Simple Linear Regression

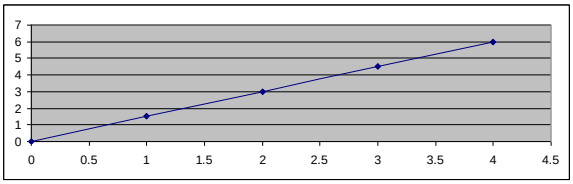
11.1 Creating the Least Squares Equation

Probabilistic Models

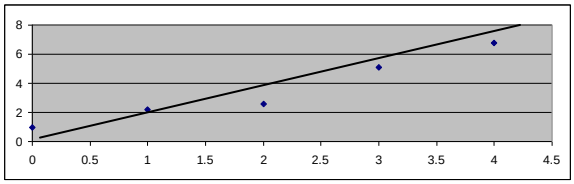
In this section, we will try to model the relationship between two variables. In algebra, you worked with many models that were ***deterministic*** in nature. For example, the model: $y = 1.06x$ is a deterministic model that will give the after tax price for an item purchased in Florida. X here represents the pre-tax price of an item. Y is the final price of the item post-tax. This model is deterministic because given a pre-tax price we can exactly (there is no error in this prediction) determine the value of the item after tax. Recall that the y variable is called the dependent variable because it depends upon the independent variable x.

Deterministic models are great when we can get them, but many times we do not know all of the factors affecting the dependent variable. In those cases, we will not be able to predict y without error. This means we will need to create a ***probabilistic*** model:

$$y = \text{deterministic model} + \textit{Random error}$$

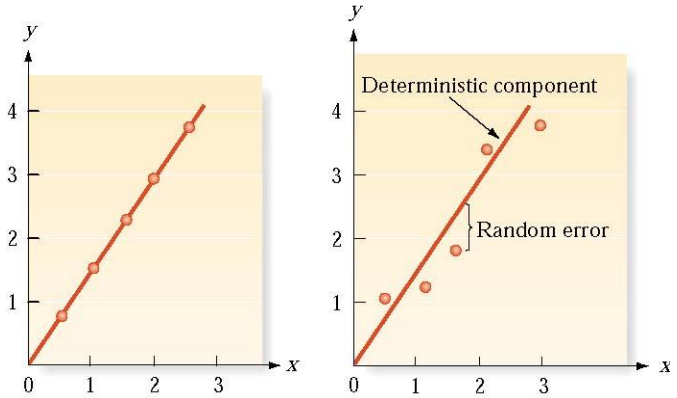


Deterministic model



Probabilistic Model

In the probabilistic model graph, the difference between the height of the line and the height of our individual points is due to the random error.



a. Deterministic relationship: $y = 1.5x$

b. Probabilistic relationship: $y = 1.5x + \text{Random error}$

In this chapter, we look at the simplest form of a probabilistic model:

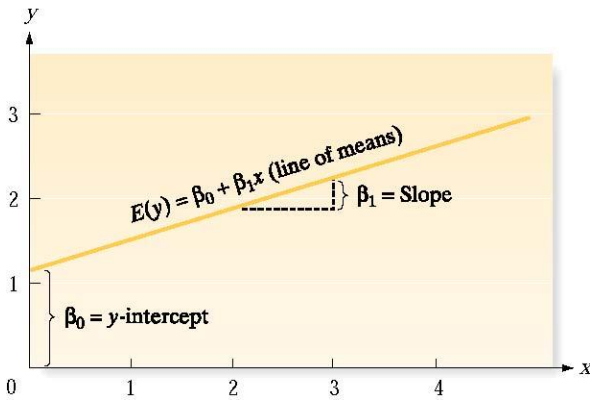
A First-Order (Straight-Line) Probabilistic Model

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

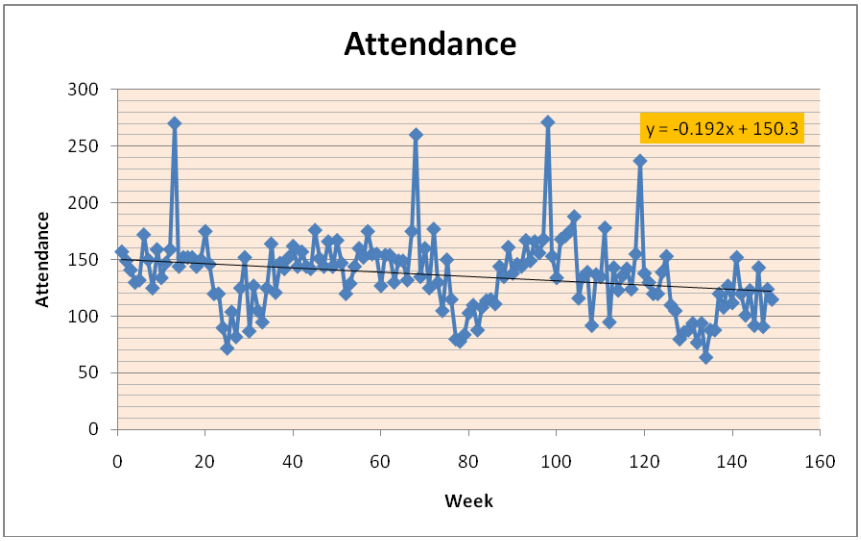
where: y = Dependent variable & x = independent variable

$E(y) = \beta_0 + \beta_1 x$ = Deterministic component,

β_1 = slope, β_0 = y-intercept, and ε = random error component.



Using data to come up with estimates of the parameters β_0 and β_1 in order to form an equation is called **regression analysis**. The goal of **regression analysis** is to find the straight line that comes closest to all of the points in a scatter plot simultaneously.



Fitting the Model: The Least Squares Approach

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

To fit the straight line model we need to find a way to estimate the unknown parameters:

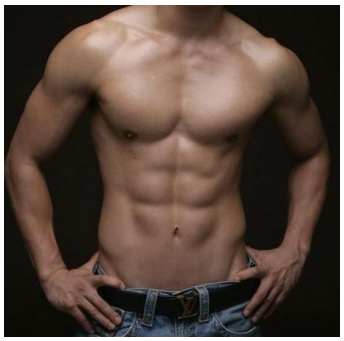
β_1 & β_0.

Consider this simple **example:**

In earlier research it was found that the body-mass index (BMI, weight scaled for height) was the main determinant of female physical attractiveness as judged by men. This raised the question: can women's attraction to men be so easily explained? To explore this hypothesis, researchers replicated the experiments on female attractiveness, but substituted male bodies and female raters.

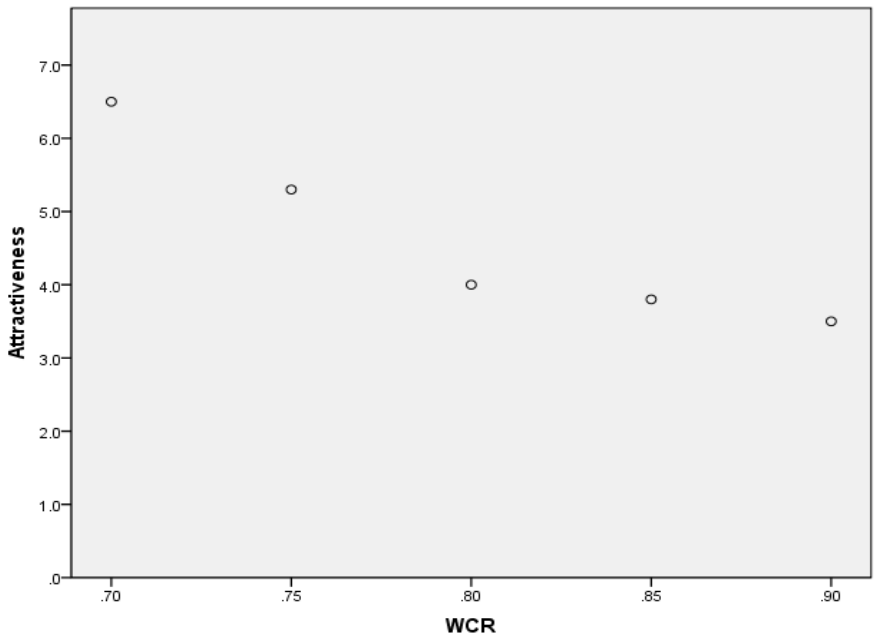
The results show that a woman's ratings of male attractiveness can be explained by simple physical characteristics, in particular the waist-to-chest ratio. Women prefer men who have a narrow waist and a broad chest and shoulders. Below are some of the waist-to-chest ratios and the attractiveness scores issued by the female judges:

WCR and Attractiveness		
Subject	WCR	Attractiveness Rating
1	0.70	6.5
2	0.75	5.3
3	0.80	4.0
4	0.85	3.8
5	0.90	3.5

