Chapter 22

Simple Linear Regression: Continued

Chapter 21 presented quite a few methods for describing bivariate numerical data. Section 1 below presents a few more descriptive methods. (This material should have been in Chapter 21, but I decided that chapter was already too long!) Beginning with Section 2, this chapter will present methods of inference.

22.1 Is the Regression Line Any Good?

George Edward Pelham Box (1919–2013) was a great statistician and founded the Department of Statistics at the University of Wisconsin–Madison in 1960. He was my friend and taught me a great deal about what's important in Statistics. Being both famous and brilliant, George was often asked for pithy statements; one of my favorites of his is:

> Just because something is optimal, it doesn't mean it's any good!

George would especially direct this remark at mathematical statisticians who were solving problems that were not useful to scientists.

In the present context, we have found that the regression line is the best line for describing our data; but is it any good? I will explore this question in detail.

As stated in the previous chapter, each case enters the data set with two numbers: x_i and y_i, its values of the predictor and the response. After the regression line is determined, the case has two additional numbers associated with it: its predicted response $\hat{y}_i$ and its residual $e_i = y_i - \hat{y}_i$.

I want to examine the residuals. Each case has a residual; thus, there are n residuals:

$$e_1, e_2, e_3, \ldots, e_n.$$

With this one set of numbers, we are back in the realm of Chapters 1 and 2. We could draw a picture of them: a dot plot, a histogram or, perhaps, a kernel density histogram. Also, we could calculate their mean and their standard deviation.

To make this easier to follow, I will focus on two data sets from Chapter 21:

- The data on midterm and final exam scores for $n = 35$ students; and

Figure 22.1: Frequency histogram of the $n = 35$ residuals for the regression of final exam score on midterm exam score.

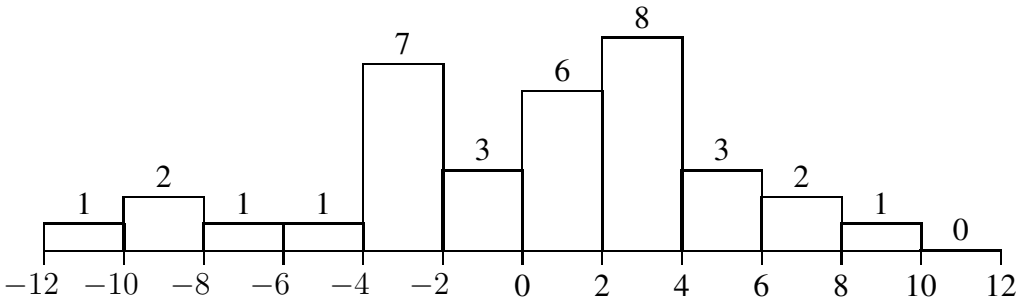
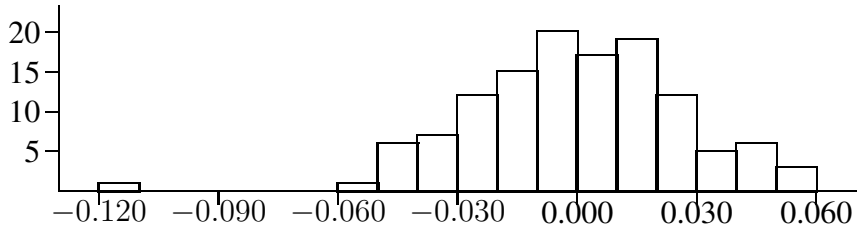


Figure 22.2: Frequency Histogram of the Residuals for the Batting Averages Data.



- The batting averages data for $n = 124$ American League baseball players.

Figure 22.1 presents a frequency histogram of the residuals for the exam scores and Figure 22.2 presents a frequency histogram of the residuals for the batting averages data. Let me make a few comments about these histograms.

1. For the exam scores' residuals, there are no outliers—the dot plot, not shown here, also revealed no outliers. Otherwise, the shape of the histogram is not recognizable to me. If I combine adjacent class intervals and have six class intervals instead of 12 (12 is a large number of intervals for $n = 35$) the histogram becomes a bit smoother.

2. For the batting averages data, I note one small outlier, a residual equal to approximately -0.120. Do you remember who this is?

   Answer: A negative residual means that the 'O' for the case is below the regression line. The fact that it's an outlier means that its distance below the regression line is much larger than any other case's distance below the regression line. It's Floyd Rayford!

   Excluding Mr. Rayford, the remainder of the histogram is approximately bell-shaped and symmetric.

3. Each histogram gives us a picture of the *sizes* of the residuals. For example:

(a) For the exam scores, 69% ($7 + 3 + 6 + 8 = 24$ of 35) of the residuals are between -4 and $+4$. Remembering what a residual measures, this means that by using the midterm score to predict the final score via the regression line, the predicted final is within four points of the actual final for 69% of the students. I view this as a *glass half full* statement. As a *glass half empty guy*, I prefer to say that for 31% of the students, the actual final deviates from the predicted final by more than four points.

(b) For the batting averages data, with the exception of Mr. Rayford, all predictions are within 60 points (remember, this is *baseball speak* for 0.060) of the actual value of the corresponding y. Although you cannot verify the following from Figure 22.2:

- 77% (95 of 124) residuals are between -30 and $+30$ points;
- 68% (84 of 124) residuals are between -25 and $+25$ points;
- 57% (71 of 124) residuals are between -20 and $+20$ points; and
- 30% (37 of 124) residuals are between -10 and $+10$ points;

Given that I **always** perform a regression analysis by using a computer software package, the counts illustrated in item 3 above are easy to obtain. They are very helpful for the researcher who is trying to decide whether the regression line is of value to the scientific problem being considered. Regarding the batting averages study, as a baseball fan I am disappointed with the residuals. It seems to me that predicting a batting average within 20 points is not particularly accurate and it is disappointing to learn that for 43% of the players the prediction fails to meet this modest goal. If you are a baseball fan, form your own opinion; you need not agree with me. Regarding the exams scores, I conjecture that as a student you have stronger feelings about exam scores than I do. Thus, I will leave it to you to decide whether predicting within four points for 69% of the students is a noteworthy achievement.

Let me make one more comment about the above; this is an issue that sometimes causes confusion. In Chapter 21, when I examined the scatterplot for the batting averages data, I labeled three players as being *isolated cases*. They are isolated in comparison to the other cases in the scatterplot. Now, we see that one case is an outlier. Thus, my first remark is to note that being isolated and being an outlier are two different notions. Thus, be careful about your use of these terms; sadly, many persons are careless and use them interchangeably, which leads to confusion. Below are some comments on how to keep these ideas separate.

1. We use the term *isolated* when considering two (or more) variables simultaneously; we use the term *outlier* for one variable. Note that we will **not** consider more than two variables simultaneously in these *Course Notes*.

2. A case is *isolated* if it is far away from other cases (with the possible exception that it could be close to another isolated case or cases). A case's residual is an *outlier* if it is unusually far away—in terms of vertical deviation—from the regression line. Often people are lazy and say *a case is an outlier* instead of the more accurate *a case's residual is an outlier*. Do not interpret *lazy* as a pejorative; if I am talking with a statistician—or if I simply forget—I will call a case an outlier.

3. If you think about the previous item, you will note that if a case's residual is an outlier, then it is an isolated case, but a case can be isolated without its residual being an outlier. For example, the residual for Wade Boggs [Don Mattingly] is 29 [52] points. They are far from other cases, but close to the line. Well, Boggs is reasonably close to the line; Mattingly had the second largest residual (third largest absolute residual) among the 124 players; thus, he is *almost* an outlier.

22.1.1 The Mean and Standard Deviation of the Residuals

It can be shown that for every set of data, there are two restrictions on the values of the residuals.

$$\sum e_i = 0 \text{ and } \sum x_i e_i = 0. \tag{22.1}$$

(If you read the optional Appendix in Chapter 21, the above is simply Equation 21.8.) From the first of these equations, we see that the mean of the residuals, $\bar{e}$, equals 0 for every set of data. In symbols,

$$\bar{e} = 0, \text{ for every set of data.}$$

Let's **not** rush on to the next topic; this is a very important equation. Similar to my feelings about the Law of the Preservation of Mediocrity in Chapter 21, I would not like the regression line very much if this equation were not true. Here is why.

The fact that $\sum e_i = 0$ means that the regression line passes through the *center* of the data in the sense that: some cases are above the line and some are below, but the sum of the distances above the line cancel exactly the sum of the distances below the line.

Now that we know the center—mean—of the distribution of the residuals, as in Chapter 1 we turn to the determination of the amount of spread. In Chapter 1, for each x_i in the data set, we compared it, via subtraction, to its mean in order to obtain its deviation: $(x_i - \bar{x})$. Residual e_i is its own deviation because $\bar{e} = 0$. Thus, the sum of squared deviations of the residuals is simply the sum of squared residuals, which we denote by SSE:

$$\text{SSE } = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2. \tag{22.2}$$

Enrichment: Note that statisticians are a bit confused about their usage of the letters 'e' and 'r.' Why do I say this?

- We use r to denote the correlation coefficient and, hence, cannot use it for residuals. We use e for residuals, that, historically, were also called errors. Hence, the sum of the squared residuals is denoted SSE, where the E is for the word error. One reason statisticians replaced the name *error* with residual is we got tired of the following exchanges between a statistician (S) and a client (C):

  S: Let's look at a list of your errors.

  C: I didn't make any errors!

  or

  S: Here is a list of the errors from our regression analysis.

