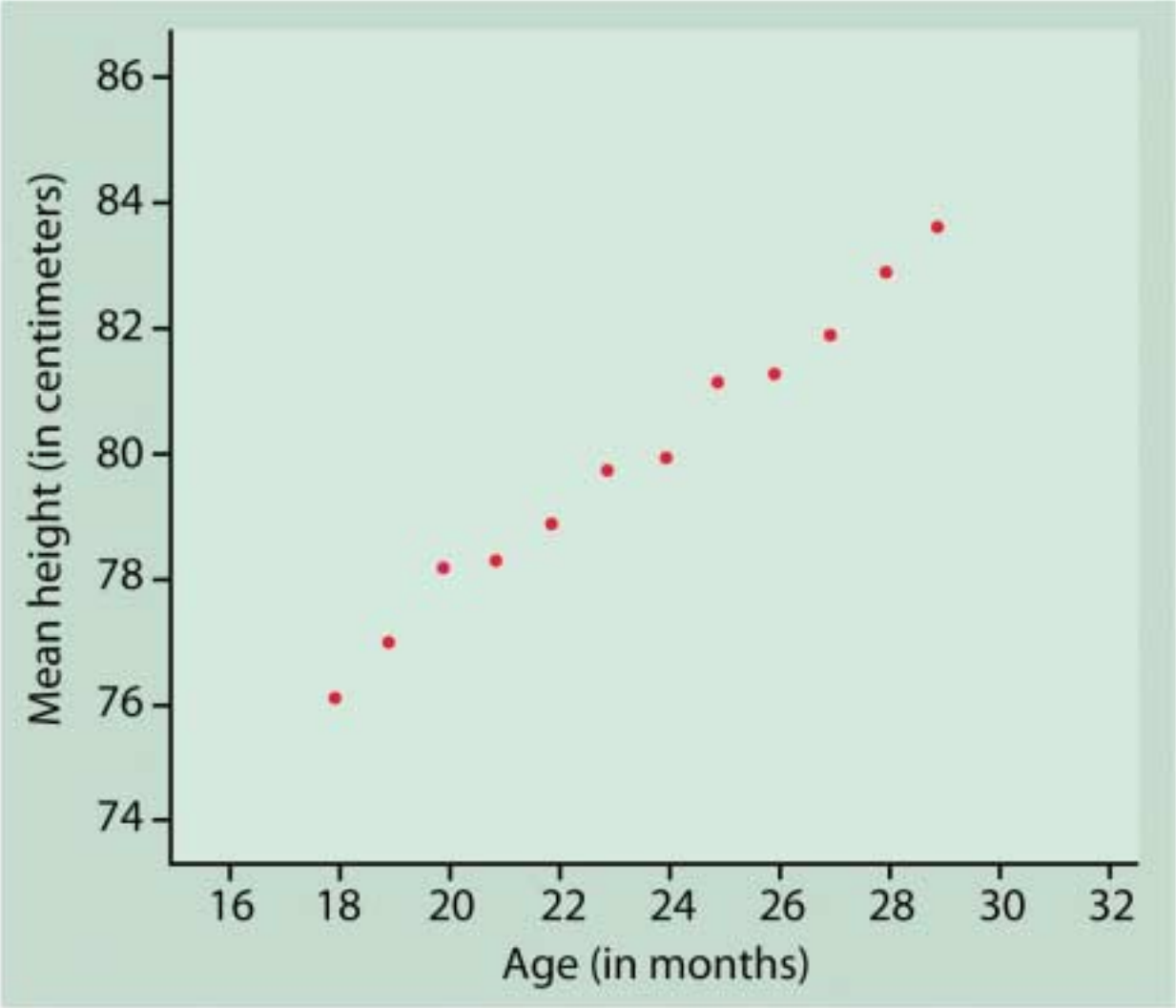
The process of fitting a line is usually accomplished with the help of a computer. In

SPSS Example One, the data is analyzed using SPSS. See the introduction to SPSS manual for another example using the data below that describes the relationship between children's age and height.

TABLE 2.7 Mean height of Kalama children

Age x in months	Height y in centimeters
18	76.1
19	77.0
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

A graph of the data follows on the next page.

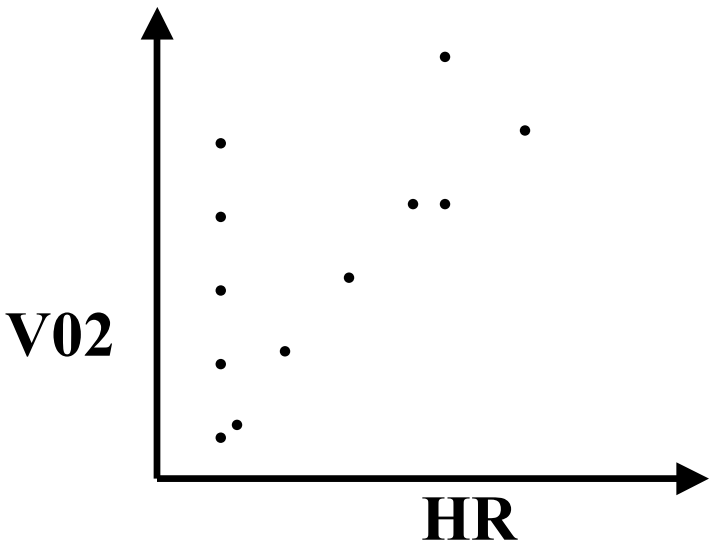


Example

A researcher is interested in seeing how oxygen uptake and heart rate are related.

She suspects there is a linear relationship between the two variables. The SPSS commands in Chapter 2 show how to read and

plot the data. The plot below shows a clear linear relationship.



The following SPSS example shows how to answer questions 1, 2, 3 below:

1) Fit a line to the data

2) Find 95% prediction intervals for heart rates 95 and 110

3) Construct residual plots to check assumptions.

SPSS EXAMPLE TWO

Click SPSS example two above to view a program that will answer the questions in the example.

Summary

The **model** for simple linear regression has been introduced. There are two unknowns in the model or population **parameters,** the slope and the intercept. From the data we calculate **least squares regression** estimates of the parameters.

We use the sample estimates to make inferences about the population parameters. **Tests of hypothesis** and **confidence intervals** are introduced using the t distribution.

The **Analysis of Variance** (ANOVA) table is introduced to show how the data is split into two pieces, model and error. If the model is not a

reasonable one then the **F ratio** of model over error components will equal 1.0.

Remember: Tips for Success

1) Read the text

2) Read the notes.

3) Try the assignment.

4) If needed, try the exercise questions.

5) Try the simulations and view the videos if you need more help with a concept.

6) Try the self tests for practice on each chapter of the text at: www.whfreeman.com/ips

7) Steady Work = SUCCESS