

Inference for Slope

Well, now that we can describe the linear relationship between two quantitative variables, it's time to conduct inference.

The Big Idea

There are several parameters in linear regression. The big idea is that there is a real linear relationship, coupled with random variation, that produces the data.

$y = \alpha + \beta x + \varepsilon$, where α is the y -intercept of the true linear relationship, β is the true slope of the linear relationship, and ε represents random variation. This can also be written $\mu_{y_i} = \alpha + \beta x_i$, which indicates that for an particular value of x (x_i), the linear relationship produces the mean response (for that value of x).

There are two parameters of interest—the intercept, and the slope. Of these, the slope is more interesting. The slope is about change, and much of mathematics (Calculus in particular) is concerned with change.

Test of Significance for the Slope

The Requirements

[1] There must be evidence of a linear relationship. We check this with our initial scatterplot, r and r^2 , checking the fit of the line, and the residual plot.

[2] observations must be independent—for us, the best evidence of this will be a random sample.

[3] the variation about the line (the variation in the set of responses for each particular value of x) must be constant. We check this with the residual plot.

[4] the response variable (for any value of x) must have a normal distribution centered on the line. We check this with a histogram or normal probability plot.

The Formula

The formula is the same as most test statistics that we have used— $\frac{\text{statistic} - \text{parameter}}{\text{measure of variation}}$.

Specifically, $t = \frac{b - \beta}{SE_b}$. Our null hypothesis must be for no relationship; thus, it will always be β

$= 0$. In light of this, we often write this statistic this way: $t = \frac{b}{SE_b}$.

Example

[1.] Here are some data on body fat and age. The data show the age of the subject (18 randomly selected subjects) and the percentage of body fat for that subject.

Table 1 - Age and Fat

Age	%Fat
-----	------

23	9.5
23	27.9
27	7.8
27	17.8
39	31.4
41	25.9
45	27.4
49	25.2
50	31.1
53	34.7
53	42.0
54	29.1
56	32.5
57	30.3
58	33.0
58	33.8
60	41.1
61	34.5

Let's see if there is any evidence of a useful linear relationship between these variables.

$H_0: \beta = 0$ (no useful relationship)

$H_a: \beta \neq 0$ (a useful relationship)

This calls for a t test for slope. There are many requirements that must be met in order to conduct this test. We've been told that these data represent a random sample. Let's check for linearity in the relationship.

