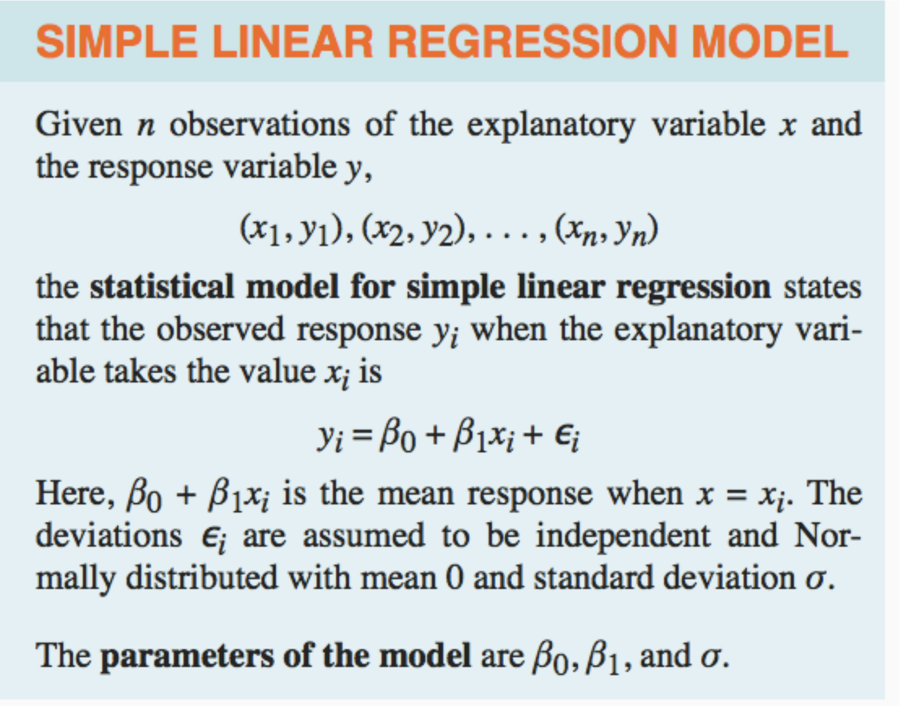
10.1 Simple Linear Regression



SIMPLE LINEAR REGRESSION MODEL

Given n observations of the explanatory variable x and the response variable y,

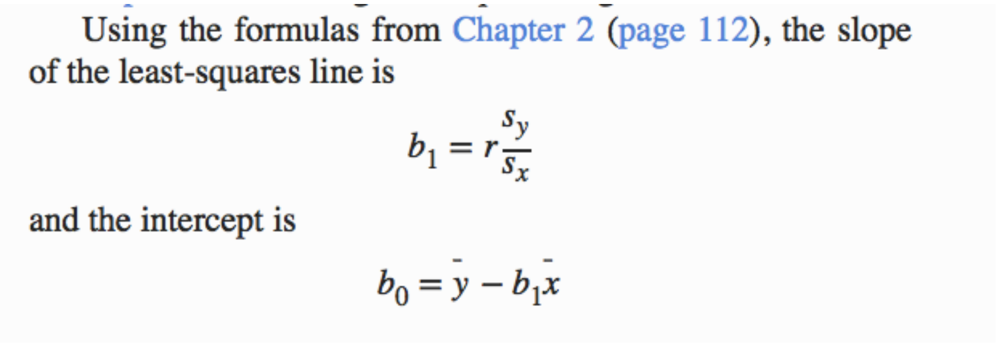
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

the **statistical model for simple linear regression** states that the observed response y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Here, $\beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$. The deviations ϵ_i are assumed to be independent and Normally distributed with mean 0 and standard deviation σ.

The **parameters of the model** are β_0, β_1, and σ.