  C: Fix those errors, then get back to me!

Also, we use SSE for sum of squared residuals because SSR—see below—is reserved for the *sum of squares due to regression*.

Back in Chapter 1, I remarked that statisticians and mathematicians disagree on what to do with the sum of squared deviations: the mathematicians divide it by n and the statisticians divide it by the degrees of freedom, $(n-1)$. Recall, also, that there are $(n-1)$ degrees of freedom for the deviations because they are subject to one constraint: they must sum to zero. As stated above in Equation 22.1, the residuals have two constraints; hence, they have $(n-2)$ degrees of freedom. As a result, statisticians define the variance and standard deviation of the residuals as follows.

Definition 22.1 (The variance and standard deviation of the residuals.) *The variance of the residuals is*

$$s^2 = \frac{SSE}{n-2}, \tag{22.3}$$

where SSE is defined in Equation 22.2.

The standard deviation of the residuals is

$$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}. \tag{22.4}$$

We now have three standard deviations associated with a regression analysis:

- The standard deviation of the values of x, denoted by s_1.
- The standard deviation of the values of y, denoted by s_2.
- The standard deviation of the residuals, denoted by s.

Note that it would *make sense* to denote the last of these by s_e with the subscript e to remind us that we are measuring the spread in the residuals. Statisticians don't do this, for at least two reasons:

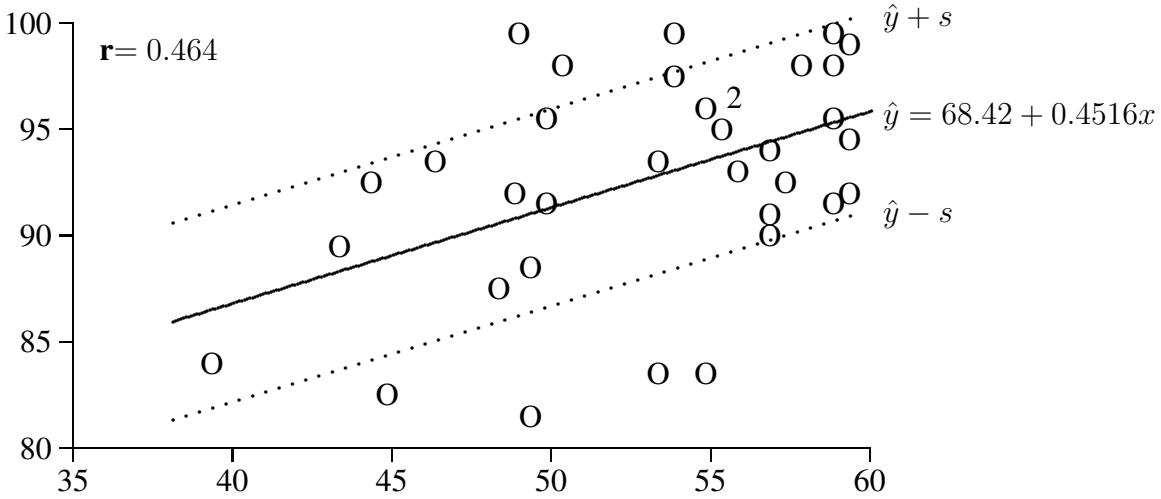
1. We are lazy and want to save the effort of typing a subscript whenever possible.
2. The lack of a subscript *honors* the standard deviation of the residuals as being the most important of the three standard deviations. This is analogous to how certain pop stars—and numerous Brazilian soccer players—have only one name—Cher, Prince, Pelé, etc. (I apologize for my pop culture references being so dated. Also, of course, Prince became the favorite of all math-ophiles when he replaced his name with a symbol!)

For the exam scores data, $s = 4.635$. Because $\bar{e} = 0$, the Empirical Rule from Chapter 2 (Result 2.2) tells us that approximately 68% of the residuals are between -4.635 and $+4.635$; i.e., approximately 68% of the predicted values are within 4.635 of the actual response.

The Empirical Rule is actually a pretty bad approximation for the exam scores data. As we saw above, 69% of the residuals are between -4 and $+4$. Here's another way to see that it is bad; by actual count—details not given— fully 80% (28 of 35) residuals are between -4.635 and $+4.635$. As noted, in Chapter 2, the Empirical Rule is unreliable for small values of n or for distributions that are not bell-shaped; and the exam scores' residuals suffer both of these maladies.

By the way, if you prefer pictures to counting, please refer to Figure 22.3. This figure presents the scatterplot of the 35 pairs of exams scores with three parallel lines superimposed:

Figure 22.3: Final Exam Score Versus Midterm Exam Score for 35 Students.



- The regression line $\hat{y} = 68.42 + 0.4516x$, drawn as a solid line.
- The line $\hat{y} + s$, drawn as a dotted line.
- The line $\hat{y} - s$, drawn as a dotted line.

A quick examination of this picture shows that four cases fall below the line $\hat{y} - s$ and three cases fall above the line $\hat{y} + s$; thus, the remaining 28 cases fall between the dotted lines. In words, 28 cases have a residual between the value $-s$ and $+s$, as I reported earlier.

For the batting averages data, $s = 0.0268$, just under 27 points. By actual count—details not given—slightly more than 72% (87 of 124) of the residuals are between -0.0268 and $+0.0268$. Here, n is pretty large and, excepting the outlier, the residuals have a bell-shaped distribution. The 72% is larger than the Empirical Rule's 68% because Mr. Rayford's residual inflates the value of s. In particular, if one deletes Mr. Rayford from the data set, and runs the regression program again, s becomes 0.0248, a reduction of 7.5% from the earlier $s = 0.0268$. Sadly, however, for the new data set with $n = 123$, only 65% (80 of 123) of the residuals are between -0.0248 and $+0.0248$; approximations can be so annoying! To further confuse matters, if we round s to 25 points, then 67% (82 of 123) of the residuals fall between $-s$ and $+s$. The moral: The Empirical Rule is a useful guide, but it's not exact; if you want exact, look at the list of residuals and count!

To summarize the above, for the exam scores and batting average data sets, we look at the value of s and, using the Empirical Rule, can make a subjective assessment as to whether the predictions from the regression line are scientifically useful.

Another popular approach is to measure how well the regression line predictions—which are obtained from the best line for using X to predict Y—compare to predictions that ignore X.

Well, if we ignore X, then our data set becomes a collection of n values of Y. Based on the Principle of Least Squares the best predictor of each y_i using only the values of

$$y_1, y_2, y_3, \ldots, y_n,$$

is $\bar{y}$. (Allow me to casually call $\bar{y}$ the best predictor. As I show in the optional Appendix to Chapter 21, for data on one variable, according to the Principle of Least Squares the mean is the best predictor/describer of the data. If you did not read this material, that is fine, but you will need to take my use of $\bar{y}$ on faith.)

If I predict y_i by $\bar{y}$, the difference is $(y_i - \bar{y})$. Squaring these differences and summing them, we get

$$\text{SSTO} = \sum (y_i - \bar{y})^2, \tag{22.5}$$

where the symbol SSTO is called the *total sum of squares*. (In very old textbooks, our SSTO is called the *adjusted* total sum of squares.) From Chapter 2 we know that

$$s_2^2 = \text{SSTO}/(n-1) \text{ and } s_2 = \sqrt{\text{SSTO}/(n-1)}.$$

Recall that for the exam scores data, $s_2 = 5.154$ and for the batting averages data, $s_2 = 0.0320$. Also, as discussed above, for the exam scores data, $s = 4.635$ and for the batting averages data, $s = 0.0268$. For the exam scores data:

$$s/s_2 = 4.635/5.154 = 0.899;$$

thus, using the midterm to predict the final reduces the standard deviations of the errors in the predictions by 10.1%. For the batting averages data:

$$s/s_2 = 0.0268/0.0320 = 0.838;$$

thus, using the 1985 batting average to predict the 1986 batting average reduces the standard deviations of the errors in the predictions by 16.2%.

In my opinion, comparing s to s_2, as I have done above, is a good way to measure the relative usefulness of using X versus not using X. I need to mention, however, that this is not the only comparison that scientists make. Indeed, I must admit that my personal evidence is overwhelming that the comparison below is more popular than my favored comparison of s versus s_2. First, however, we must make a side trip into the next subsection.

22.1.2 The Analysis of Variance Table

In Table 21.3 in Chapter 21 I presented *edited* Minitab output for the regression analysis of the exam scores data for $n = 35$ students. This table, with an additional column added—SE(Fit)—is reproduced in Table 22.1 because we will need it often later in this chapter. In both of these tables I *deleted* the **Analysis of Variance Table** given in the Minitab output. By the way, we abbreviate the *Analysis of Variance Table* as the ANOVA table. We could get through this chapter without

Table 22.1: Edited Minitab output for the regression of final exam score on midterm exam score for 35 students.