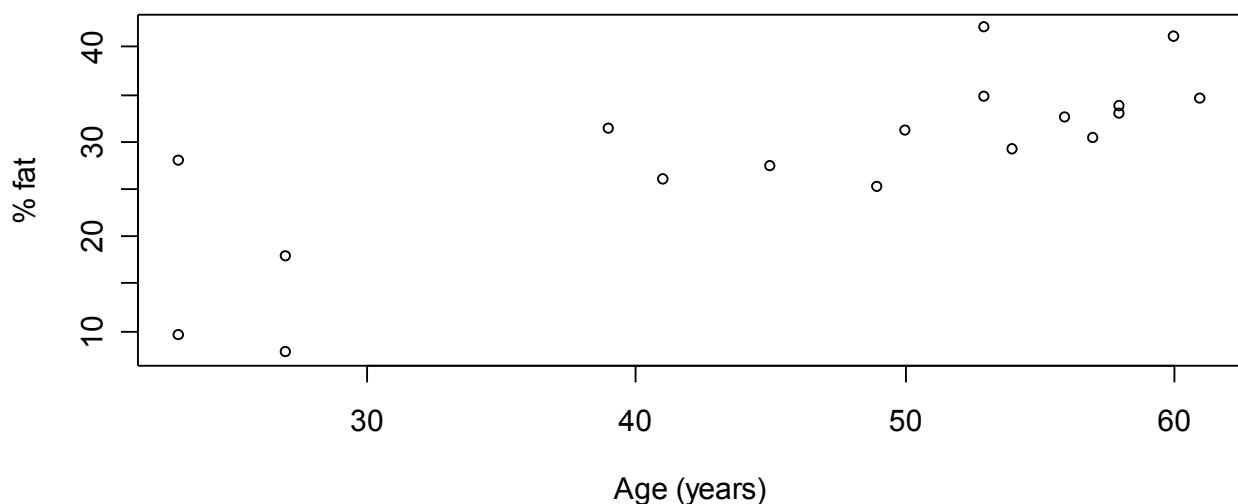


Figure 1 - Scatterplot for Example 1

Looks pretty linear, with strong positive association. $r = 0.7921$, which confirms our observations of a strong linear relationship; and $r^2 = 0.6274$, which means that 62.74% of the variation in percentage of body fat can be explained by the least squares regression of body fat on age. That's not bad...let's look at the fit.

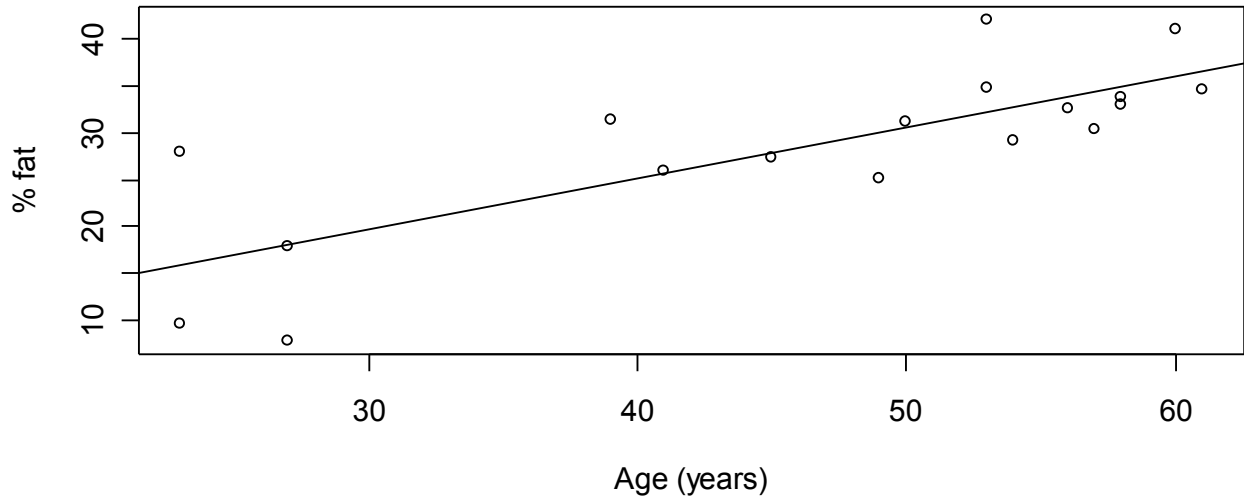


Figure 2 - Least Square Line for Example 1

Looks pretty good; the gap from 30 to 40 shouldn't be a problem...on to the residuals.