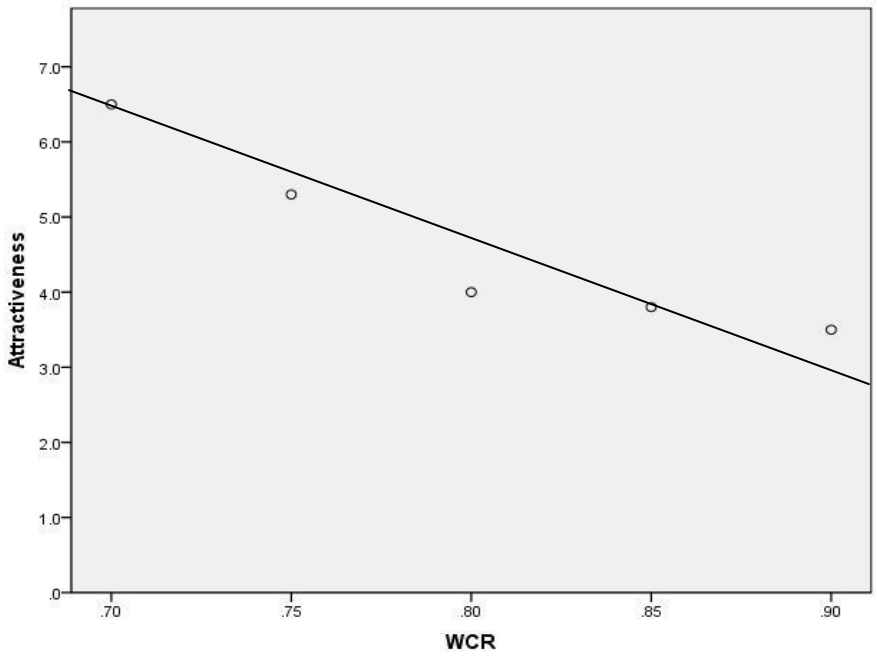


If we assume an adequate model for this situation is the first-order linear model $E(y) = \beta_0 + \beta_1 x$, we can try to use the sample data to estimate the missing parameters β_1 & β_0.

Consider the scatter plot of the data below:



We could try to fit some arbitrary line to the points above. For Example:



It looks like this line passes through two points: (0.70, 6.5) and another point of (0.85, 3.8) which gives a slope of $\frac{(y_2 - y_1)}{(x_2 - x_1)} = \frac{(6.5 - 3.8)}{(0.70 - 0.85)} \approx -18$. Using the point-slope formula from Algebra, we get the linear equation: $\tilde{y} = -18x + 19.1$. This model was obtained visually (i.e.-we guessed). We could have made several other guesses at the appropriate equation, so we should assess our guess.

Let's then compare the observed and predicted values for the visual model we found. In the table below, the X and Y are the actual values represented by the dots in our graph above. The $\tilde{y}$ (y-tilda) is the value that results when we plug the x value from the left most column of the table into our model. The first of the last two columns gives us the difference between the actual y-value from the point and the predicted value from our line. We call that difference the ***error*** of our prediction. For example, our line model says that when x is 0.75 we should have y at 5.6, but in reality the y value at x = 0.75 was 5.3. This means our error is -0.3. The last column squares these differences (or errors).

X	Y	$\tilde{y} = -18x + 19.1$	$(y - \tilde{y})$	$(y - \tilde{y})^2$
0.70	6.5	6.5	0	0
0.75	5.3	5.6	-0.3	0.09
0.80	4.0	4.7	-0.7	0.49
0.85	3.8	3.8	0	0
0.90	3.5	2.9	0.6	0.36
Sum of Errors:			**-0.4**	**0.94**

One way to determine quantitatively how well a straight line fits a set of points is to note the extent to which the data points deviate from the line. The quantity $\sum(y - \tilde{y})$ in the table above gives us the total deviation between our observed values and our predicted values. The $\sum(y - \tilde{y})$ should equal zero (the sum of errors should equal zero). In our visual model, that is not the case, which is something that would eliminate it as a candidate model for this set of data points. We will want all of our prediction lines to have the property that **the sum of errors equals zero**. This will ensure on average our error of prediction is zero. The quantity $\sum(y - \tilde{y})^2$ is called the ***sum of squares of the errors (SSE)*** gives another measure of deviation which gives a greater emphasis to larger deviations from the line.

Remember we visually selected the model (line) above, so it is no wonder that the sum of errors was not zero. However, there are usually multiple models possible that have the property that $\sum(y-\tilde{y})$=0. Since, we can find more than one model with the property $\sum(y-\tilde{y})$ = 0, we need an additional criteria to choose the best fitting line. It turns out that it can be shown that there is one (and only one) line for which the SSE is a *minimum*. This line is called the **least squares line**.

The **least squares line** has the following properties:

1. The sum of errors (SE) equals zero.
2. The sum of squared errors (SSE) is smaller than that for any other straight line model.

The following formulas will give the Least Squares Estimates for the population β_1 & β_0. (We will use a "hat" symbol to denote the estimates, that is to say $\hat{\beta}_1$ estimates β_1 and $\hat{\beta}_0$ estimates β_0).

Formulas for the Least Squares Estimates

Slope: $\hat{\beta}_1 = \dfrac{SS_{xy}}{SS_{xx}}$

y-intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

where $SS_{xy} = \sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \sum x_i y_i - \dfrac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$

and $SS_{xx} = \sum\left(x_i - \bar{x}\right)^2 = \sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}$

Example 160**:** Now let us find the least squares prediction line for our WCR vs. Attractiveness Rating example above:

Let $E(y) = \beta_0 + \beta_1 x$ be our straight line model where **y = attractiveness rating** on an 8-point scale **and x = waist-to-chest ratio**.

Preliminary computations for the male attractiveness problem				
	x_i	y_i	x_i^2	$x_i y_i$
	0.70	6.5	0.49	4.55
	0.75	5.3	0.5625	3.975
	0.80	4.0	0.64	3.2
	0.85	3.8	0.7225	3.23
	0.90	3.5	0.81	3.15
Totals	$\sum x_i = 4$	$\sum y_i = 23.1$	$\sum x_i^2 = 3.225$	$\sum x_i y_i = 18.105$

Using the numbers from above we can get:

$$SS_{xy} = \sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 18.105 - (4)(23.1)/5 = -0.375$$

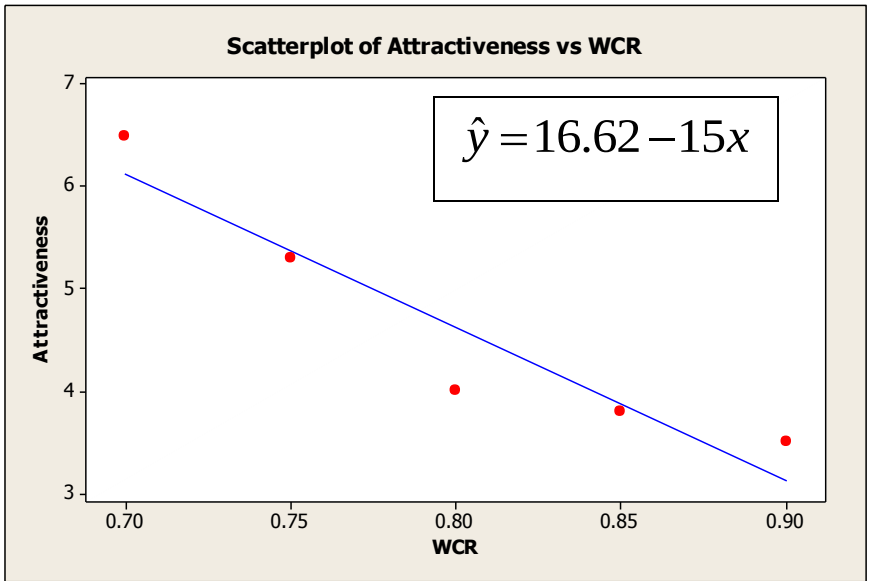
$$SS_{xx} = \sum\left(x_i - \bar{x}\right)^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 3.225 - \frac{(4)^2}{5} = 3.225 - 3.2 = 0.025$$

then $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = -0.375/0.025 = -15$

and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\left(\sum y_i\right)}{n} - \hat{\beta}_1 \frac{\left(\sum x_i\right)}{n} = \frac{23.1}{5} + 15\left(\frac{4}{5}\right) = 16.62$

The **least squares line** is then given by:

$$\hat{y} = 16.62 - 15x$$



Let us find the SSE for this line to determine if it beats our visual model:

X	Y	$\hat{y} = 16.62 - 15x$	$(y - \hat{y})$	$(y - \hat{y})^2$
0.70	6.5	6.12	0.38	0.1444
0.75	5.3	5.37	-0.07	0.0049
0.80	4.0	4.62	-0.62	0.3844
0.85	3.8	3.87	-0.07	0.0049
0.90	3.5	3.12	0.38	0.1444
Sum of Errors:			**0**	**0.683**

We can now confirm our Least Squares Model is better fitting than out visual model because our LSM has a Sum of Errors = 0, and even if both models had that trait, the least squares model has a lower SSE.

The work we performed to create the least squares model above is typically done using software. The output below is from the software package called Minitab.

Minitab Display:

```
Coefficients

Term          Coef  SE Coef  T-Value  P-Value
Constant     16.62     2.42     6.86    0.006
WCR         -15.00     3.02    -4.97    0.016
```

Finally, we can now use the above model for prediction.