The regression equation is: Final = 68.4 + 0.452 Midterm

Predictor	Coef	SE(Coef)	T	P
Constant	68.420	7.963	8.59	0.000
Midterm	0.4516	0.1501	3.01	0.005

S = 4.635 R-Sq = 21.5%

Obs	Midterm	Final	Fit	SE(Fit)	Residual
1	39.0	83.5	86.032	2.213	-2.532
2	43.0	89.0	87.838	1.665	1.162
3	44.0	92.0	88.290	1.534	3.710
4	44.5	82.0	88.515	1.470	-6.515
5	46.0	93.0	89.193	1.285	3.807
6	48.0	87.0	90.096	1.063	-3.096
7	48.5	91.5	90.322	1.014	1.178
8	49.0	99.5	90.548	0.968	8.952
9	49.0	88.0	90.548	0.968	-2.548
10	49.0	81.0	90.548	0.968	-9.548
11	49.5	91.0	90.773	0.926	0.227
12	49.5	95.0	90.773	0.926	4.227
13	50.0	97.5	90.999	0.888	6.501
14	53.0	83.0	92.354	0.784	-9.354
15	53.0	93.0	92.354	0.784	0.646
16	53.5	97.0	92.580	0.791	4.420
17	53.5	99.0	92.580	0.791	6.420
18	54.5	95.5	93.031	0.825	2.469
19	54.5	83.0	93.031	0.825	-10.031
20	55.0	94.5	93.257	0.851	1.243
21,22	55.5	96.0	93.483	0.883	2.517
23	55.5	92.5	93.483	0.883	-0.983
24	56.5	93.5	93.934	0.962	-0.434
25	56.5	89.5	93.934	0.962	-4.434
26	56.5	90.5	93.934	0.962	-3.434
27	57.0	92.0	94.160	1.007	-2.160
28	57.5	97.5	94.386	1.056	3.114
29	58.5	99.0	94.838	1.162	4.162
30	58.5	95.0	94.838	1.162	0.162
31	58.5	91.0	94.838	1.162	-3.838
32	58.5	97.5	94.838	1.162	2.662
33	59.0	91.5	95.063	1.218	-3.563
34	59.0	98.5	95.063	1.218	3.437
35	59.0	94.0	95.063	1.218	-1.063

Table 22.2: Analysis of Variance table from the Minitab analysis of the regression of final exam score on midterm exam score for 35 students.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	194.38	194.38	9.05	0.005
Residual Error	33	708.81	21.48		
Total	34	903.19			

mentioning ANOVA tables, but if you do any statistical analyses beyond this chapter, you may well run into them.

Table 22.2 is the ANOVA table from Minitab's regression analysis of the exam scores data. All statistical software packages I have seen give a similar ANOVA table for regression. Let me take a few minutes to explain the connection between this table and our current work.

First, ignore the last three columns—those headed *MS*, *F* and *P*. The *DF* column presents degrees of freedom and the *SS* column presents various sum of squares, both identified with the *feature* in the *Source* column. If you recall that $n = 35$, you see that the degrees of freedom for the total sum of squares (SSTO) is indeed $(n - 1) = (35 - 1) = 34$, as Minitab states. Also, the degrees of freedom for the error sum of squares, SSE, is $(n - 2) = (35 - 2) = 33$, as Minitab states. This table leads to some obvious questions:

1. What is the *regression sum of squares*? Why does it have one degree of freedom?

2. Why do the first two sums of squares sum to the third? ($194.38 + 708.81 = 903.19$.)

To answer these questions, I must go back to the beginning. We start with an arbitrary case 'i' in our data set and look at the deviation of its response y_i from the mean response, $\bar{y}$:

$$(y_i - \bar{y}).$$

We rewrite this as

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}). \tag{22.6}$$

In words, this equation states:

> The deviation of a response from its mean is the sum of two terms: The deviation of the response from its predicted value **and** the predicted value minus the (overall) mean response.

This equation is so important that I want to illustrate it with two cases from our exam scores data:

- Consider the case with $x = 57.5$ and $y = 97.5$. You may verify that for this case, $\hat{y} = 94.386$. (You may verify this by plugging $x = 57.5$ into the equation of the regression line, or, more

easily, locate this case as observation 28 in Table 22.1.) Also, recall that $\bar{y} = 92.257$. For this case, Equation 22.6 becomes:

$$(97.5 - 92.257) = (97.5 - 94.386) + (94.386 - 92.257) \text{ or}$$

$$5.243 = 3.114 + 2.129 = 5.243.$$

- Consider the case with $x = 56.5$ and $y = 89.5$. You may verify that for this case, $\hat{y} = 93.934$. (This case is observation 25 in Table 22.1.) For this case, Equation 22.6 becomes:

$$(89.5 - 92.257) = (89.5 - 93.934) + (93.934 - 92.257) \text{ or}$$

$$-2.757 = -4.434 + 1.677 = -2.757.$$

If you take Equation 22.6 and sum both sides over all values of i, then obviously the equality is preserved:

$$\sum(y_i - \bar{y}) = \sum(y_i - \hat{y}_i) + \sum(\hat{y}_i - \bar{y}).$$

If we square the three terms in Equation 22.6 we on longer get equality; in my cases above:

$$(5.243)^2 = 27.489 \text{ does not equal } (3.114)^2 + (2.129)^2 = 14.230 \text{ and}$$

$$(-2.757)^2 = 7.601 \text{ does not equal } (-4.434)^2 + (1.677)^2 = 22.473.$$

If, however, we square the three terms and then sum over all cases, the equality is preserved:

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2. \tag{22.7}$$

This seems to be a magical result, but it is simply the n dimensional version of the Pythagorean Theorem. You should recognize two of the terms in Equation 22.7; it becomes:

$$\text{SSTO} = \text{SSE} + \sum(\hat{y}_i - \bar{y})^2.$$

I name this last term the sum of squares due to regression and write it as SSR. Thus, we see that the identity

$$\text{SSTO} = \text{SSE} + \text{SSR}$$

in Table 22.2 is not an accident; this equation will be true for every regression analysis.

By the way, I will ask you to take it on faith that the degrees of freedom for SSR is 1. It's easy to remember: both sum of squares and degrees of freedom sum in an ANOVA table.

For any regression analysis, all three of our sums of squares must be nonnegative numbers. In addition:

- SSTO is positive, because $s_2 > 0$.
- SSE is zero if, and only if, all points lie on a straight line which happens if, and only if $r = \pm 1$.

- SSR is zero if, and only if, the regression line has slope equal to zero which happens if, and only if, $r = 0$.

Thus, for example,

$$\text{SSE} \leq \text{SSTO}\,.$$

We can also see this result by applying *logic*, as follows. SSE minimizes the sum of squared errors around all lines; thus, it cannot exceed SSTO, which is the sum of squared errors around a *particular line*, namely, the horizontal line with intercept equal to $\bar{y}$.

The above considerations lead to the following definition.

Definition 22.2 (The coefficient of determination, R^2.) *The coefficient of determination is denoted by R^2 and is given by:*

$$R^2 = (\frac{\textit{SSTO} - \textit{SSE}}{\textit{SSTO}}) \textit{ which also equals } (\frac{\textit{SSR}}{\textit{SSTO}}) \textit{ or } (1 - \frac{\textit{SSE}}{\textit{SSTO}}). \tag{22.8}$$

Let's look at the first expression for R^2. The numerator is the total squared error when ignoring X (SSTO) minus the the total squared error when using X via the best possible line (SSE). Various picturesque ways to describe this difference include:

- SSTO minus SSE measures the amount of squared error in the y's that can be _____

  (Choose one): explained, removed, accounted for, explained, ...

  by a linear relationship with the x's.

Thus, for example if you have regression data and obtain SSTO $= 100$ and SSE $= 20$, then 80 of the 100 total squared error is explained (my choice of verb) by a linear relationship between y and x. But 80 what? How do we interpret this number? **Answer:** Look at the denominator of R^2. We compare, by dividing, the *squared error removed* with the original amount of squared error. Thus, for my current fictional numerical example,

$$R^2 = \frac{\text{SSTO} - \text{SSE}}{\text{SSTO}} = \frac{100 - 20}{100} = \frac{80}{100} = 0.80.$$

Usually, R^2 is reported as a percentage, for a reason you will see in a moment. Thus, instead of saying $R^2 = 0.80$ it is standard to say $R^2 = 80\%$. Returning to my picturesque statement for this value, we get:

> Eighty percent of the squared error in the y's can be explained by x.

Let's return to real data. Using the sums of squares in Table 22.2 we obtain

$$R^2 = \text{SSR}/\text{SSTO} = 194.38/903.19 = 0.215.$$

Recall that for the exam scores data, the correlation coefficient equals 0.464. Thus, $r^2 = (0.464)^2 = 0.215$, rounded to three digits. Is this agreement between r^2 and R^2 an accident? No.

Result 22.1 *For every regression analysis,*

$$r^2 = R^2. \tag{22.9}$$

Be careful with this result. Its main benefit is that it gives us another interpretation of the correlation coefficient r; namely, the square of correlation coefficient has the same interpretation as R^2 in terms of *explaining* squared errors. This helps us see the validity of property four of the correlation coefficient given in Result 21.1. Among other things, this property said that $r = +0.60$ reflects the same strength of a linear relationship as $r = -0.60$; we can see that this is true in the sense they both give the same value, 36%, of R^2.

There is a bizarre misinterpretation of Result 22.1 and I must comment on it. Let me illustrate this with a fictional conversation between Researchers A and B.

> A: I published my regression analysis; now all the world knows that my r is 60%.
>
> B: You are a bad person!
>
> A: Huh?
>
> B: You deliberately deceive people. It would be more honest to report that R^2 equals 36%. You are trying to trick people into believing you have somehow accounted for 60% when, in fact, you have accounted for only 36%. Shame, shame on you!

If you haven't guessed my position, Researcher B is misguided. I disagree with Researcher B for two reasons:

1. There is nothing *natural* about squaring errors. Indeed, when we squared deviations in Chapters 1 and 2 to obtain the variance, we quickly found that to get a summary that *has meaning* we need to take the square root to obtain the standard deviation. Thus, arguably, R^2 is not a natural measure.