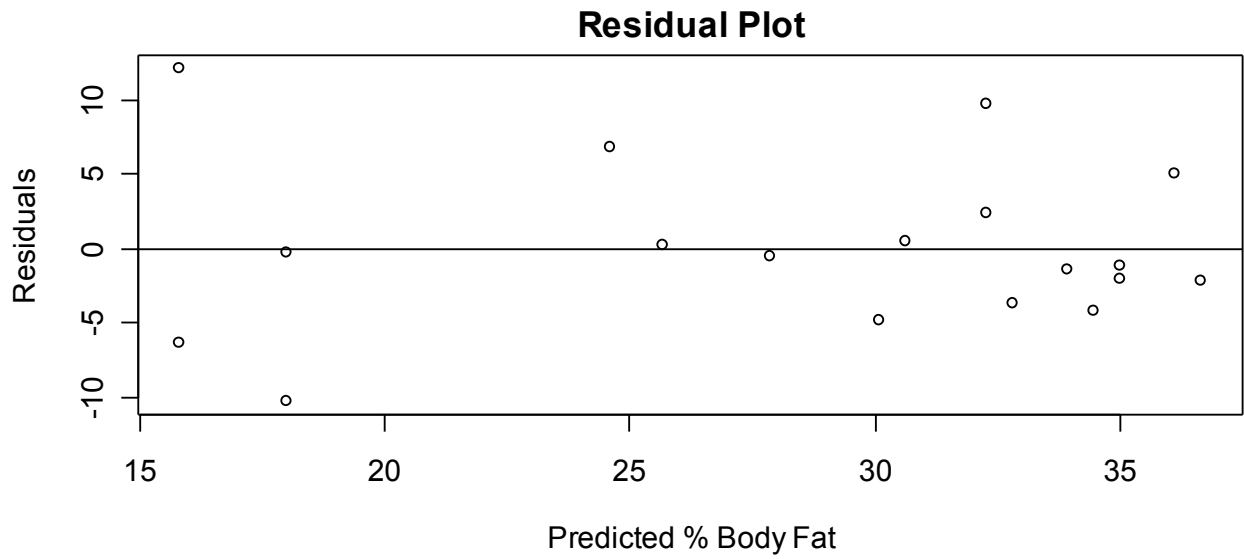


Figure 3 - Residual Plot for Example 1

Looks pretty good. How about normality?

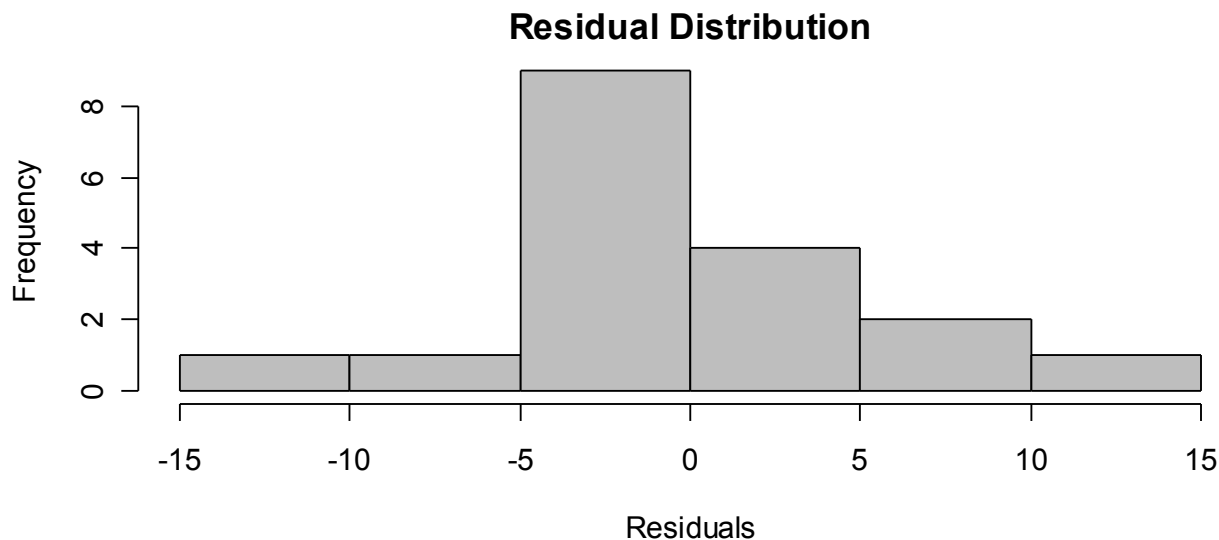


Figure 4 - Histogram of Residuals for Example 1

That'll do. We should be able to continue with the test.

I'll choose a 5% level of significance.

$t = 5.1910$. $2P(b > 0.5480) = 2P(t > 5.1910) = 0$.

If there is no significant slope, then I can expect to get a sample slope of 0.5480 or greater (or -0.5480 or lower) in almost no samples. This happens too rarely to attribute to chance at the 5% level; it is significant, and I reject the null hypothesis.

It appears that there is a useful linear relationship between age and % body fat.