Example 161: (tech) Using the model, what would the average attractiveness score be for a 45 inch chest and a 32 inch waist (which produces a WCR of 0.71)?

Example 162: (tech) Find the least squares prediction line for the following pre-owned Corvette data

Age	6	6	5	2	2	5	4	5	1	4
Price	27000	26000	27500	40500	36400	29500	33500	30800	40500	30500

Here is the same work performed by Minitab (the price values are expressed in ten thousands):

```
Coefficients

Term          Coef  SE Coef  T-Value  P-Value
Constant     4.386    0.128    34.21    0.000
Age        -0.2911   0.0296    -9.84    0.000
```

Example 163: (tech) What would the average price be for a Corvette that was three years old?

11.2 Finding S for the Random Error Terms

Model Assumptions

Our model in the above section had a deterministic component and a random error component. In this section, we will consider the assumptions that we make about that random error component (ε).

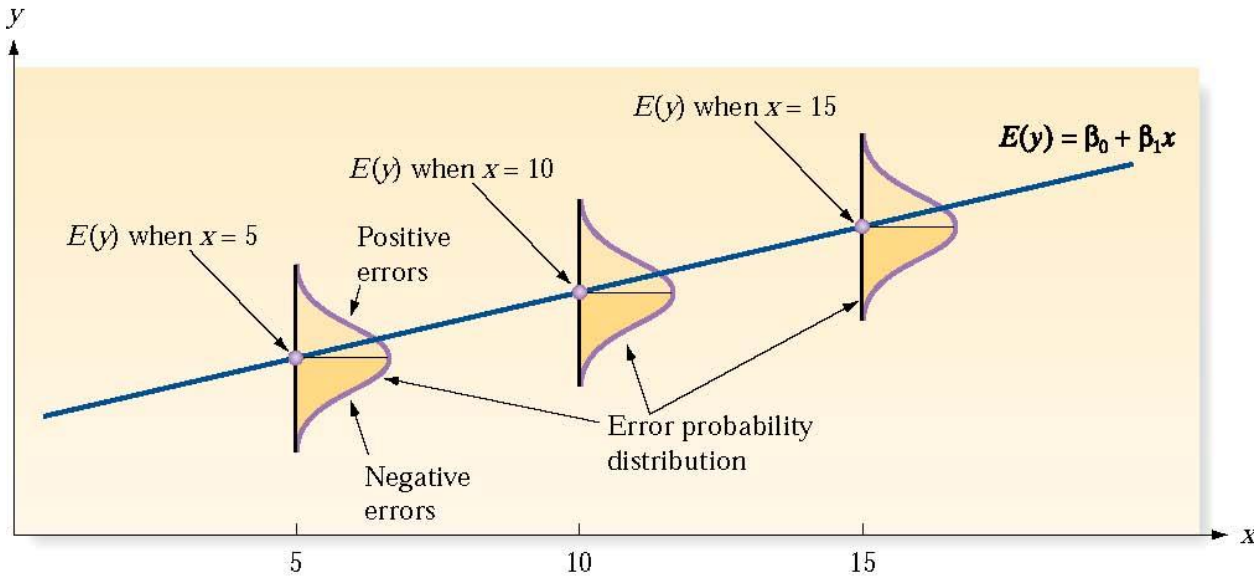
Assumption 1: The mean of the probability distribution for (ε) is 0. That is why $E(y) = \beta_0 + \beta_1 x$. Recall the original model was $y = \beta_0 + \beta_1 x + \varepsilon$.

Assumption 2: The variance for (ε) is a constant denoted by σ^2. No matter what x value we use in the model the distribution of the random error has the same variance.

Assumption 3: The probability distribution of (ε) is normal.

Assumption 4: The values of (ε) associated with any two observed values of y are independent.

Study the figure below: notice how the shape of the error distribution is normal and the variance is also constant.



An estimator of σ^2

$$S^2 = \frac{SSE}{Degrees\ of\ \ Freedom\ for\ Error} = \frac{SSE}{n-2}$$

*(We used up two degrees of freedom estimating our two parameters β_1 & β_0. That is why d.f. = n – 2.)

where:

$$SSE = \sum (y_i - \hat{y}_i)^2 = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

where:

$$SS_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} \text{ and recall } \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

Finally, if we want an estimate of σ, we will use $\sqrt{S^2} = \sqrt{\frac{SSE}{n-2}}$

Let's find $S^2 = \frac{SSE}{n-2}$ for the following example:

Example 164: Find $S^2 = \frac{SSE}{n-2}$ for the data below:

Number of Weeks on a Low Carb Diet X	Weight Change Y
1	-5
2	-9
3	-12
4	-17
5	-20

Steps to finding S^2 an estimate of σ^2:

1. Create and fill in the preliminary calculation table below:

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	⋮	⋮	⋮	⋮	⋮
Totals	$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum y_i^2$	$\sum x_i y_i$

2. Calculate $SS_{xy} = \sum x_i y_i - \dfrac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$
3. Calculate $SS_{xx} = \sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}$
4. $SS_{yy} = \sum y_i^2 - \dfrac{\left(\sum y_i\right)^2}{n}$
5. Find $\hat{\beta}_1 = \dfrac{SS_{xy}}{SS_{xx}}$
6. Calculate $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$
7. Finally, $S^2 = \dfrac{SSE}{n-2}$

As with our previous examples, it is entirely possible to find the standard deviation of the error term by using software. The following is the complete output provided by Minitab:

Regression Analysis: Weight Change versus Number of Weeks on the Diet

```
Analysis of Variance

Source       DF   Adj SS   Adj MS  F-Value  P-Value
Regression    1  144.400  144.400   541.50    0.000
No.of Weeks   1  144.400  144.400   541.50    0.000
Error         3    0.800    0.267
Total         4  145.200

Model Summary

       S    R-sq  R-sq(adj)
0.516398  99.45%     99.27%


Coefficients

Term           Coef  SE Coef  T-Value  P-Value
Constant     -1.200    0.542    -2.22    0.114
No. of Weeks -3.800    0.163   -23.27    0.000


Regression Equation: Weight Change = -1.200 - 3.800*(No.of Weeks)
```

The MSE from the ANOVA display is the S^2 described above!

Here, they have given us S.

Interpretation of s, the estimated standard deviation of (ε):

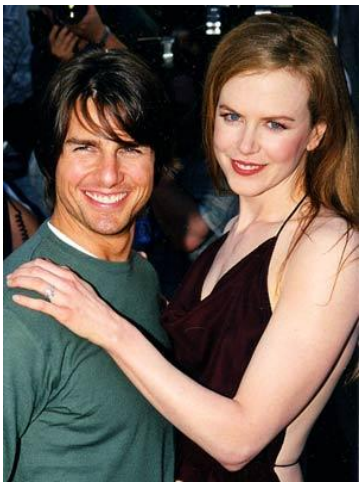
We expect approximately 95% of the observed y values to lie within 2s of their respective least squares predicted y – values, $\hat{y}$.

For example, if s = 0.103 we would expect 95% of all of our observed values to lie within 0.206 of the least squares line.

Incidentally, the difference (y - $\hat{y}$) between an observed sample y-value and the value of $\hat{y}$, which is the value of y that is predicted by using the regression equation, is called a residual for a sample of paired (x, y) data.

Example 165: (tech) Find the $S^2 = \frac{SSE}{n-2}$ for the following data and state the largest deviation we would expect between any of the actual data points and our least squares line.

Female Height (in Meters) X	Ideal Height of Mate Y
1.52	1.69
1.60	1.74
1.68	1.80
1.75	1.93
1.83	2.00



Accompanying Minitab output:

```
Model Summary

        S    R-sq  R-sq(adj)
0.0282013  96.46%     95.29%


Coefficients

Term       Coef  SE Coef  T-Value  P-Value
Constant  0.076    0.194     0.39    0.721
Female Ht 1.048    0.116     9.05    0.003


Regression Equation: Ideal Ht of Mate = 0.076 + 1.048 * (Female Ht)
```

11.3 Finding the Standard Error of the Slope Estimator

In order to perform a hypothesis test or form a confidence interval to make an inference about β_1 (the slope), we need to know the sampling distribution of our estimator $\hat{\beta}_1$.

Sampling Distribution of $\hat{\beta}_1$

If we make the four assumptions about ε (see section 11.2), the sampling distribution of the least squares estimator $\hat{\beta}_1$ of the slope will be **normal** with mean β_1 (the true slope) and standard deviation

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$$

We estimate $\sigma_{\hat{\beta}_1}$ by $s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$ and refer to this quantity as the **estimated standard error of the least squares slope** $\hat{\beta}_1$ (recall $S = \sqrt{\frac{SSE}{n-2}}$).

Example 165.5 (tech) Use the data from example 165 and the Minitab results below to find the standard error of the slope estimator ($S_{\hat{\beta}_1}$).