2. When we discussed the regression effect in Chapter 21, we found that Equation 21.6 justifies referring to $r = 0.60$ as 60%; because 60% of the advantage in x is inherited by $\hat{y}$.

Don't get me wrong; I think that R^2 is an interesting summary of a regression, but it is not the whole story. If you decide that you prefer R^2 to r, that is fine; just don't use a questionable argument to bully people!

One final comment on this section. I have shown you two ways to decide whether using X via the regression line is better than not using it. The first was to compare s to s_2; and the second was to compute R^2. You may have noticed that these ways are mathematically equivalent. I will spare you the algebra, but note that

$$(s/s_2)^2 = [(n-1)/(n-2)](1 - R^2).$$

For the exam scores data,

$$(s/s_2)^2 = (4.635/5.154)^2 = 0.815057 \text{ and}$$

$$[(n-1)/(n-2)](1 - R^2) = (34/33)(1 - (194.38/903.19)^2 = 0.808566,$$

which are the same, except for my round-off error. (Sorry.)

22.2 The Simple Linear Regression Model

We now turn to inference for simple linear regression. As you might imagine, inference will require us to make one of the following three assumptions:

- We have a smart random sample from a finite population;
- We have a dumb random sample from a finite population; or
- We assume that we have i.i.d. trials.

Indeed, our approach in Part II of these notes always has been to make one of these assumptions in order to perform inference. Recall also that:

- Having a dumb random sample yields i.i.d. trials;
- Provided the sample size, n, is 5% or fewer of the population size, N, probabilities for dumb sampling are a good approximation to probabilities for smart sampling; and
- Literal random samples are very rare in science. Usually, the best we can do is feel comfortable in making the WTP (Willing to Pretend) assumption of Chapter 10 (Definition 10.3) on page 240.

Before I talk about random sample assumptions, however, there is another feature of regression that I need to introduce. This feature is similar to the observational study versus experimental study dichotomy of Chapter 15.

In the three main studies of Chapter 21—spiders, baseball players and Statistics 371 students—each case entered the study with two numerical features that the researcher measured/determined. In the vernacular, both X and Y are random. While Y is always random in regression analysis, in many studies the n values of X are selected by the researcher and assigned by randomization to the n cases available for study. Let me give you an example:

Example 22.1 (A hypothetical study on crop yield.) *A researcher has n one-acre* plots *of land available for study. The goal is to investigate the effect of a particular fertilizer on the yield of a particular crop. Before conducting the study, the researcher selects n values of the* concentration of the fertilizer *for study. The concentration values need not be distinct. The n concentration values are assigned to plots by randomization. Let x_i denote the concentration assigned to plot number i and let y_i be the yield—say, in bushels or pounds—of the crop from the plot.*

To summarize, there are n cases (one-acre plots) and each case yields two numbers: the concentration of the fertilizer, x, and the crop yield, y. In other words, we have regression data *with Y random and X not random.*

I am now ready to show you the **simple linear regression model**. We assume the following relationship between X and Y:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ for } i = 1, 2, 3, \ldots n. \tag{22.10}$$

Note the following features of this equation.

1. I use upper case letters—i.e., Y_i and X_i—to emphasize that we write down this model **before** we collect any data. This means, in particular, that we will consider **probabilities** for the values taken on by
$$Y_1, Y_2, Y_3, \ldots, Y_n.$$

2. Note that even though we use upper case letters for the values of the predictor, we **do not** consider **probabilities** for the values taken on by
$$X_1, X_2, X_3, \ldots, X_n,$$
for one of the following two reasons:

   (a) In the non-random X case introduced above, it makes no sense to calculate probabilities for the values of the predictor; they are deliberately (i.e., non-randomly) selected by the researcher.

   (b) In the random X case, our inference **conditions** on the n values of X in the data set.

   A thorough exploration of why statisticians and scientists condition on the values of X in the *random X situation* is beyond the scope of these notes. I will remark, however, that while the reasons include mathematical necessity—some of the formulas later in this chapter become invalid for random X without conditioning—they also include scientific usefulness. For example, in most scientific applications the main interest is on the behavior of the response, conditional on the value of the predictor. Suppose that you want to use height (X) to predict weight (Y) for a *particular randomly selected man of interest* then you want to condition on (i.e., use) *his height* to make your prediction; i.e., you want to condition on his being 76 inches tall or 66 inches tall rather than have an *unconditional prediction* over all possible heights!

   Let me note that the idea of conditioning is familiar to the reader of these notes. The Skeptic's Argument in Part I has the effect of conditioning on the response values actually obtained. Even statisticians who don't like randomization-based inference condition on the marginal totals in the 2×2 contingency table in order to perform Fisher's Test. I could list more examples of conditioning, but I prefer to *stick to our task*.

3. After collecting data, the researcher will have the n observed pairs of values of the predictor and response.

4. The values β_0 and β_1 are parameters of the model; this means, of course, that they are numbers and by changing either or both of their values we get a different model. The actual numerical values of β_0 and β_1 are known by Nature and unknown to the researcher; thus, the researcher will want to estimate both of these parameters and, perhaps, test hypotheses about them.

5. The ϵ_i's are random variables with the following properties. They are i.i.d. with mean 0 and variance σ^2. Thus, σ^2 is the third parameter of the model. Again, its value is known

by Nature but unknown to the researcher. It is very important to note that we assume these ϵ_i's, which are called **errors**, are statistically independent. In addition, we assume that every case's error has the same variance.

Oh, and by the way, not only is σ^2 unknown to the researcher, the researcher does not get to observe the ϵ_i's. Remember this: all that the researcher observes are the n pairs of values of (X, Y).

6. Our inference procedures below are based on the assumption that the simple linear regression model is **true** or **correct**. A scientist won't know whether the model is correct. Indeed, another of George Box's pithy statements is:

   All models are wrong, but some are useful.

   This reflects the belief that we should be careful about using the simple linear regression model, as well as more complicated models. If you study regression beyond this chapter, I hope that a good amount of time is spent on how to check whether the assumptions of the model are close enough to being correct for the model to be useful.

Now, we look at some consequences of our model. The results below follow quite easily from the rules of means and variances familiar to the undergraduate Statistics major. In these notes, I have made only vague references to these rules; thus, don't worry if the algebra below is confusing.

Remember, the Y_i's are random variables; the X_i's are viewed as constants. The mean of Y_i given X_i is denoted by $\mu_{Y_i|X_i}$. First, we note that

$$\mu_{Y_i|X_i} = \beta_0 + \beta_1 X_i + \mu_{\epsilon_i},$$

because the mean of a constant ($\beta_0 + \beta_1 X_i$) is the constant. Finally, remembering the mean of the error term, ϵ_i equals 0, we get:

$$\mu_{Y_i|X_i} = \beta_0 + \beta_1 X_i. \tag{22.11}$$

The variance of Y_i is

$$\sigma^2_{Y_i} = \sigma^2_{\epsilon_i} = \sigma^2, \tag{22.12}$$

because the variance of a constant ($\beta_0 + \beta_1 X_i$) is 0.

The important facts for you to know are:

- The relationship between X and Y is such that the mean of Y given the value of X is a linear function of X with y-intercept given by β_0 and slope given by β_1.

- The variance of the Y's around their means (remember the mean depends on X) is σ^2, for every case regardless of its value of X.

22.2.1 Point Estimates of the Slope, Intercept and Variance

The first issue we turn to is: How do we use data to estimate the values of β_1, β_0 and σ^2? This turns out to be quite easy.

First, I will refer to Equation 22.11 as **the equation of the population regression line**. We estimate the slope, β_1, and intercept, β_0, by using the Principle of Least Squares, as we did in Chapter 21. In particular, the point estimate of β_1 is

$$b_1 = r(s_2/s_1), \tag{22.13}$$

and the point estimate of β_0 is

$$b_0 = \bar{y} - b_1\bar{x}. \tag{22.14}$$

The point estimate of σ^2 is a bit trickier. First, by rewriting Equation 22.10 we obtain:

$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i. \tag{22.15}$$

If we were Nature, then we could calculate the n observed values of ϵ and calculate their variance. Sadly, we are not Nature and must proceed as follows. Replace Y_i and X_i by their observed values and replace the unknown β_0 and β_1 by their point estimates. After these four replacements, the right side of Equation 22.15 becomes:

$$y_i - b_0 - b_1 x_i = y_i - \hat{y}_i = e_i.$$

Thus, in a sense, the residual, e_i, is the *sample version* of the error, ϵ_i. Thus, it makes sense to use the variance of the residuals to estimate the variance of ϵ_i. We do this and our result is that the point estimate of σ^2 is

$$s^2 = \frac{\text{SSE}}{n-2}.$$

22.3 Three Confidence Intervals, a Prediction Interval and a Test

It is possible to estimate β_1, β_0 and σ^2 with confidence. In this section I will address these problems along with two closely related problems. Finally, I will present a test of hypotheses for the slope of the population regression line.

First, let's deal with σ^2. There is a confidence interval formula for σ^2, but I will not present it, for a variety of reasons that (sadly?) I have no time to discuss.

There exist algebraic formulas for confidence intervals for both β_1 and β_0, but, frankly, they are no fun and nobody uses them by hand. Instead, I will show you how to obtain these confidence intervals from Minitab output.