Confidence Interval for the Slope

The Requirements

The same! Hoo-ray!

The Formula

The same basic formula still applies—statistic \pm (critical value)(measure of variation). In particular, $b \pm t^* SE_b$, where b is the slope of the least squares regression line, t^* is the upper $\frac{1-C}{2}$ critical value from the $t(n-2)$ distribution (which means $df = n - 2$), and SE_b is the **standard error of the slope**.

$$SE_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}, \text{ and } s = \sqrt{\frac{1}{n-2} \sum (y_i - \hat{y}_i)^2} . s \text{ is the } \mathbf{standard\ deviation\ about\ the}$$

least squares regression line. You'll notice that it looks a lot like standard deviation for the response variable—there are two differences; the denominator, and the fact that we're subtracting the predicted response, not the overall mean response.

Example

[2.] Here are some data from a random sample of baseball teams. The data show the team's batting average and the total number of runs scored for the season.

Table 2 - Batting Average and Runs Scored

Avg.	Runs
0.294	968
0.278	938
0.278	925
0.270	887
0.274	825
0.271	810
0.263	807
0.257	798
0.267	793
0.265	792
0.256	764
0.254	752
0.246	740
0.266	738
0.262	731
0.251	708

Let's estimate the slope of the true linear relationship between these variables with 99% confidence.

Well, before we begin plugging in numbers, we should check the requirements.
First up: a scatterplot of the data.

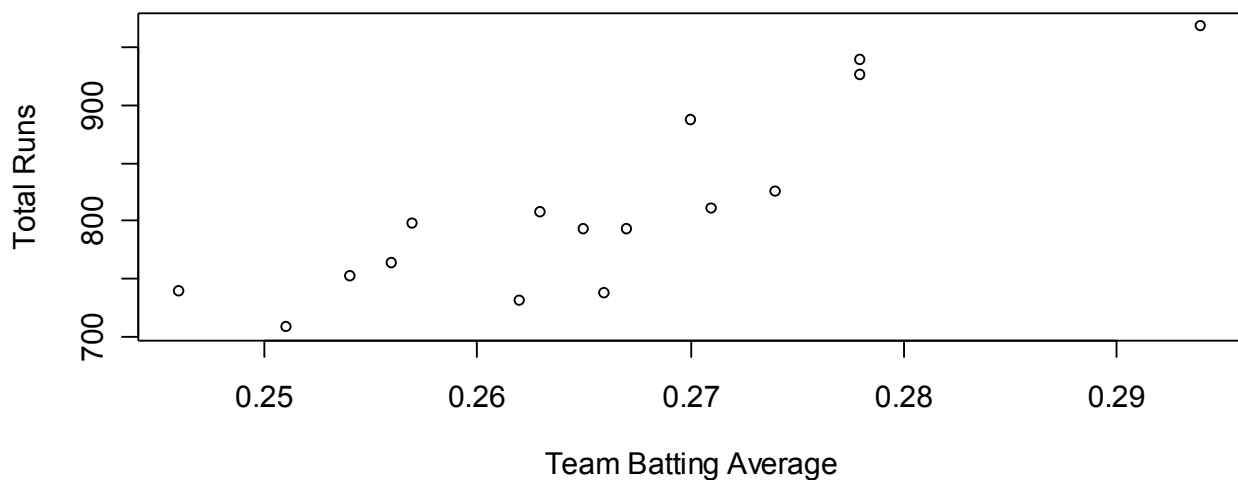


Figure 5 - Scatterplot for Example 2

Looks pretty linear, with strong positive association. There seems to be a gap between 0.28 and 0.29.

$r = 0.8655$, which supports our earlier observations. $r^2 = 0.7491$, which indicates that 74.91% of the variation in Total Runs can be explained by the least squares regression of Total Runs on Team Average. Let's check the fit of the regression line.