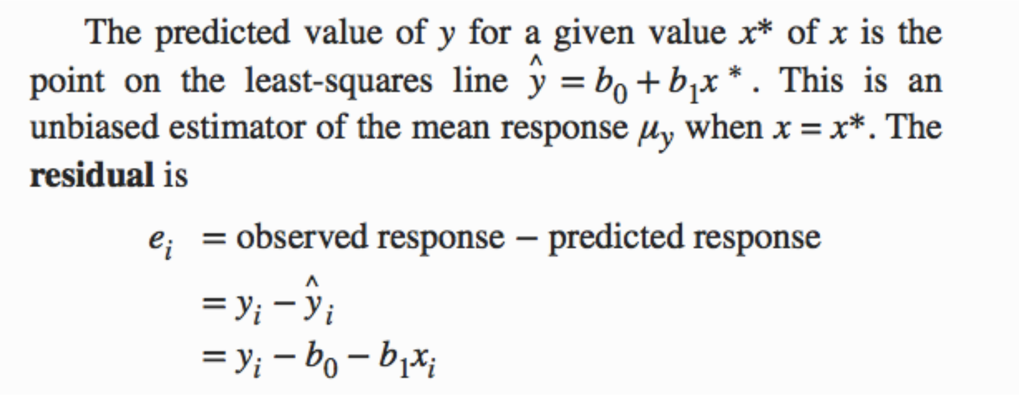


Using the formulas from Chapter 2 (page 112), the slope of the least-squares line is

$$b_1 = r\frac{s_y}{s_x}$$

and the intercept is

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



The predicted value of y for a given value x^* of x is the point on the least-squares line $\hat{y} = b_0 + b_1 x^*$. This is an unbiased estimator of the mean response μ_y when $x = x^*$. The **residual** is

$$\begin{aligned} e_i &= \text{observed response} - \text{predicted response} \\ &= y_i - \hat{y}_i \\ &= y_i - b_0 - b_1 x_i \end{aligned}$$

The residuals e_i correspond to the linear regression model deviations ϵ_i. The e_i sum to 0, and the ϵ_i come from a population with mean 0. Because we do not observe the ϵ_i, we use the residuals to check the model assumptions of the ϵ_i.

sample variance, p. 38

$$\begin{aligned} s^2 &= \frac{\sum e_i^2}{n-2} \\ &= \frac{\sum (y_i - \hat{y}_i)^2}{n-2} \end{aligned}$$

model standard deviation σ

We average by dividing the sum by $n - 2$ in order to make s^2 an unbiased estimate of σ^2 (the sample variance of n observations uses the divisor $n - 1$ for this same reason). The quantity $n - 2$ is called the degrees of freedom for s^2. The estimate of the **model standard deviation σ** is given by

$$s = \sqrt{s^2}$$

LINEAR REGRESSION MODEL CONDITIONS

To use the least-squares line as a basis for inference about a population, each of the following conditions should be approximately met:

- The sample is an SRS from the population.
- There a linear relationship between x and y.
- The standard deviation of the responses y about the population regression line is the same for all x.
- The model deviations are Normally distributed.

Confidence intervals and significance tests

Chapter 7 presented confidence intervals and significance tests for means and differences in means. In each case, inference rested on the standard errors of estimates and on t distributions. Inference in simple linear regression is similar in principle. For example, the confidence intervals have the form

$$\text{estimate} \pm t^*\text{SE}_{\text{estimate}}$$

where t^* is a critical point of a t distribution. The formulas for the estimate and standard error, however, are more complicated.