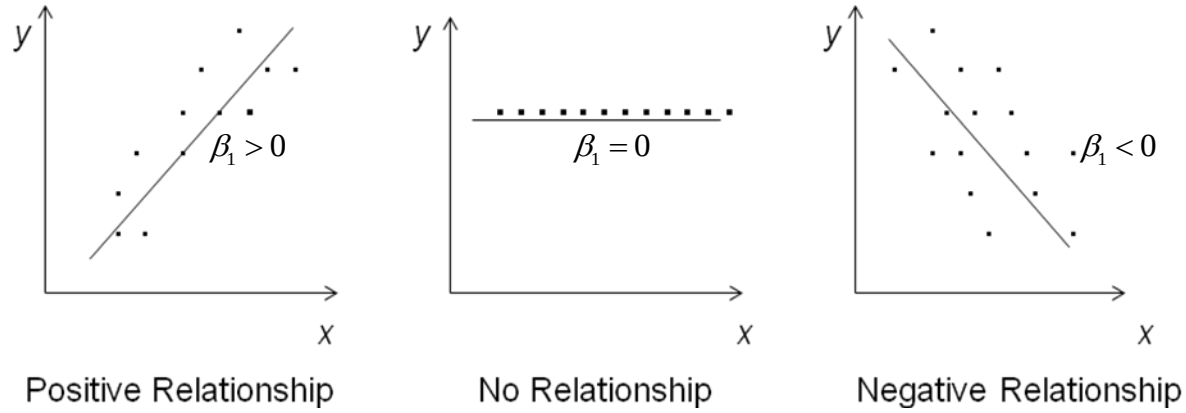
```
Analysis of Variance

Source       DF    Adj SS    Adj MS  F-Value  P-Value
Regression    1  0.065094  0.065094    81.85    0.003
Female Ht     1  0.065094  0.065094    81.85    0.003
Error         3  0.002386  0.000795
Total         4  0.067480


Model Summary

        S    R-sq  R-sq(adj)
0.0282013  96.46%     95.29%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value
Constant   0.076    0.194     0.39    0.721
Female Ht  1.048    0.116     9.05    0.003
```

Now that we know the sampling distribution of $\hat{\beta}_1$, we can perform our hypothesis test.

11.4 Hypothesis Tests about the Slope β1

Making Inferences about β_1 our slope

Recall β_1 is our slope for the linear model: $y = \beta_0 + \beta_1 x + \varepsilon$.



If the true value of the slope is equal to zero $y = \beta_0 + \beta_1 x + \varepsilon$ becomes $y = \beta_0 + 0 \bullet x + \varepsilon = \beta_0 + \varepsilon$, this means that x has no role in predicting y. If that is the case, our model is not useful. For this reason, we will want to test the claim that the slope is equal to zero. We would like to reject that claim because if we are unable to reject it we have a useless model.

A Test of Model Utility: Simple Linear Regression

One Tailed Test

$H_o : \beta_1 \geq 0$ or $(H_o : \beta_1 \leq 0)$

$H_a : \beta_1 < 0$ or $(H_a : \beta_1 > 0)$

Test Statistic: $t = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \dfrac{\hat{\beta}_1}{s / \sqrt{SSxx}}$

Rejection region: $t < -t_a$

Or ($t > t_a$ when $H_a : \beta_1 > 0$)

Two Tailed Test

$H_o : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

Rejection region: $|t| > t_{a/2}$

Where t_a and $t_{a/2}$ are based on (n-2) degrees of freedom

Example 166: At the 1% significance level, test the claim that there is a positive linear relationship between a mother's height and her daughter's height.

Mother's Height	Daughter's Height
63	58.6
67	64.7
64	65.3
60	61.0
65	65.4
59	60.9

Step 1: Form the claim symbolically

$$\beta_1 > 0$$

Step 2: Get your Hypotheses

$$H_0 : \beta_1 \le 0$$
$$H_A : \beta_1 > 0$$

Step 3: Find $s_{\hat{\beta}_1} = \dfrac{s}{\sqrt{SS_{xx}}}$ where $s = \sqrt{\dfrac{SSE}{n-2}}$, $\hat{\beta}_1 = \dfrac{SS_{xy}}{SS_{xx}}$, and $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$

$$s_{\hat{\beta}_1} = 0.3579$$

Step 4: Find the Test Stat $\quad t = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \dfrac{\hat{\beta}_1}{s/\sqrt{SSxx}} = 1.719$

Step 5: Find your critical t-value by looking up $\alpha = 1\%$ in one tail 4 df (n – 2) $t_{\alpha/2} = 3.747$

Step 6: Form your initial conclusion: Do not reject the null.

Step 7: State your final conclusion

At the 1% significance level, there is not sufficient evidence to support the claim that there is a positive relationship between a mother's height and her daughter's height.

Example 166 (Tech): Using a 1% significance level and the computer output below, test the claim that there is a positive linear relationship between a mother's height and her daughter's height.

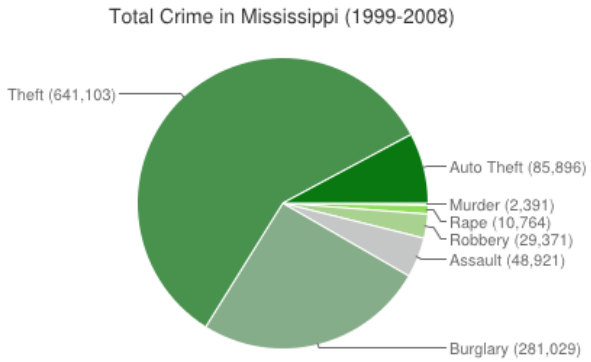
Mother's Height	Daughter's Height
63.2	58.6
67.1	64.7
64.5	65.3
60.5	61.0
65.1	65.4
59.3	60.9
64.0	60.0
62.0	59.0
66.5	64.5

Minitab output:

```
Analysis of Variance

Source       DF  Adj SS  Adj MS  F-Value  P-Value
Regression    1   26.39  26.388     5.11    0.058
Mother's Ht   1   26.39  26.388     5.11    0.058
Error         7   36.15   5.165
Total         8   62.54


Model Summary

      S    R-sq  R-sq(adj)
2.27263  42.19%     33.93%


Coefficients

Term         Coef  SE Coef  T-Value  P-Value
Constant     18.0     19.6     0.92    0.388
Mother's Ht 0.695    0.307     2.26    0.058

Regression Equation

Daughter's Ht = 18.0 + 0.695 * (Mother's Ht)
```

Example 167: At the 1% significance level, test the claim that there is a linear relationship between the number of casino employees (in thousands) working in Mississippi and the crime rate per thousand.

Number of casino employees	Crime rate
15	1.35
18	1.63
24	2.33
22	2.41
25	2.63
29	2.93



Example 167 tech: The data above was entered into Minitab, and the results of the data analysis are provided below. Use the Minitab output and a 1% significance level to test the claim that there is a linear relationship between the number of casino employees (in thousands) working in Mississippi and the crime rate per thousand.

```
Analysis of Variance

Source       DF   Adj SS   Adj MS  F-Value  P-Value
Regression    1  1.74493  1.74493    87.02    0.001
No Employees  1  1.74493  1.74493    87.02    0.001
Error         4  0.08020  0.02005
Total         5  1.82513


Model Summary

       S    R-sq  R-sq(adj)
0.141602  95.61%     94.51%


Coefficients

Term            Coef  SE Coef  T-Value  P-Value
Constant      -0.387    0.285    -1.36    0.246
No Employees  0.1173   0.0126     9.33    0.001


Regression Equation

Crime Rate = -0.387 + 0.1173 * (No Employees)
```

11.5 Confidence Interval for the Slope β1

It is also possible to form an interval estimator of β_1:

A 100(1-a)% Confidence Interval for the Sample Linear Regression Slope β_1

$$\hat{\beta}_1 \pm t_{a/2} s_{\hat{\beta}_1}$$

Where the estimated standard error of $\hat{\beta}_1$ is calculated by $s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$

And $t_{a/2}$ is based on (n-2) degrees of freedom.

Example 168: Using the data from above, we calculated $s_{\hat{\beta}_1} = 0.01257$ and $\hat{\beta}_1$ = 0.11729. Use the formula, $\hat{\beta}_1 \pm t_{a/2} s_{\hat{\beta}_1}$ to form a 99% confidence interval to estimate the true slope β_1 for the Mississippi Casino/Crime rate data above.