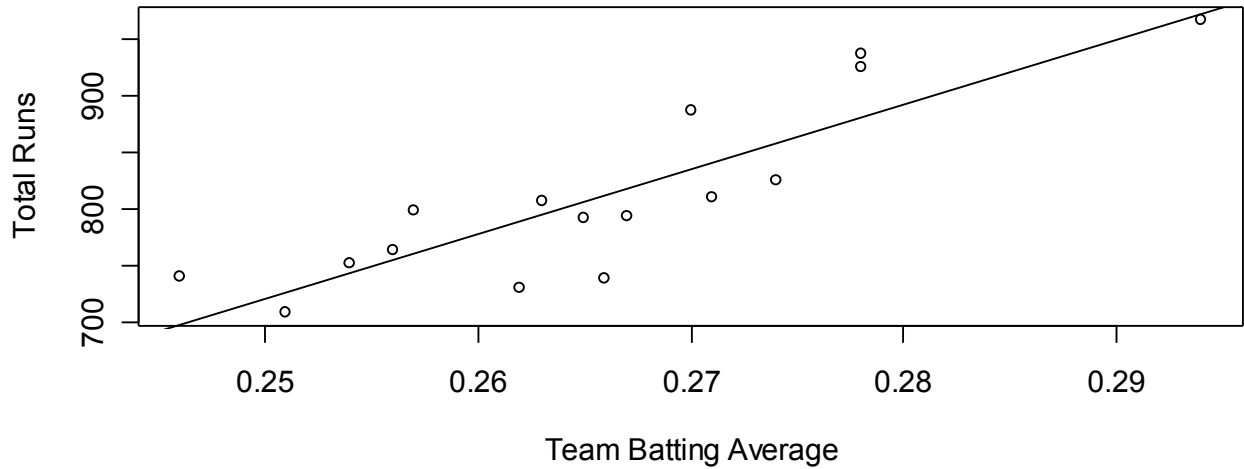


Figure 6 - Least Squares Line for Example 2

The fit seems acceptable. Let's check the residuals.

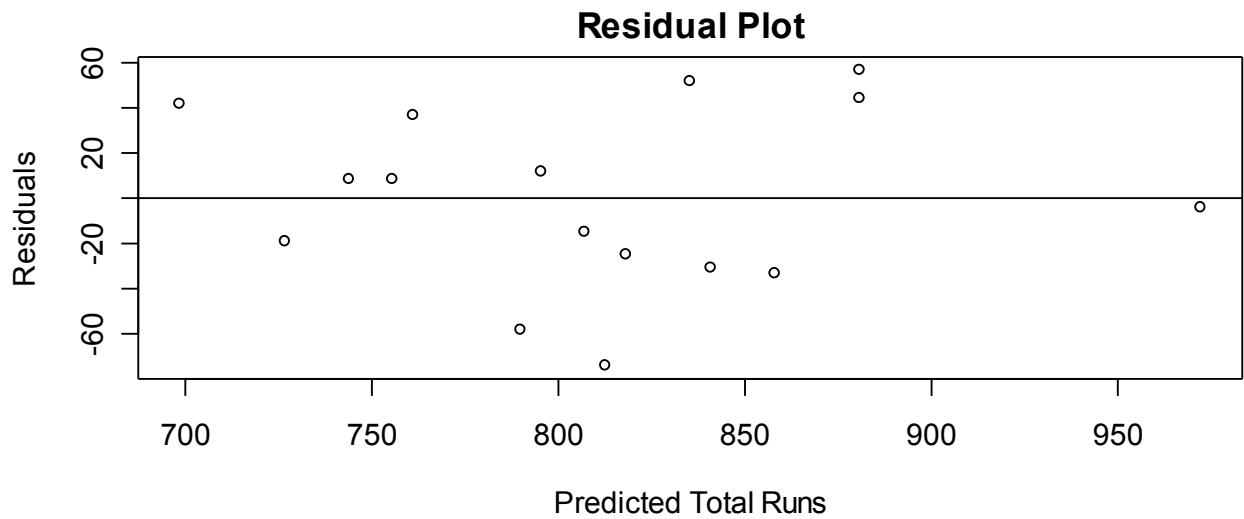


Figure 7 - Residual Plot for Example 2

The residual plot shows no obvious pattern—now to check normality.