Table 22.1 on page 592 presents our familiar Minitab output for the regression of final exam score on midterm exam score for $n = 35$ students. I will use this output to illustrate the following result.

Result 22.2 (Confidence interval estimates of the slope and intercept.) *The confidence interval estimates of the slope β_1, and intercept, β_0, of the population regression line are*

$$b_1 \pm t^*(SE(b_1)) \text{ and} \tag{22.16}$$

$$b_0 \pm t^*(SE(b_0)) \tag{22.17}$$

respectively.

In these formulas, t^* is obtained from the t-curve with $df = (n-2)$, with our usual method, first introduced for the t-curve in Chapter 17. The expression "SE(b_1)" ["SE(b_0)"] is the *estimated standard error of* b_1 [b_0] and its value must be obtained from computer output.

Based on our earlier work in these notes, it is reasonable to wonder whether the confidence level in the above result is exact or an approximation. **For all of the inference procedures in this chapter, including the above result, the confidence levels, prediction probabilities and P-values are exact with the additional assumption that the error terms, ϵ_i, have a Normal curve for their pdf.** Without this assumption, the results are approximations. Sadly, we do not have time to explore the quality of these approximations.

Suppose that I choose 95% for my confidence level; because $df = n - 2 = 35 - 2 = 33$, you may verify by using our t-curve website that $t^* = 2.035$. For the slope, using Table 22.1, we find that the estimated standard error of the estimated slope is 0.1501; it's in the *SE(Coef)* column in the *Midterm* row. Thus, the 95% confidence interval estimate of β_1, the population slope, is:

$$0.4516 \pm 2.035(0.1501) = 0.4516 \pm 0.3055 = [0.1461, 0.7571].$$

First, I conclude, qualitatively, that the population slope is a positive number. Recalling that the slope measures the change in the mean of Y for a unit change in X, this qualitative result is unsurprising; it states that a higher midterm score yields a higher mean score on the final. How much higher? This is where the endpoints of the confidence interval come into play: if the score on the midterm increases by one point, then the mean score on the final increases by at least 0.1461 and at most 0.7571 points. Subjectively, I consider this interval to be very wide; perhaps not very useful.

Next, let's consider the intercept, β_0. Minitab presents the relevant output in the *Constant* row, with *Constant* falling under the heading *Predictor*; admittedly, the terminology is a bit confusing. Anyways, for expediency, I will stick to 95% for my confidence level; thus, t^* remains equal to 2.035. Therefore, the 95% confidence interval estimate of β_0, the y-intercept of the population regression line, is:

$$68.42 \pm 2.035(7.963) = 68.42 \pm 16.20 = [52.22, 84.62].$$

I conjecture that I know what you are thinking: Bob is going to interpret this interval. Well, I am not. Here is why.

Literally, the population intercept is the mean value of Y given $X = 0$. The smallest midterm score in the data set is $X = 39.0$; thus, we have no idea whether the linear relationship in the data set extends down to $X = 0$. (Sound familiar? We talked about this idea in the Fish Activity study

in Chapter 21.) Also, as a teacher, I am not particularly interested in predicting a final exam score for a student who scores 0 on my midterm! (Nobody has ever scored 0 on any of my tests!) In particular, if a student said,

> Yeah, I scored 0 on the midterm. The final is inconvenient for me, so why not give me my predicted score, $b_0 = 68.42$?

Even though the gift of 68.42 would not save the student from an F in the course, as a matter of principle I would never seriously consider such a request! (And I have no evidence any student would make this request; my students are serious about their educations.)

To summarize the above, if $X = 0$ is outside the range of the data and/or not of scientific interest, then, even though it is possible to calculate a confidence interval estimate of β_0, I don't do it.

Be careful in your reading of the previous paragraph. I am **not saying** that estimating β_0 is unimportant. It is **important** because it is part of the population regression line. What I am saying is that unless $X = 0$ is both in the range of the data **and** of scientific interest to the researcher, then β_0, **by itself**, it is not important. If this distinction is confusing, I hope that the next confidence interval formula will help.

22.3.1 Confidence Interval for the Mean Response for a Given Value of the Predictor

Let us consider a specific possible value of X, call it x_0. Now given that $X = x_0$, the mean of Y is $\beta_0 + \beta_1 x_0$; call this μ_0. We can use the computer output to obtain a point estimate and confidence interval estimate of μ_0.

For example, suppose we select $x_0 = 49.0$. Then, the point estimate is

$$b_0 + b_1(49.0) = 68.42 + 0.4516(49.0) = 90.548$$

If, however, you look at the output in Table 22.1, you will find this value, 90.548, in the column *Fit* and the row *Midterm* $= 49.0$. (This is observation 8, 9 or 10.) Of course, this is not a huge aid because, frankly, the above computation of 90.548 was pretty easy. But it's the next entry in the output that is useful. Just to the right of *Fit* is *SE(Fit),* where, as before, SE is the abbreviation for estimated standard error. Thus, we are able to calculate a confidence interval estimate of μ_0:

$$\text{Fit } \pm t^*(\text{SE(Fit)}). \tag{22.18}$$

For the current example, the 95% confidence interval estimate of the mean of Y given $X = 49.0$ is

$$90.548 \pm 2.035(0.968) = 90.548 \pm 1.970 = [88.578, 92.518].$$

The obvious question is: We were pretty lucky that $X = x_0 = 49.0$ was in our computer output; what do we do if our x_0 isn't there? It turns out that the answer is easy: trick the computer. Here is how.

Suppose we want to estimate the mean of Y given $X = 51.0$. An inspection of the *Midterm* column in the computer output in Table 22.1 reveals that there is no row for $X = 51.0$. Go back to the data set and add a 36th *student*. For this student, enter 51.0 for Midterm and a **missing value** for Final. (For Minitab, my software package, this means you enter a *.) Then rerun the regression analysis. In all of the computations the computer will ignore the 36th student because it has no value for Y and therefore cannot be used in all of the needed computations. Thus, it is not used in any computations. As a result, the computer output is unchanged by the addition of this extra student. But, and this is the key point, the computer output includes observation 36 in the last section of the output, creating the row:

```
Obs  Midterm  Final     Fit SE(Fit)  Residual
 36     51.0      *  91.451   0.828         *
```

From this we see that the point estimate of the mean of Y given $X = 51.0$ is 91.451. (This, of course, is easy to verify.) But now we also have the *SE(Fit)*, so we can obtain the 95% confidence interval estimate of the mean of Y given $X = 51.0$:

$$91.451 \pm 2.035(0.828) = 91.451 \pm 1.685 = [89.766, 93.136].$$

22.3.2 Prediction of the Response for a Given Value of the Predictor.

Suppose that beyond our n cases for which we have data, we have an additional case. For this new case, we know that $X = x_{n+1}$, for a known number x_{n+1}, and we want to predict the value of Y_{n+1}. Now, of course,

$$Y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \epsilon_{n+1}.$$

We assume that ϵ_{n+1} is independent of all previous ϵ's, and, like the previous errors, has mean 0 and variance σ^2.

The natural prediction of Y_{n+1} is obtained by replacing the β's by their estimates and ϵ_{n+1} by its mean, 0. The result is

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}.$$

We recognize this as the Fit for $X = x_{n+1}$; as such, its value and its SE are both presented in (or can be made to be presented in) our computer output.

Following our approach to prediction in Chapter 14, we compare the actual response to the predicted response via subtraction, giving us:

$$W = Y_{n+1} - \hat{y}_{n+1}.$$

Our Result 14.1 tells us that the estimated variance of W is

$$s^2 + [\text{SE(Fit)}]^2.$$

(It is ok if you don't bother with verifying that this result applies here; it's late in the semester! This variance, however, has a nice interpretation. The estimated variance of the prediction is the sum of two terms: The estimated variance of an observation around the population regression line:

s^2; and the estimated variance due to the population regression line being estimated at the point x_{n+1}: [SE(Fit)]^2.)

The prediction interval for Y_{n+1} is:

$$\text{Fit } \pm t^* \sqrt{s^2 + [\text{SE(Fit)}]^2}. \tag{22.19}$$

For example, suppose that $x_{n+1} = 49.0$. From our computer output, and our earlier work, the point prediction of y_{n+1} is 'Fit,' which is 90.548. The estimated variance of W is

$$(4.635)^2 + (0.968)^2 = 22.4202;$$

thus, the estimated standard error is $\sqrt{22.4202} = 4.735$. Thus, the 95% prediction interval for Y_{n+1} is

$$90.548 \pm 2.035(4.735) = 90.548 \pm 9.636 = [80.912, 100.184].$$

In words, for an additional student who scores 49.0 on the midterm, at the 95% probability level, we predict that this student's final exam score will be between, roughly, 81.0 and 100 points, inclusive. This is not a particularly useful prediction interval.

22.3.3 A Test of Hypotheses

The results of this short subsection are similar to the results we had in Section 18.1.

In many regression analyses, a key question is whether a regression is needed. In particular, researchers often want to test the null hypothesis that the slope of the population regression line, β_1, equals zero:

$$H_0 : \beta_1 = 0.$$

There are, as usual, three possible alternatives:

$$H_1 : \beta_1 > 0; H_1 : \beta_1 < 0; \text{ and } H_1 : \beta_1 \neq 0.$$

I recommend using the *Inconceivable Paradigm* to select the appropriate alternative.