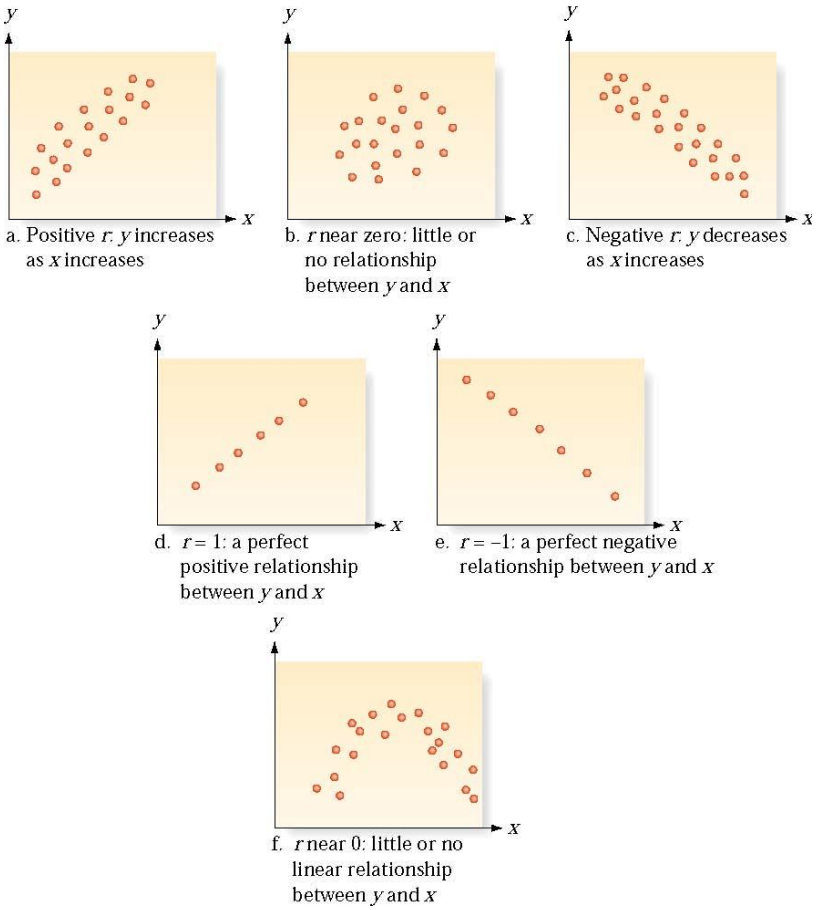
11.6 Finding r the Coefficient of Correlation

The Coefficient of Correlation

The coefficient of correlation, $r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$ is a measure of the strength of the linear relationship between two variables x and y.

Recall that the numerator for $\hat{\beta}_1$ is SS_{xy}, this is the same as the numerator for r. This means that when SS_{xy} = 0, both r and $\hat{\beta}_1$ will be equal to zero. When SS_{xy} =0, there is no linear relationship between x and y.

Values of ***r*** and their implications:



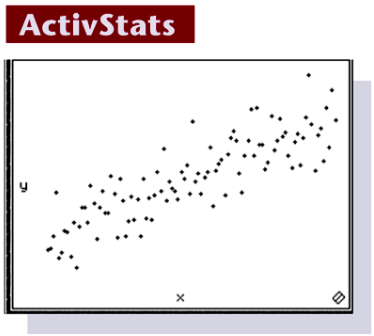
People sometimes misinterpret r. Please remember that if r = 0 it does not mean there is no relationship between x and y it just means there does not seem to be a **linear** relationship between them. Look at drawing f. above, it clearly has a relationship (perhaps quadratic), but it has no linear relationship. Also, if |r| is close to one, it does not mean that x causes y or that y causes x. It only means there is some linear relationship between the two variables, but the relationship could be due to some other unknown cause. For example, in the 1970's it might have been possible to show a positive correlation between number of hours spent flying and the incidence of lung cancer. This doesn't mean that flying causes lung cancer. In fact, in the 1970's people were allowed to smoke on planes. This meant frequent flyers were inundated with second hand smoke, which was more likely the cause of the higher rates of cancer.

Example 169: Calculate the correlation coefficient for the following data set which gives the waist measurement and body fat percentage for males who weigh 165 pounds. Do these quantities have a linear relationship? If so, is it positive or negative?

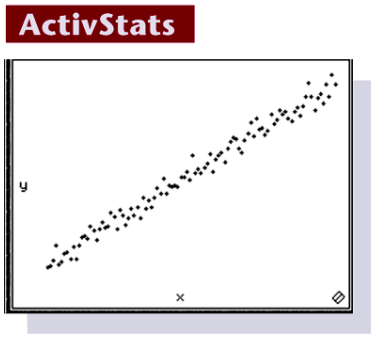
Male Waist (in inches)	Body fat %
30	8.5
32	12.0
34	18.5
36	25.0
38	27.0

Note: $SS_{xx} = 40, SS_{xy} = 100,$ and $SS_{yy} = 256.3$

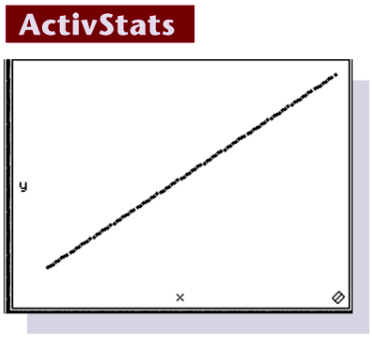
Here are some more examples of r values and scatter plots:



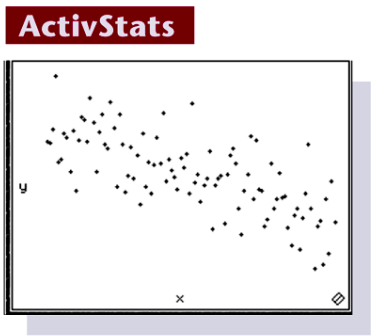
(a) Positive correlation: $r = 0.851$



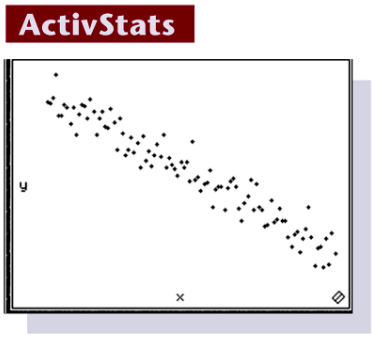
(b) Positive correlation: $r = 0.991$



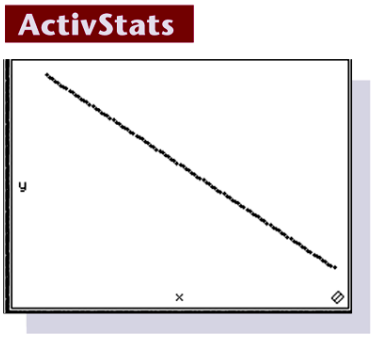
(c) Perfect positive correlation: $r = 1$