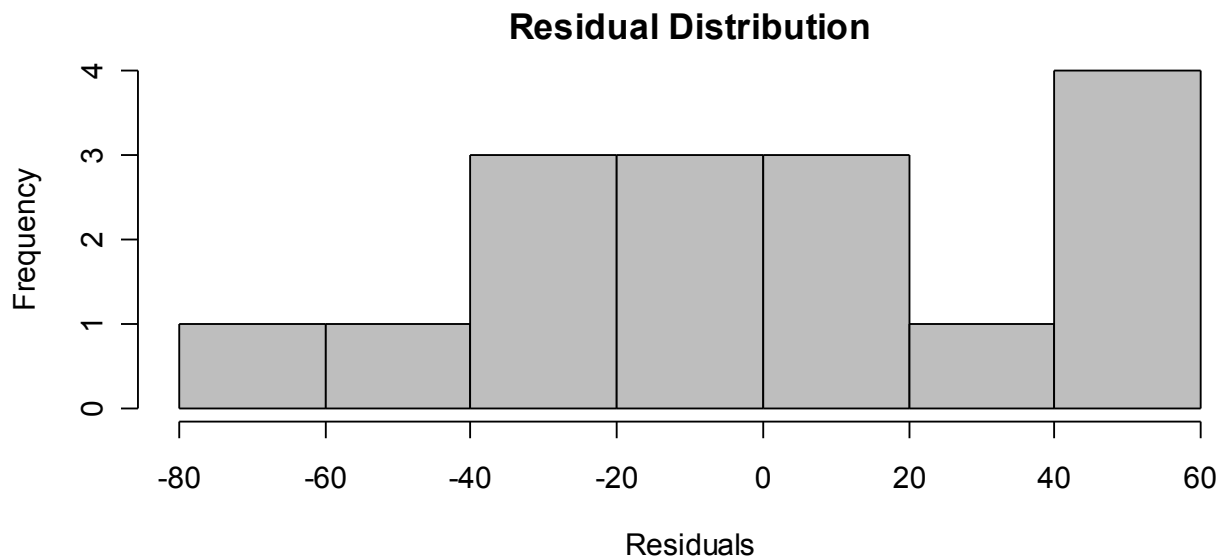


Figure 8 - Histogram of Residuals for Example 2

Hmmm—I'll look at the Normal Probability Plot, too.