CONFIDENCE INTERVAL AND SIGNIFICANCE TEST FOR THE REGRESSION SLOPE

A **level C confidence interval for the slope** β_1 is

$$b_1 \pm t^*\text{SE}_{\text{b}_1}$$

In this expression, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

To test the hypothesis H_0: $\beta_1 = 0$, compute the **test statistic**

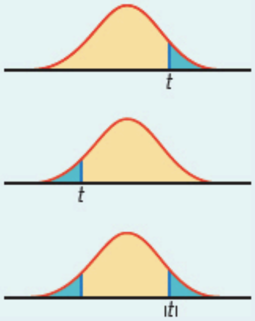
$$t = \frac{b_1}{\text{SE}_{b_1}}$$

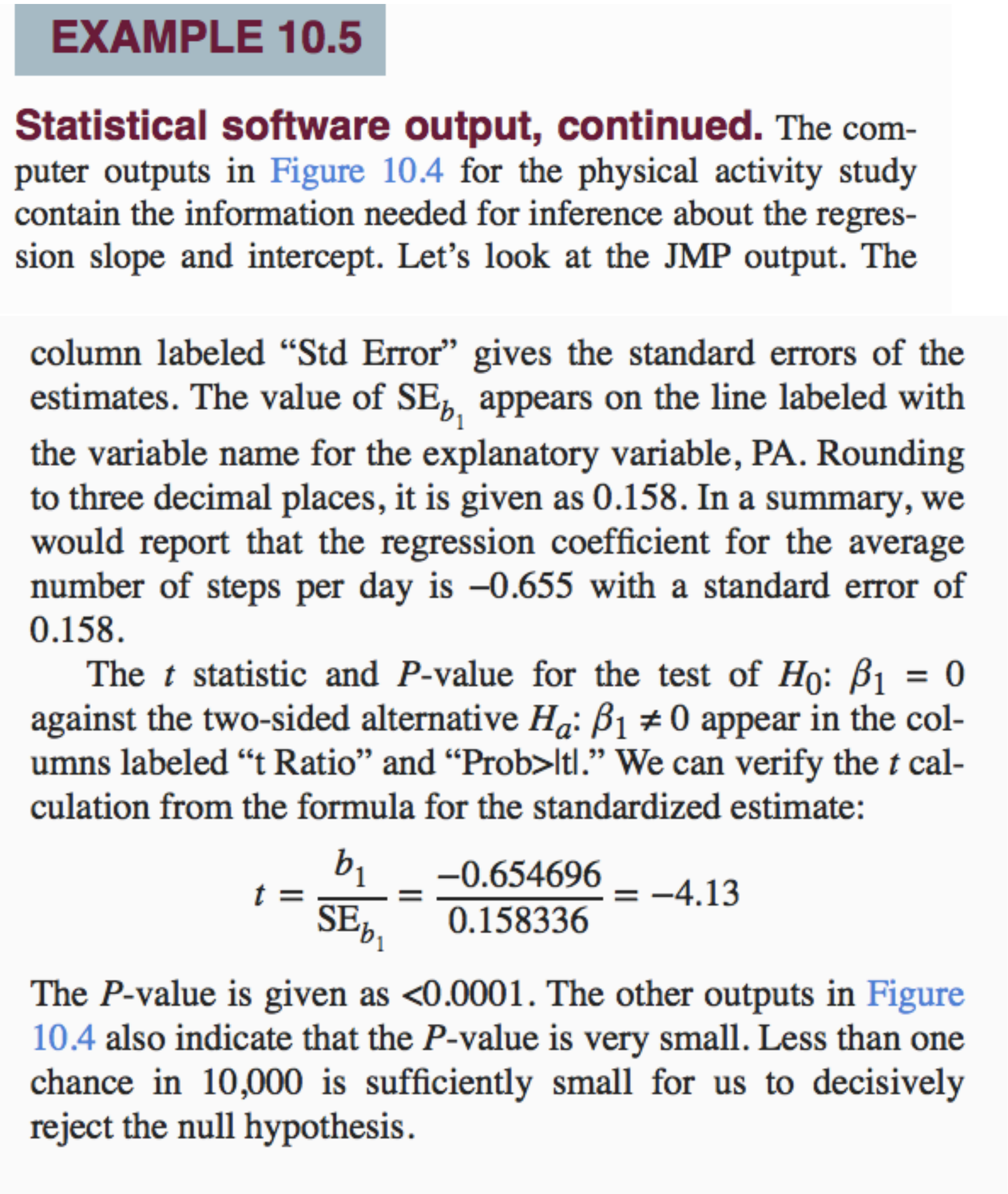
The **degrees of freedom** are $n - 2$. In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against