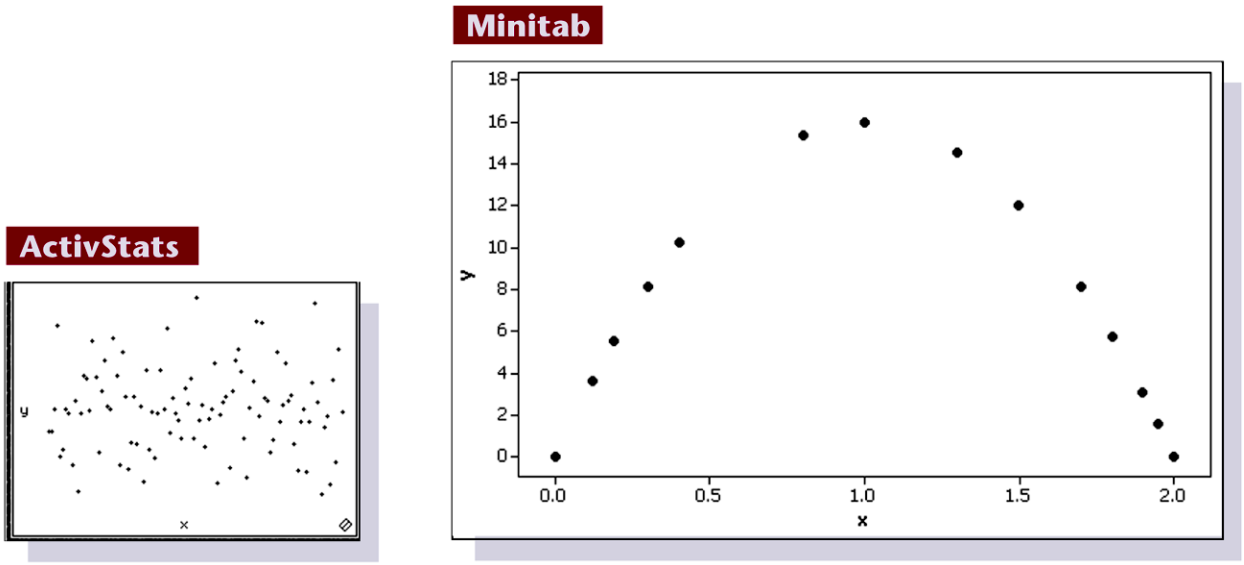
(d) Negative correlation: $r = -0.702$



(e) Negative correlation: $r = -0.965$

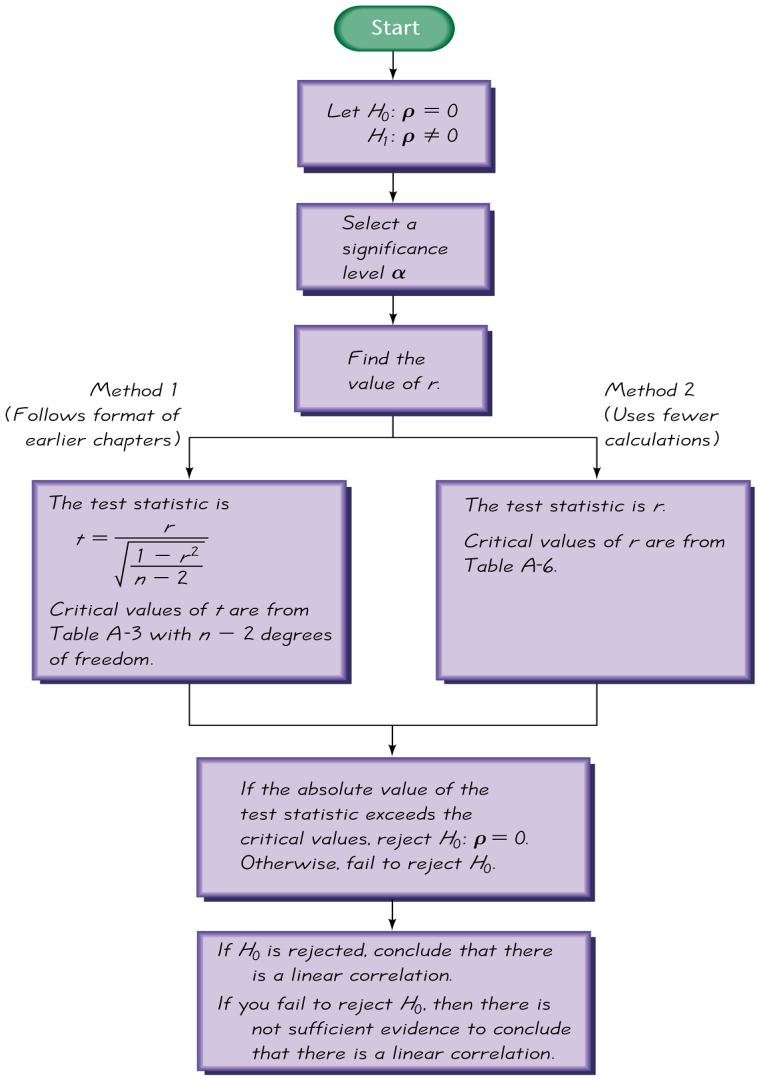


(f) Perfect negative correlation: $r = -1$



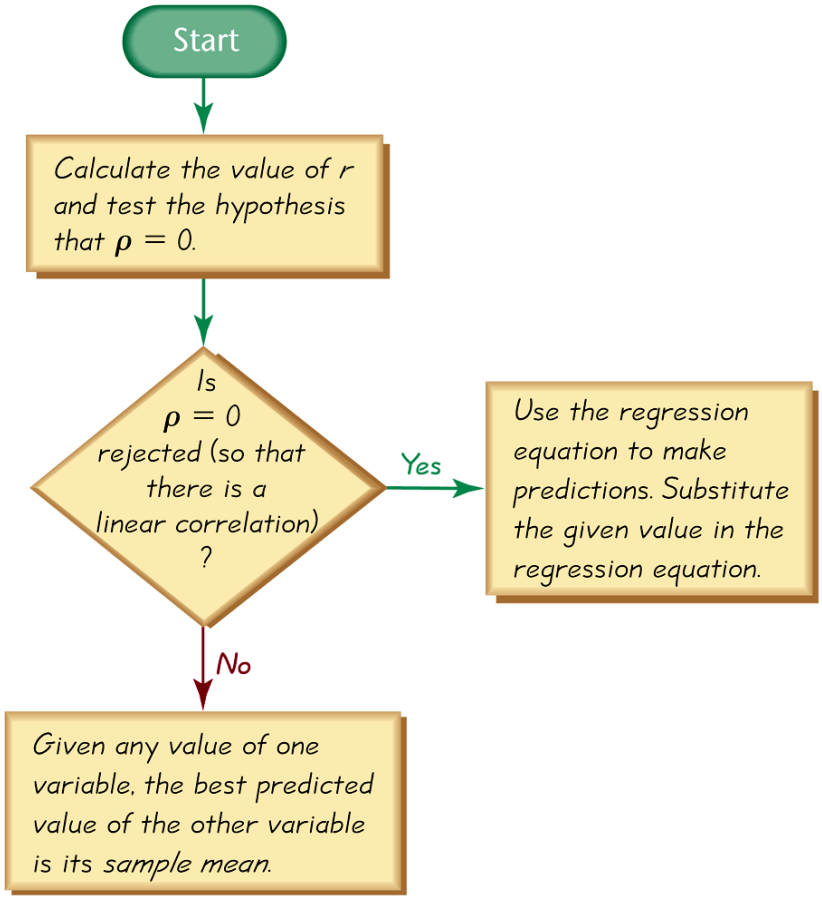
(g) No correlation: $r = 0$

(h) Nonlinear relationship: $r = -0.087$



How do we know if ***r*** is close enough to -1 or 1 to conclude that linear correlation exists? What about the natural variability that will occur from sample to sample? Well, normally in this case we would conduct a hypothesis test. This is actually quite simple. I have included a flow chart that describes the process. The only thing you need to know is that the population symbol for the correlation coefficient is "rho" (ρ).

Once we know if there seems to be enough evidence to assume there is linear correlation we can start to use our model:



11.7 Finding r-squared the Coefficient of Determination

The coefficient of determination

Another way to measure the usefulness of the model is to measure the contribution of x in predicting y. To do this, we calculate how much the errors of prediction of y were reduced by using the information provided by x.

Recall that $SS_{yy} = \sum \left(y_i - \bar{y} \right)^2$ = **total sample variation** of the observations around the sample mean for y, and $SSE = \sum \left(y_i - \hat{y}_i \right)^2$ = the **remaining unexplained sample variability** after fitting the line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

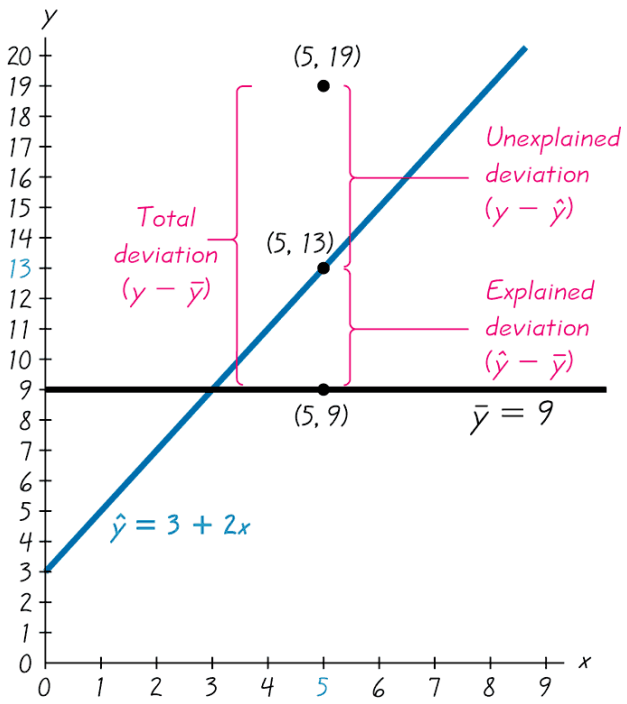
Then recall that if the model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, and x contributes nothing to the prediction of y, the best model becomes: $\hat{y} = \bar{y}$. *(If x does not contribute to the prediction of y then why is it that the best model becomes $\hat{y} = \bar{y}$? Answer: Mathematically, it is because the slope $\hat{\beta}_1$ for the model will be zero. Intuitively, imagine if I pointed to a guy on the street and said to you, "what's his IQ given that he wears a size ten shoe." You would ignore the shoe size info because it's probably useless here. Then the most logical guess for the guy's IQ is the average IQ since you have no other useful information.)*

Therefore, if x does not contribute to the prediction of y, $SS_{yy} \approx SSE$. However, if x does contribute to the prediction of y, $SS_{yy} > SSE$. Then a simple measure of the usefulness of the model could be formed as follows:

The **coefficient of determination** is

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}} = \frac{\text{Explained sample variability}}{\text{Total sample variability}}$$

The **coefficient of determination**, r^2, represents the proportion of the total sample variability around the mean of *y* that is explained by the linear relationship between *x* and *y* (see the diagram below).



Example 170: The following table gives male human heights and shoe sizes, find both r and r^2:

Heights	65	72	60	59	59
Shoe Sizes	6.5	9.5	7.5	7	5

Steps to finding r and r^2:

1. Create and fill in the preliminary calculation table below:

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	⋮	⋮	⋮	⋮	⋮
Totals	$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum y_i^2$	$\sum x_i y_i$

2. Calculate $SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$ = 28
3. Calculate $SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$ = 126
4. $SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$ = 10.7
5. Find $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{28}{126} \approx 0.222$
6. Calculate $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$
7. Calculate $r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$ and calculate $r^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$

Example 170 tech: Add a question with an interesting interpretation

Example 171 Use the following data to find r and r^2 (the table below will help you considerably)

Data

x	3	1	3	5
y	5	8	6	4

Table 11-1 Finding Statistics Used to Calculate r

	x	y	$x \cdot y$	x^2	y^2
	3	5	15	9	25
	1	8	8	1	64
	3	6	18	9	36
	5	4	20	25	16
Total	**12**	**23**	**61**	**44**	**141**
	↑	↑	↑	↑	↑
	Σx	Σy	Σxy	Σx^2	Σy^2

On the exam, you may have some of the work done for you like in this example.