The obvious starting point for the test statistic is the point estimator of β_1, denoted, when viewed as a random variable, by B_1, which has observed value b_1. As in our earlier work, the standardized version of B_1 is:

$$\frac{B_1 - \beta_1}{\sqrt{\text{Variance}(B_1)}}.$$

To obtain our test statistic we replace β_1 by its hypothesized value, 0, and estimate the variance in the denominator. The result is our test statistic, denoted by T with observed value t:

$$T = \frac{B_1}{\text{SE}(B_1)} \text{ and}$$

$$t = \frac{b_1}{\text{SE}(b_1)}. \tag{22.20}$$

In the formulas below, t is given above (Equation 22.20) and areas are computed under the t-curve with $df = (n - 2)$.

1. For the alternative $>$, the P-value equals the area to the right of t.

2. For the alternative $<$, the approximate P-value equals the area to the left of t.

3. For the alternative $\neq$, the approximate P-value equals twice the area to the right of $|t|$.

The Minitab output *anticipates* that you will want to do this test. Let's look at part of the output again for the exam scores' data:

Predictor	Coef	SE(Coef)	T	P
Constant	68.420	7.963	8.59	0.000
Midterm	0.4516	0.1501	3.01	0.005

The observed value of the test statistic

$$t = 0.4516/0.1501 = 3.01,$$

is given in the fourth column (headed 'T'). The P-value for the alternative $\neq$ is given in the fifth column (headed 'P') and is equal to 0.005. Thus, for the alternative $>$, the P-value equals $0.005/2 = 0.0025$. More precisely, using our t-curve website, the area under the t-curve with $df = 33$ to the right of 3.01 is 0.00249.

Often in regression, the researcher has a special possible value of interest for the population slope, denoted by β_{10} and read as *beta-one-zero* or *beta-one-naught*, but never as *beta-ten*. In this situation the null hypothesis becomes:

$$H_0 : \beta_1 = \beta_{10}.$$

The three possible alternatives are:

$$H_1 : \beta_1 > \beta_{10}; H_1 : \beta_1 < \beta_{10}; \text{ and } \beta_1 \neq \beta_{10}.$$

(Obviously, if $\beta_{10} = 0$ this new problem reduces to the problem we solved above.)

For this more general situation, the test statistic is

$$T = \frac{B_1 - \beta_{10}}{\text{SE}(B_1)}$$

with observed value

$$t = \frac{b_1 - \beta_{10}}{\text{SE}(b_1)}. \tag{22.21}$$

The earlier rules for the P-value apply to this new situation.

For this course—and our final exam—the above is the only test you are required to learn. Let me mention in passing that a researcher could adapt the above method to test a null hypothesis on the value of the intercept, β_0, or on the mean value of the response for a given value of the predictor. Indeed, the Minitab output gives the observed value of the test statistic and the two-sided P-value for the test of the null hypothesis that the intercept equals 0.

22.4 Extensions

The simple linear regression model begins with Equation 22.10, reproduced below:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

for $i = 1, 2, 3, \ldots n$. The model includes the assumptions about the sequence of ϵ's given earlier; namely, they are independent, mean equal to zero and variance equal to the unknown σ^2.

The word *simple* signifies that there is only one predictor. Also—although this fact often causes confusion—*linear* refers to the model being linear in the parameters β_0 and β_1 and **not** to the fact that the mean of Y is a linear function of X. This leads to the first generalization of our model.

Suppose that for some reason, you believe that the relationship between a nonnegative response Y and a nonnegative predictor X is

$$Y_i = \beta_0 + \beta_1 X_i^2 + \epsilon_i. \tag{22.22}$$

Literally, this is not the same as the simple linear regression model, but it can be **transformed** into the simple linear regression model quite easily. All we do is define a new predictor X_i^* to equal X_i^2. With this substitution, Equation 22.22 becomes

$$Y_i = \beta_0 + \beta_1 X_i^* + \epsilon_i,$$

which **is** the simple linear regression model with predictor denoted by X_i^*.

In Chapter 13 you learned about the family of Poisson distributions. Recall that for a Poisson distribution the mean equals the variance. In many applications to science the variance of the response will increase with the mean, although they won't necessarily be equal. In particular, consider the following modification of the simple linear regression model for a predictor that must be positive.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{22.23}$$

In this equation, conditional on the value of X_i, the error term, ϵ_i, has mean 0 and variance $\sigma^2 X_i^2$. Thus, this is **not** the simple linear regression model because the error terms do **not** have constant variance. How can we fix this? The answer is quite simple. Divide both sides of Equation 22.23 by X_i (remember, it must be positive), to get:

$$Y_i/X_i = \beta_0/X_i + \beta_1 + \epsilon_i/X_i.$$

Next, make the following definitions:

$$Y_i^* = Y_i/X_i; X_i^* = 1/X_i; \beta_0^* = \beta_1; \beta_1^* = \beta_0; \text{ and } \epsilon_i^* = \epsilon_i/X_i.$$

Thus, Equation 22.23 becomes

$$Y_i^* = \beta_0^* + \beta_1^* X_i^* + \epsilon_i^*,$$

the simple linear regression model for response Y_i^* and predictor X_i^* because the errors ϵ_i^* are independent, mean 0 with constant variance σ^2.

Sometimes, researchers assume a multiplicative error rather than the additive error in the simple linear regression model. In particular, define $\epsilon_i^* = \exp(\epsilon_i)$, where the ϵ_i's satisfy the assumptions of the simple linear regression model. Consider the new model:

$$Y_i = \exp(\beta_0 + \beta_1 X_i) \times \epsilon_i^*. \tag{22.24}$$

If we take the natural logarithm ($\ln$) of both sides of this equation, we get:

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i,$$

which is the simple linear regression model for response $Y_i^* = \ln(Y_i)$.

Finally, with the definition of ϵ_i^* in the previous paragraph, consider the model with $\beta_0 > 0$ and the predictor and response both constrained to be positive:

$$Y_i = \beta_0 X_i^{\beta_1} \times \epsilon_i^*. \tag{22.25}$$

If we take the natural logarithm ($\ln$) of both sides of this equation, we get:

$$\ln(Y_i) = \ln(\beta_0) + \beta_1 \ln(X_i) + \epsilon_i,$$

which is the simple linear regression model for:

$$Y_i^* = \ln(Y_i); X_i^* = \ln(X_i); \beta_0^* = \ln(\beta_0); \text{ and } \beta_1^* = \beta_1.$$

Let me end with two comments about the above list of examples.

1. Rather obviously, the list above is not an exhaustive list of models that can be easily transformed to the simple linear regression model.

2. Each example began with an equation relating the original response to the original predictor. A scientist typically *obtains* such an equation in one of two ways:

   - **Empirically:** By looking at a scatterplot of the data.
   - **Theoretically:** Some scientific theory leads to the belief that the relationship between Y and X should have the form given in the equation.

   Note that for either of these methods, not only should you focus on how Y varies with X, but on how the error terms enter the relationship: additive or multiplicative; constant or nonconstant variance.

22.5 Summary

In a regression analysis, each case has four numbers associated with it:

$$x, y, \hat{y} = b_0 + b_1 x \text{ and } e = y - \hat{y}.$$

In words, its predictor, response, predicted response and residual, respectively.

The researcher should draw a picture—dot plot or histogram—of the residuals. Note that a case's residual being an outlier is a different notion than a case being isolated, although the former implies the latter.

For any regression analysis, the mean of the residuals equals zero: $\bar{e} = 0$. The sum of squared residuals is denoted by SSE:

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

The residuals have $(n - 2)$ degrees of freedom and the variance of the residuals is

$$s^2 = \frac{\text{SSE}}{n - 2}.$$

The standard deviation of the residuals is, of course, $s = \sqrt{s^2}$.

For each individual case, the residual tells us *how well* the regression line performed, in the following sense.

- If the residual equals 0, then $y = \hat{y}$ and the regression line's prediction is perfect.
- If the residual is greater than 0, then $y > \hat{y}$ and the regression line's prediction is too small.
- If the residual is less than 0, then $y < \hat{y}$ and the regression line's prediction is too large.
- The farther the value of the residual is from 0, in either direction (positive or negative), the worse the regression line predicts the actual response.
- In short, a residual that is *close to zero*, positive or negative, indicates that the regression line did a *good job* predicting the response. A residual that is *far from zero*, positive or negative, indicates that the regression line did a *bad job* predicting the response. The distinction between being *close to zero* or *far from zero* should be based on the scientific goals of the study.

The bulleted items above are concerned with evaluating the regression line for individual cases. We also want an overall evaluation of the quality of the regression line.

If a scientist has a specific number that distinguishes between *close to zero* and *far from zero*, then the regression line can be evaluated with simple counting: count the number of cases for which the prediction is good and compare it to the number of cases for which the prediction is bad. I did this earlier in the chapter where I (subjectively) decided that the boundary between a good and bad prediction of a batting average was 20 points (0.020). I found that with this boundary, (only) 57% of the cases were predicted well.

Table 22.3: The ANOVA table for simple linear regression.

Source	DF	SS
Regression	1	SSR
Residual Error	$n-2$	SSE
Total	$n-1$	SSTO

However useful it is for a scientist in a particular study to specify the boundary between *good* and *bad* predictions, this activity does **not** lend itself to mathematical analysis of a general nature. Instead, we focus on the value of s, the standard deviation of the residuals, and use the Empirical Rule for interpreting s to measure the effectiveness of the regression line. For example, according to the Empirical Rule approximately 68% of the residuals will be between $-s$ and s; in other words, for approximately 68% of the cases, the actual response will be within s of the predicted response.