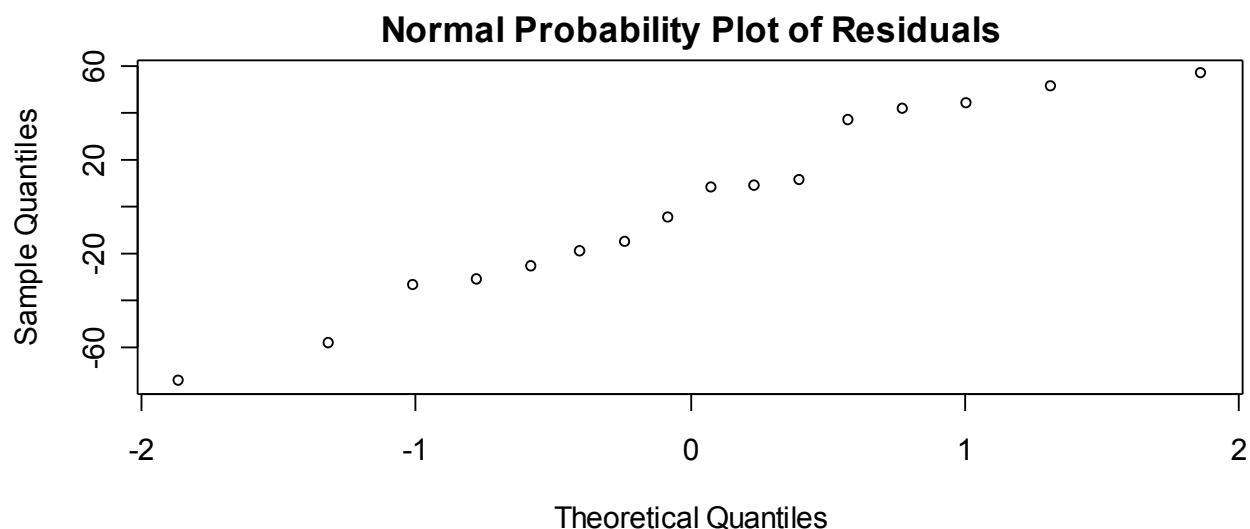


Figure 9 - Normal Probability Plot for Example 2

Well, that's not too bad.

I think that we can safely continue.

99% confidence and 14 degrees of freedom gives a t^* of 2.9768. The sample slope is 5709.2, and the standard error of the slope...uh-oh. How are we going to calculate that? The calculator just gives s —it sure would be a pain to have to use that formula to find SE_b .

Well, let's use a trick. There is one other place where we've used SE_b —in the test statistic!

Since $t = \frac{b}{SE_b}$, and since we can get the calculator to give us t and b (from the test), we can do

this: $SE_b = \frac{b}{t}$.

THIS IS A TRICK! NEVER write this formula where someone (e.g., an AP grader) might see it! If you were really doing this out in the world, you'd be using software, and you'd have everything you needed.

So—we get $SE_b = 883.1$. Plug in!

$$5809.2 \pm 2.9768 \cdot 883.1 = (3080.393, 8338.093).$$

I am 99% confident that the true slope of this relationship is between 3080.393 and 8338.093.

(since this slope represents the change in the Total Runs for a change of 1 in the Team Average, and since the Team Average cannot be greater than 1, this slope actually represents the maximum total runs in a year!)

Reading Computer Output

Often, you will be required to read standard computer output in order to obtain values for these items. Fortunately, almost all computer output looks alike. Here are some data, followed by several examples of computer output. Since you know how to get the calculator to give you what you need to know, you should be able to determine where those items are located in these examples...

The data relate the mass of a plant (g) with the quantity of volatile compounds (hundreds of nanograms) emitted by each plant.

Table 3 - Mass and Volatiles

mass	volatiles
57	8.0
85	22.0
57	10.5
65	22.5
52	12.0
67	11.5
62	7.5
80	13.0
77	16.5
53	21.0
68	12.0

Here is the output from a freeware program called **R**:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.5237	10.2995	0.342	0.74
mass	0.1628	0.1547	1.053	0.32

Residual standard error: 5.418 on 9 degrees of freedom
 Multiple R-Squared: 0.1096, Adjusted R-squared: 0.01067
 F-statistic: 1.108 on 1 and 9 DF, p-value: 0.32

Here is the output from **Microsoft Excel**:

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.331066792
R Square	0.109605221
Adjusted R Square	0.010672467
Standard Error	5.417706998
Observations	11

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	32.51787616	32.51787616	1.107875977	0.319982024
Residual	9	264.163942	29.35154911		
Total	10	296.6818182			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	3.523687722	10.29948909	0.342122575	0.740112225
mass	0.162848458	0.154717015	1.052556876	0.319982024

And now, output from **Data Desk**:

Dependent variable is: **volatiles**
No Selector

R squared = 11.0% R squared (adjusted) = 1.1%
s = 5.418 with 11 - 2 = 9 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	32.5179	1	32.5179	1.11
Residual	264.164	9	29.3515	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	3.52369	10.3	0.342	0.7401
mass	0.162848	0.1547	1.05	0.3200

And finally, output from **Statcrunch.com** (which looks a lot like Minitab):

Simple linear regression results:

Dependent Variable: volatiles

Independent Variable: mass

Sample size: 11

Correlation coefficient: 0.3311

Estimate of sigma: 5.417707

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value	
Intercept	3.5236878	10.299489	9	0.34212258	0.7401	You should
mass	0.16284846	0.15471701	9	1.0525569	0.32	see some

similarities there...