Example 171 tech: A study looked at the relationship between the number of hours spent sitting each day and the LDL cholesterol number for each participant. The study, which used a prospective cohort study design, followed 40 participants for ten years. At the start of the study, all participants had similar LDL numbers but varied with respect to the number of hours they spent being sedentary. The results of the analysis are given below. Use the results to answer the questions that follow.

```
Model Summary

      S    R-sq  R-sq(adj)
32.5974  52.18%     50.92%


Coefficients

Term                    Coef  SE Coef  T-Value  P-Value
Constant               -23.4     20.5    -1.14    0.261
No.Sedentary hrs       14.00     2.17     6.44    0.000


Regression Equation: LDL = -23.4 + 14.00 * (No.Sedentary hrs)
```

a) Is there a significant linear relationship between these variables? If so, is the correlation positive or negative?
b) Interpret the coefficient of determination, r^2
c) Find and interpret the correlation coefficient, r.
d) Interpret the slope of the provided regression equation.
e) Do these results imply that being sedentary causes higher LDL levels?

11.8 Using the Model to Create an Estimation Interval

Using the model for estimation and prediction

Two common uses of our regression model are:

1. To estimate the mean value of y for a specific x value (section 11.8)
2. To predict an individual value of y for a specific x value (section 11.9)

In this section, we will consider the first of the two above cases. The formula below is the standard deviation of the sampling distribution of the estimator $\hat{y}$, when $\hat{y}$ is being used to estimate the mean value of y for a specific x value:

Standard error of $\hat{y}$: $\sigma_{(\hat{y})} = \sigma\sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$, recall that we use S = $\sqrt{S^2} = \sqrt{\frac{SSE}{n-2}}$ to estimate the value of σ.

Now that we know what our sampling error will be, we can form our *estimation interval to estimate the average value of y for a specific value of x:*

$$\hat{y} \pm t_{\alpha/2} S\sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

where $t_{\alpha/2}$ is based on n-2 degrees of freedom.

Example 171.5: A manager is worried about hiring older workers because he fears that they may be more likely to call out sick. He decides to look at the attendance records for sample of employees from the previous year. The data is given below. Find a 95% confidence interval for the average number of sick days used by a 53 year old employee.

Employee Age	18	28	38	48	58
Sick Days Last Year	15	12	9	5	2

S = 0.31623, $\hat{\beta}_0 = 21.14$, $\hat{\beta}_1 = -0.33$, $\bar{x} = 38$, $SS_{xx} = 1{,}000$

11.9 Using the Model to Create a Prediction Interval

The formula below is the standard deviation of the prediction error for the predictor $\hat{y}$, when $\hat{y}$ is being used to estimate the value of y for a specific x value:

Standard error of prediction: $\sigma_{(y-\hat{y})} = \sigma\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{SS_{xx}}}$ Recall that we use S = $\sqrt{S^2} = \sqrt{\frac{SSE}{n-2}}$ to estimate the value of σ.

Now that we know what our sampling error will be, we can form:

Our prediction interval to estimate a specific y value for a given x value

$$\hat{y} \pm t_{\alpha/2} S\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{SS_{xx}}}$$

where $t_{\alpha/2}$ is based on n-2 degrees of freedom.

Example 172: Find the 95% prediction interval for the height of a woman given that she has a shoe size of 6.

Heights (Y)	65	72	60	59	59
Shoe Size	6.5	9.5	7.5	7	5

S = 4.1924, $\hat{\beta}_0 = 44.421$, $\hat{\beta}_1 = 2.617$, $\bar{x} = 7.1$, $SS_{xx} = 10.7$

Step 1: Use the least squares line to find $\hat{y}$

$$\hat{y} = 44.421 + 2.617(6) = 60.123$$

Step 2: Find $t_{\alpha/2}$

Our degrees of freedom is n – 2 = 3, so our $t_{\alpha/2}$ = 3.182

Step 3: Find $S\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{SS_{xx}}}$

$$4.1924\sqrt{1+\frac{1}{5}+\frac{\left(6-7.1\right)^2}{10.7}} \approx 4.804$$

Step 4: Find the Margin of Error = ME = $t_{\alpha/2}\ S\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{SS_{xx}}}$

= 3.182 (4.804) = 15.286

Step 5: Finish by getting: $[\hat{y}-ME, \hat{y}+ME]$

= [60.123 – 15.286, 60.123 + 15.286]

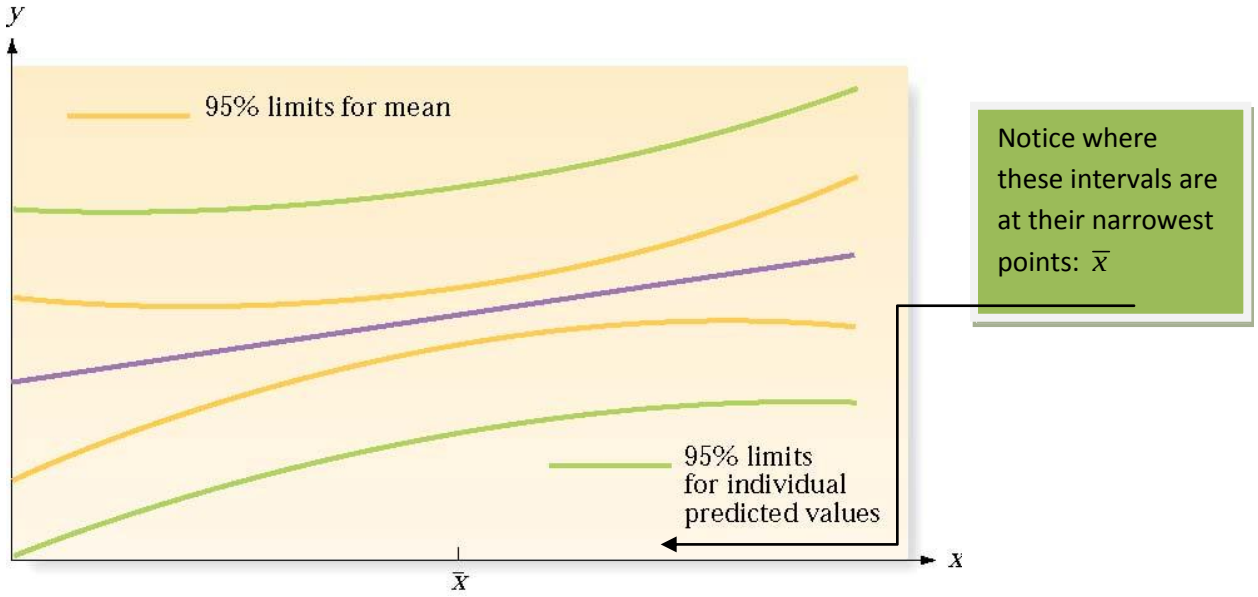
= [44.837, 75.409]

The interval above is not very good. It is very wide. How could we improve our results? The margin of error determines the interval width, so we need to reduce the size of:

$$t_{\alpha/2}\ S\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{SS_{xx}}}$$

If our sample size was larger we would have more degrees of freedom which would make $t_{\alpha/2}$ smaller, also the quantity 1 / n would be smaller. We could get a shorter interval by picking an x_p that is closer to $\bar{x}$. Finally, if SSE was smaller, S would be smaller, but to get a smaller SSE we need a better model. Recall however that the least squares line has the minimum SSE. This means that we need a better predictor variable (x), or we need to use a more complicated model.

Let's study the confidence interval widths and the prediction interval widths by looking at the following figure and by looking at the formulas themselves. Try to think about how these intervals can be made more narrow (i.e.-better).



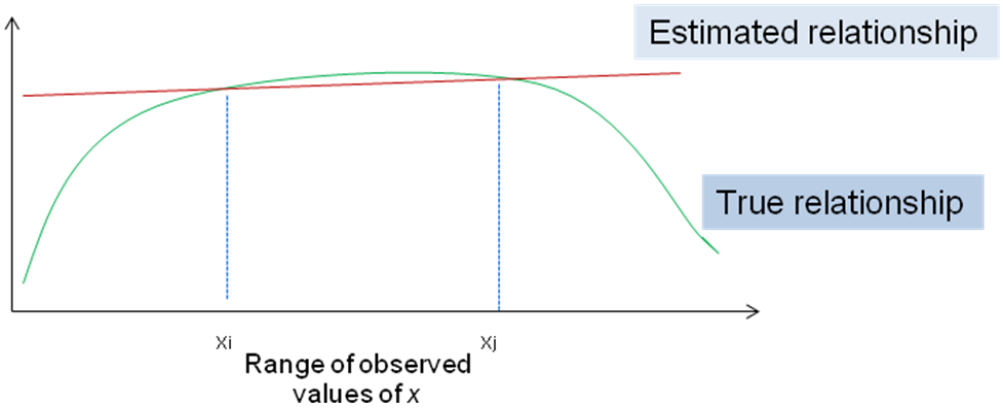
Now, that we have spent a good deal of time talking about linear regression, let's lay out some important guidelines.