I recommend comparing the standard deviation of the residuals, s, to the standard deviation of the responses, s_2. This amounts to comparing the best predictions using X with the best predictions that ignore X.

Mostly in work beyond this chapter, the Analysis of Variance Table is useful. It is presented in Table 22.3. The various sum of squares are defined by:

$$\text{SSR } = \sum(\hat{y}_i - \bar{y})^2; \text{ SSE } = \sum(y_i - \hat{y}_i)^2; \text{ and SSTO } = \sum(y_i - \bar{y})^2.$$

Note that in this table, both the degrees of freedom and sums of squares *sum*, in the sense that:

$$1 + (n-2) = (n-1); \text{ and SSR } + \text{ SSE } = \text{ SSTO}.$$

The coefficient of determination, R^2 is a popular summary statistic for a regression analysis:

$$R^2 = \frac{\text{SSTO } - \text{ SSE}}{\text{SSTO}}.$$

The denominator of R^2 equals the total squared error in the responses. The numerator of R^2 equals the amount of squared error that is removed by using X to predict Y via the regression line. Thus, the ratio of R^2 equals the proportion (usually reported as a percentage) of the total squared error in the response that can be explained by a linear relationship with the predictor.

For every regression analysis, R^2 equals the square of the correlation coefficient:

$$R^2 = r^2.$$

This identity gives us another interpretation of the correlation coefficient, r.

For inference, we assume that the simple linear regression model is true:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ for } i = 1, 2, 3, \ldots n.$$

In this equation, β_0 and β_1 are parameters whose values are known to Nature, but unknown to the researcher. The error terms,

$$\epsilon_1, \epsilon_2, \epsilon_3, \ldots, \epsilon_n,$$

are assumed to be independent random variables, each with mean equal to 0 and variance equal to σ^2, the third unknown parameter of the model. Statistical inference for this model conditions on the values of the predictor.

A consequence of this model is that the mean value of Y for a given value of $X = x$ is equal to:

$$\beta_0 + \beta_1 x.$$

The point estimates of β_1 and β_0 are obtained by applying the Principle of Least Squares, as we did in Chapter 21, yielding

$$b_1 = r(s_2/s_1); \text{ and } b_0 = \bar{y} - b_1\bar{x}, \text{ respectively.}$$

The point estimate of σ^2 is the variance of the residuals, $s^2 = \text{SSE}/(n-2)$.

I do not give you algebraic formulas for confidence intervals, prediction intervals and testing. Instead, I show you how to obtain answers—intervals and P-values—from Minitab computer output. In particular, confidence intervals for β_1 and β_0 are given in Result 22.2 on page 601.

For a specified value of the predictor, call it x_0, the mean value of the response is

$$\mu_0 = \beta_0 + \beta_1 x_0.$$

The point estimate of μ_0 is

$$b_0 + b_1 x_0,$$

which Minitab denotes as *Fit*. The estimated standard error of the *Fit* is denoted by SE(Fit). The confidence interval for μ_0 is

$$\text{Fit } \pm t^*(\text{SE(Fit)}).$$

Suppose that beyond our n cases for which we have data, we have an additional case. For this new case, we know that $X = x_{n+1}$, for a known number x_{n+1}, and we want to predict the value of Y_{n+1}. The point prediction is:

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}.$$

We recognize this as the Fit for $X = x_{n+1}$; as such, its value and SE(Fit) are both presented in (or can be made to be presented in) our computer output. The prediction interval for Y_{n+1} is:

$$\text{Fit } \pm t^* \sqrt{s^2 + [\text{SE(Fit)}]^2}.$$

Tests of hypotheses also are possible for the simple linear regression model. Sometimes the researcher has a special possible value of interest for the population slope, denoted by β_{10}. The null hypothesis is:

$$H_0 : \beta_1 = \beta_{10}.$$

The three possible alternatives are:

$$H_1 : \beta_1 > \beta_{10}; H_1 : \beta_1 < \beta_{10}; \text{ and } \beta_1 \neq \beta_{10}.$$

The test statistic is

$$T = \frac{B_1 - \beta_{10}}{\text{SE}(B_1)}$$

with observed value

$$t = \frac{b_1 - \beta_{10}}{\text{SE}(b_1)}.$$

The rules for obtaining the P-value are given on page 605.

Finally, Section 22.4 presents several models that can be transformed easily into the simple linear regression model.

22.6 Practice Problems

1. A simple linear regression analysis with $n = 5$ yields the numbers in the following table.

x:	-2	-1	0	1	2
e:	1	-2	$+2$	b	c

   Determine the values of b and c. (Hint: Use Equation 22.1.)

2. A simple linear regression analysis with $n = 10$ yields the following (partial) ANOVA table.

Source	DF	SS
Regression	a	496
Residual Error	b	800
Total	c	d

   (a) Determine the values of a–d in the ANOVA table.

   (b) Calculate the values of s and s_2. Explain what you have found.

   (c) Calculate the value of R^2; interpret the number you obtain.

   (d) What can you say about the value of the correlation coefficient, r?

3. A simple linear regression analysis yields the following (partial) ANOVA table.

Source	DF	SS
Regression	a	b
Residual Error	20	c
Total	d	2000

   In addition, $R^2 = 0.800$.

   (a) Determine the values of a–d in the ANOVA table.

   (b) Calculate the values of s and s_2. Explain what you have found.

4. Table 22.4 presents edited Minitab regression output for the exam scores data for all $n = 36$ students; i.e., it includes the isolated case $(35.5, 95.5)$.

   (a) Calculate the 95% confidence interval estimate of the slope of the regression line. Compare your answer to the answer earlier in this chapter for $n = 35$ students and comment.

   (b) Calculate the P-value for the alternative $\beta_1 > 0$. Compare your answer to the answer earlier in this chapter for $n = 35$ students and comment.

   (c) Calculate the P-value for the alternative $\beta_1 < 0.75$.

(d) Calculate the 95% confidence interval estimate of the mean response for $X = 49.0$. Compare your answer to the answer earlier in this chapter for $n = 35$ students and comment.

(e) Calculate the 95% prediction interval for a future response for $X = 49.0$. Compare your answer to the answer earlier in this chapter for $n = 35$ students and comment.

(f) Determine the ANOVA table for this analysis.

5. Table 22.5 presents edited Minitab regression output for the batting averages data for $n = 123$ players, after Floyd Rayford has been deleted from the data set. Note also that for arithmetic convenience (for me!) I have multiplied all batting averages by 1000 that, for example, converts $x = 0.269$ to $x = 269$.

(a) Calculate the 95% confidence interval estimate of the slope of the regression line.

(b) Calculate the P-value for the alternative $\beta_1 > 0$.

(c) Calculate the P-value for the alternative $\beta_1 < 1$.

(d) Calculate the 95% confidence interval estimate of the mean response for $X = 309$.

(e) Calculate the 95% prediction interval for a future response for $X = 309$.

Table 22.4: Edited Minitab output for the regression of final exam score on midterm exam score for 36 students.

```
The regression equation is: Final = 76.5 + 0.302 Midterm

Predictor        Coef    SE(Coef)          T        P
Constant       76.536       7.239      10.57    0.000
Midterm        0.3023      0.1375       2.20    0.035
S = 4.850       R-Sq = 12.4%

Obs   Midterm   Final       Fit  SE(Fit)  Residual
  1      39.0    83.5    88.325    2.000    -4.825
  2      43.0    89.0    89.534    1.514    -0.534
  3      44.0    92.0    89.837    1.399     2.163
  4      44.5    82.0    89.988    1.344    -7.988
  5      46.0    93.0    90.441    1.186     2.559
  6      48.0    87.0    91.046    1.002    -4.046
  7      48.5    91.5    91.197    0.963     0.303
 10      49.0    81.0    91.348    0.927   -10.348
 12      49.5    95.0    91.499    0.896     3.501
 13      50.0    97.5    91.650    0.868     5.850
 15      53.0    93.0    92.557    0.814     0.443
 17      53.5    99.0    92.708    0.825     6.292
 19      54.5    83.0    93.011    0.863   -10.011
 20      55.0    94.5    93.162    0.889     1.338
 23      55.5    92.5    93.313    0.920    -0.813
 26      56.5    90.5    93.615    0.993    -3.115
 27      57.0    92.0    93.766    1.035    -1.766
 28      57.5    97.5    93.917    1.079     3.583
 32      58.5    97.5    94.220    1.174     3.280
 35      59.0    94.0    94.371    1.225    -0.371
 36      35.5    95.5    87.267    2.449     8.233
```

Table 22.5: Edited Minitab output for the regression of 1986 batting average multiplied by 1000 on 1985 batting average multiplied by 1000 for 123 baseball players. Note the Floyd Rayford has been deleted from the data set.