Guidelines for Using the Regression Equation

1. If there is no linear correlation, don't use the regression equation to make predictions.
2. When using the regression equation for predictions, stay within the scope of the available sample data.*
3. A regression equation based on old data is not necessarily valid now.
4. Don't make predictions about a population that is different from the population from which the sample data were drawn.

*We should expand on the second guideline given above. If the data used to create our model used x values ranging from 6 to 20, don't try to make predictions about x values such as 1 or 45 because they are outside of the interval [6, 20] that we used to create our model.

- Estimating *y* beyond the range of values associated with the observed values of *x* can lead to large prediction errors.
- Beyond the range of observed *x* values, the relationship may look very different.



Example 173: Consider the following data:

X	1	2	3	4	5	6	7
Y	3	5	4	6	7	7	10

$\sum x = 28, \sum x^2 = 140, \sum y = 42, \sum y^2 = 284, \sum xy = 196, \bar{x} = 4, \bar{y} = 6$, n = 7

a. Find the least squares line

b. Calculate SSE

c. Calculate s^2

d. Find a 95% confidence interval for the mean value of y when $x_p = 2.5$

e. Find a 95% prediction interval for the value of y when $x_p = 4.6$

f. Would it be wise to make a prediction interval for y when $x_p = 10$?

Example 173 tech: The Minitab output below is from a study looking at the sum of gingival pocket depths and arterial plaque scores for 30 individuals. A high value for the sum of the pocket depths indicates gum disease, and a high value for the arterial plaque score indicates heart disease. The results of the analysis are given below. Use the results to answer the questions that follow.

```
Analysis of Variance

Source          DF   Adj SS   Adj MS  F-Value  P-Value
Regression       1  3388.73  3388.73    86.04    0.000
Error           28  1102.74    39.38
Total           29  4491.47

Model Summary

      S    R-sq  R-sq(adj)
6.27563  75.45%     74.57%


Coefficients

Term            Coef  SE Coef  T-Value  P-Value
Constant        7.77     2.23     3.48    0.002
PocketDepth  0.03161  0.00341     9.28    0.000


Regression Equation

PlaqueScore = 7.77 + 0.03161 * (PocketDepth)
```

a. Find the least squares line. Is there a significant linear relationship? If so, is it positive or negative?

b. Identify the MSE in the computer output

c. Find the correlation coefficient. Interpret this value.

d. Identify the coefficient of determination. Does it indicate that the sum of pocket depths is a strong predictor of arterial plaque scores?

e. The following interval is a 95% confidence interval for the mean arterial plaque score for patients with a pocket depth sum of 538. Interpret the interval: (22.1, 27.4)

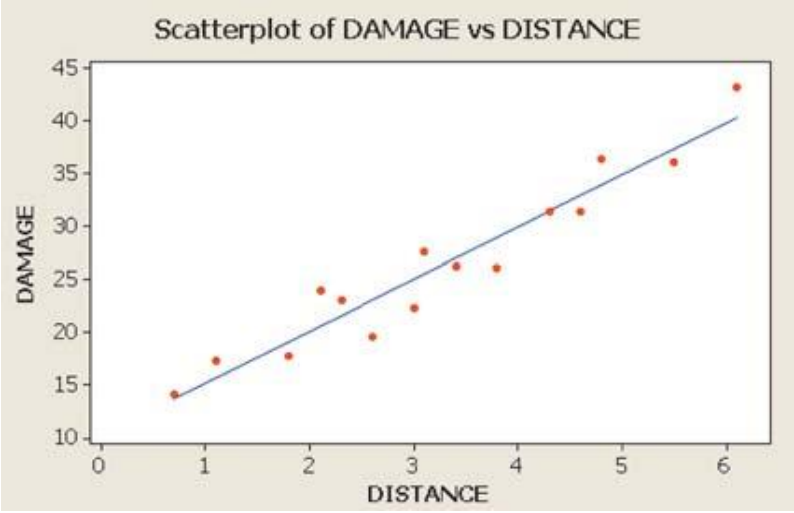
f. The following interval is a 95% prediction interval for the arterial plaque score for a patient with a pocket depth sum of 538. Interpret the interval: (11.7, 37.9)

g. The 30 values for the variable sum of pocket depths used in this problem ranged from 192 mm to 1,152 mm. Would it be wise to use this data to form a prediction interval for arterial plaque score for patients with a sum of pocket depths value of 1,325 mm?

A complete example

How does the proximity of a fire station (x) affect the damages (y) from a fire?

To answer this question we would need some real world data. After collecting the data, we would construct a scatter plot to see if it seems appropriate to try to fit a linear model to the data. Based on the scatter plot below, it seems plausible to try to fit a straight line model to the data.



There are 15 data points here. For example it appears that at 6 miles from the nearest fire station a home or building would have ~$44,000 worth of damage.

In the real world, the calculations are left up to a computer. Here is the output from SAS a popular software package:

Dependent Variable: DAMAGE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	841.76636	841.76636	156.89	<.0001
Error	13	69.75098	5.36546		
Corrected Total	14	911.51733			

Root MSE	2.31635	R-Square	0.9235
Dependent Mean	26.41333	Adj R-Sq	0.9176
Coeff Var	8.76961		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	10.27793	1.42028	7.24	<.0001	7.20960	13.34625
DISTANCE	1	4.91933	0.39275	12.53	<.0001	4.07085	5.76781

Output Statistics

Obs	DISTANCE	Dep Var DAMAGE	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	3.4	26.2000	27.0037	0.5999	21.8344	32.1729	-0.8037
2	1.8	17.8000	19.1327	0.8340	13.8141	24.4514	-1.3327
3	4.6	31.3000	32.9068	0.7915	27.6186	38.1951	-1.6068
4	2.3	23.1000	21.5924	0.7112	16.3577	26.8271	1.5076
5	3.1	27.5000	25.5279	0.6022	20.3573	30.6984	1.9721
6	5.5	36.0000	37.3342	1.0573	31.8334	42.8351	-1.3342
7	0.7	14.1000	13.7215	1.1766	8.1087	19.3342	0.3785
8	3	22.3000	25.0359	0.6081	19.8622	30.2097	-2.7359
9	2.6	19.6000	23.0682	0.6550	17.8678	28.2686	-3.4682
10	4.3	31.3000	31.4311	0.7198	26.1908	36.6713	-0.1311
11	2.1	24.0000	20.6085	0.7566	15.3442	25.8729	3.3915
12	1.1	17.3000	15.6892	1.0444	10.1999	21.1785	1.6108
13	6.1	43.2000	40.2858	1.2587	34.5906	45.9811	2.9142
14	4.8	36.4000	33.8907	0.8450	28.5640	39.2175	2.5093
15	3.8	26.1000	28.9714	0.6320	23.7843	34.1585	-2.8714
16	3.5	.	27.4956	0.6043	22.3239	32.6672	.

The data produces the following estimates (in thousands of dollars):

$$\hat{\beta}_0 = 10.28$$

$$\hat{\beta}_1 = 4.91$$

The estimated damages equal \$10,280 + \$4910 for each mile from the fire station, or

$$\hat{y} = 10.28 + 4.92x$$

The estimate of the standard deviation for our random error component of the model (ε) is highlighted above under root MSE: $s = 2.31635$. Most of the observed fire damages will be within $2s \cong 4.64$ thousand dollars of their respective predicted values when using the least squares line.

We would now like to check the usefulness of the model: Test that the true slope β_1 is 0.

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 > 0$$

You can see (next to the *Distance* variable) on our computer output that SAS automatically performs a two-tailed test, with a reported p-value < .0001. The one-tailed p-value is < .00005, which provides strong evidence to reject the null. Next to that line in the SAS printout you will see a 95% confidence interval on β_1 from the SAS output is 4.071 ≤ β_1 ≤ 5.768. This confirms that β_1 does not seem to be zero.

Another important measure of the model's utility is our coefficients of determination and correlation. Both of the values below indicate a strong linear relationship:

- The coefficient of determination, r^2, is .9235. (also highlighted on the SAS output)
- The coefficient of correlation, *r*, is $r = \sqrt{r^2} = \sqrt{.9235} = .96$

Now, we can use our model for predictions. You will see at the bottom of the SAS output we have asked the software to predict the damages for a fire that occurs 3.5 miles from the fire station. The following results were obtained:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$\hat{y} \cong 10.2 + 4.92(3.5) = 27.5$$
95% prediction interval is (22.324, 32.667)

We're 95% sure the damage for a fire 3.5 miles from the nearest station will be between $22,324 and $32,667. Finally, since the *x*-values in our sample range from .7 to 6.1, predictions about damages for *x*-values beyond this range will be unreliable.