The regression equation is: 1986BA = 83.0 + 0.681 1985BA

Predictor	Coef	SE(Coef)	T	P
Constant	83.03	21.55	3.85	0.000
1985BA	0.68114	0.08055	8.46	0.000

S = 24.82 R-Sq = 37.1%

Analysis of Variance

Source	DF	SS
Regression	1	44060
Residual Error	121	74562
Total	122	118621

Obs	1985BA	1986BA	Fit	SE(Fit)	Residual
1	265	264	263.53	2.24	0.47
2	309	296	293.50	4.12	2.50
3	268	240	265.57	2.24	-25.57
4	243	229	248.54	2.91	-19.54
5	289	289	279.88	2.90	9.12
6	266	286	264.21	2.24	21.79
7	231	238	240.37	3.61	-2.37
8	275	309	270.34	2.35	38.66
9	304	300	290.09	3.78	9.91

22.7 Solutions to Practice Problems

1. Let's expand the given table:

x:	-2	-1	0	1	2	Total
e:	1	-2	2	b	c	0
xe:	-2	2	0	b	$2c$	0

   Thus,

$$1 + b + c = 0 \text{ and } b + 2c = 0.$$

   Rewrite the first of these equations as $b = -c - 1$ and substitute it into the second equation:

$$-c - 1 + 2c = 0 \text{ or } c = 1; \text{ which yields } 1 + b + 1 = 0 \text{ or } b = -2.$$

2. (a) We know that a always equals 1; $b = (n-2) = (10-2) = 8$; $c = a + b = 1 + 8 = 9$ or $c = (n-1) = (10-1) = 9$; and $d = 496 + 800 = 1296$.

   (b) First, $s^2 = 800/b = 800/8 = 100$; thus, $s = 10$. Next, $s_2^2 = 1296/9 = 144$; thus, $s_2 = 12$. Finally, $s/s_2 = 10/12 = 0.833$; thus, the standard deviation of prediction errors using the regression line is 16.7% smaller than the standard deviation of prediction errors not using the regression line.

   (c) $R^2 = 496/1296 = 0.383$. In words, 38.3% of the squared error in the response is explained by a linear relationship with the predictor.

   (d) The correlation coefficient $r = \pm\sqrt{0.383} = \pm 0.619$.

3. (a) We know that $a = 1$, thus $d = 1 + 20 = 21$.
   We know that $R^2 = 0.800$; thus:

$$0.800 = \text{SSR}/\text{SSTO} = \text{SSR}/2000; \text{ which gives SSR} = b = 1600.$$

   By subtraction, $c = 2000 - 1600 = 400$.

   (b) We know that

$$s^2 = 400/20 = 20; \text{ and } s_2^2 = 2000/21 = 95.24.$$

   Thus, $s = 4.472$, $s_2 = 9.759$ and $s/s_2 = 4.472/9.759 = 0.458$. Thus, the standard deviation of predictions using X is 54.2% smaller than the standard deviation of predictions without using X.

4. First, note that in order to obtain 95% confidence (for the estimates) or probability (for prediction), we have $t^* = 2.033$ for $df = (n-2) = (36-2) = 34$.

   (a) Using the computer output, the 95% confidence interval estimate of β_1 is:

$$0.3023 \pm 2.033(0.1375) = 0.3023 \pm 0.2795 = [0.0228, 0.5818].$$

   This interval is narrower than the earlier interval (half-width is 0.2795 versus 0.3055), but is much closer to zero (center is 0.3023 versus 0.4516).

(b) The easiest way to obtain the P-value is to realize that it equals one-half of the two-sided P-value given by Minitab: $0.035/2 = 0.018$.

Alternatively, the observed value of the test statistic is

$$t = 0.3023/0.1375 = 2.1985.$$

With the help of the website the area under the t-curve with $df = 34$ to the right of 2.1985 is 0.0174. (Paraphrasing Shakespeare, "Much ado about not much.")

For the earlier analysis with $n = 35$, the P-value is much smaller, 0.0025.

(c) The observed value of the test statistic is

$$t = \frac{0.3023 - 0.75}{0.1375} = -0.4477/0.1375 = -3.256.$$

With the help of the website the area under the t-curve with $df = 34$ to the left of -3.256 is 0.0013.

(d) Using the computer output for observation 10, the 95% confidence interval estimate the mean of Y given $X = 49$ is

$$91.348 \pm 2.033(0.927) = 91.348 \pm 1.885 = [89.463, 93.233].$$

This point prediction is almost one point larger ($91.348 - 90.548 = 0.800$) than the earlier point prediction and this interval is narrower than the earlier interval (half-width is 1.885 versus 1.970).

(e) Again using the computer output for observation 10, we find that the estimated variance of the predicted value is:

$$(4.850)^2 + (0.927)^2 = 24.3818.$$

Thus, the estimated standard error of the predicted value is:

$$\sqrt{24.3818} = 4.938.$$

Thus, the 95% prediction interval is:

$$91.348 \pm 2.033(4.938) = 91.348 \pm 10.039 = [81.309, 101.387].$$

This interval is slightly narrower than the earlier one, but both intervals have little practical value because they are so wide. A score of 81.5 on the final is very different than a score of 100.

(f) First, the easy part of the ANOVA table:

Source	DF	SS
Regression	1	
Residual Error	34	
Total	35	

Next,

$$(4.85)^2 = 23.5225 = s^2 = \text{ SSE}/34; \text{ or SSE } = 34(23.5225) = 799.765.$$

Next,

$$0.124 = R^2 = 1 - (799.765/\text{SSTO}) \text{ or SSTO } = 799.765/0.876 = 912.974.$$

Thus, the ANOVA table is:

Source	DF	SS
Regression	1	113.209
Residual Error	34	799.765
Total	35	912.974

5. First note that for $df = 123 - 2 = 121$, for 95% confidence or probability, $t^* = 1.980$.

   (a) The 95% confidence interval estimate of the slope is:

   $$0.6811 \pm 1.980(0.08055) = 0.6811 \pm 0.1595 = [0.5216, 0.8406].$$

   (b) The P-value equals one-half of the value in the table, which is one-half of 0.000, or 0.000. More precisely, using Minitab for $t = 8.46$ I obtain 3.62×10^{-14}. This is a really small P-value!

   (c) The observed value of the test statistic is

   $$t = \frac{0.6811 - 1}{0.08055} = -3.959.$$

   The area under the t-curve with $df = 121$ to the left of -3.959 is equal to—with the help of Minitab—0.0000368; or approximately 37 in one million. This is a very small P-value.

   (d) The 95% confidence interval estimate of the mean response given $X = 309$ is

   $$293.50 \pm 1.980(4.12) = 293.50 \pm 8.16 = [285.34, 301.66].$$

   (e) Again using the computer output for observation 2, we find that the estimated variance of the predicted value is:

   $$(24.82)^2 + (4.12)^2 = 633.0068.$$

   Thus, the estimated standard error of the predicted value is:

   $$\sqrt{633.0068} = 25.16.$$

   Thus, the 95% prediction interval is:

   $$293.50 \pm 1.980(25.16) = 293.50 \pm 49.82 = [243.68, 343.32].$$

   This interval is very wide. As a baseball fan, I consider it to be almost totally worthless; 343 is a great batting average and 243 is—while not horrible—pretty poor.

22.8 Homework Problems for Chapter 21

1. (Hypothetical data.) Fifty students in a Statistics class take midterm and final exams. Below are selected summary statistics for these data.

Exam	Mean	Stand. Dev.
Midterm	50.00	10.00
Final	70.00	15.00

   Also, the correlation coefficient of the the two exam scores is $r = 0.48$.

   (a) Determine the equation of the regression line for using the score on the midterm exam to predict the score on the final exam.

   (b) Determine the equation of the regression line for using the score on the final exam to predict the score on the midterm exam.

   (c) Sally scores 60 on the midterm exam. Use your equation from (a) to obtain her predicted score on the final exam.

   (d) Tom scores 80 on the final exam. Use your equation from (b) to obtain his predicted score on the midterm exam.

   (e) Refer to (c). Given that Sally actually scored 82 on the final exam, calculate her residual.

   (f) Refer to (d). Given that Tom actually scored 47 on the midterm exam, calculate his residual.

2. (Hypothetical data.) We have two measurements, x and y, on each of 500 children. We use these data to obtain the regression line for using x to predict y.

   The means for the 500 children are: 110 for x and 190 for y.

   Ron's value for x is 10 less than the mean of the x values. In addition, Ron's y is 20 less than predicted from his x.

   Given that Ron's $y = 140$, calculate the regression line.

22.9 Homework Problems for Chapter 22

Below is edited output from a study on the heights and weights of 10 members of the WNBA Chicago Sky team. For the purpose of the following questions, we will view these 10 women as a random sample from the population of all female professional basketball players.

```
The regression equation is
Weight = - 304 + 6.57 Height

Predictor      Coef   SE Coef
Height         6.57    0.9256

S = 12.36       R-Sq = 86.3%

  Ht       Wt      Fit  SE Fit
75.0   184.00   188.47    4.59
69.0   162.00   149.08    5.02
74.0   162.00   181.91    4.18
78.0   200.00   208.17    6.49
80.0   240.00   221.30    8.05
```

1. Briefly explain (this means in words) the meaning of the regression equation. Make sure you interpret the number 6.57.
2. Bert looks at the regression equation and states, "This is ridiculous! A woman cannot have a negative weight!"
   (a) Can you guess why Bert made this statement? If yes, explain.
   (b) Do you agree or disagree with Bert? Explain your answer.
3. Calculate the 95% confidence interval for the slope of the simple linear regression model.
4. Calculate the 95% confidence interval for the mean weight of women in the population who are 69 inches tall.
5. Calculate the 95% confidence interval for the mean weight of women in the population who are 78 inches tall.
6. I select a woman at random from the population and note that she is 74 inches tall. Given this information, obtain the 95% prediction interval for her weight.
7. I select a woman at random from the population and note that she is 80 inches tall. Given this information, obtain the 95% prediction interval for her weight.
8. Calculate the value of the correlation coefficient r.
9. (Tricky.) Calculate the slope of the regression line for using weight to predict height.