H_a: $\beta_1 > 0$ is $P(T \geq t)$

H_a: $\beta_1 < 0$ is $P(T \leq t)$

H_a: $\beta_1 \neq 0$ is $2P(T \geq |t|)$





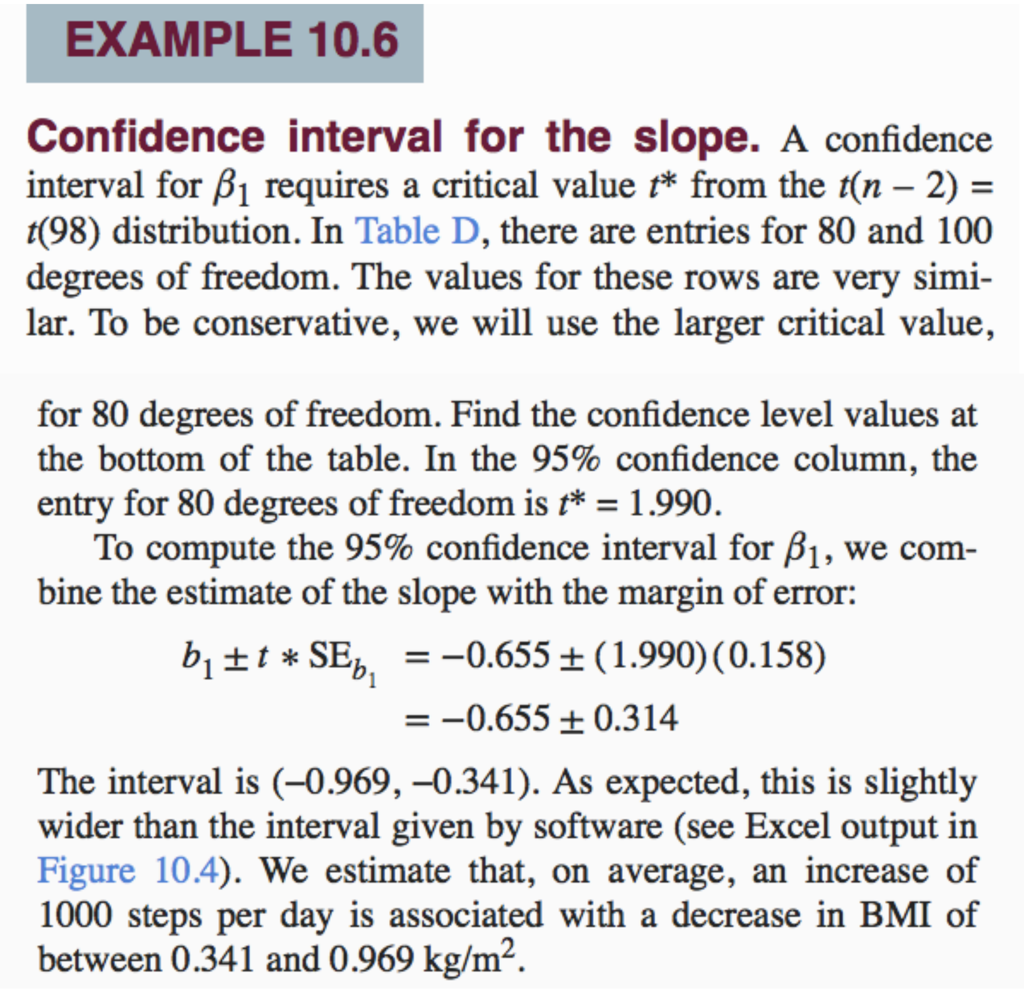
EXAMPLE 10.5

Statistical software output, continued. The computer outputs in Figure 10.4 for the physical activity study contain the information needed for inference about the regression slope and intercept. Let's look at the JMP output. The column labeled "Std Error" gives the standard errors of the estimates. The value of SE_{b_1} appears on the line labeled with the variable name for the explanatory variable, PA. Rounding to three decimal places, it is given as 0.158. In a summary, we would report that the regression coefficient for the average number of steps per day is –0.655 with a standard error of 0.158.

The t statistic and P-value for the test of H_0: $\beta_1 = 0$ against the two-sided alternative H_a: $\beta_1 \neq 0$ appear in the columns labeled "t Ratio" and "Prob>|t|." We can verify the t calculation from the formula for the standardized estimate:

$$t = \frac{b_1}{\text{SE}_{b_1}} = \frac{-0.654696}{0.158336} = -4.13$$

The P-value is given as <0.0001. The other outputs in Figure 10.4 also indicate that the P-value is very small. Less than one chance in 10,000 is sufficiently small for us to decisively reject the null hypothesis.



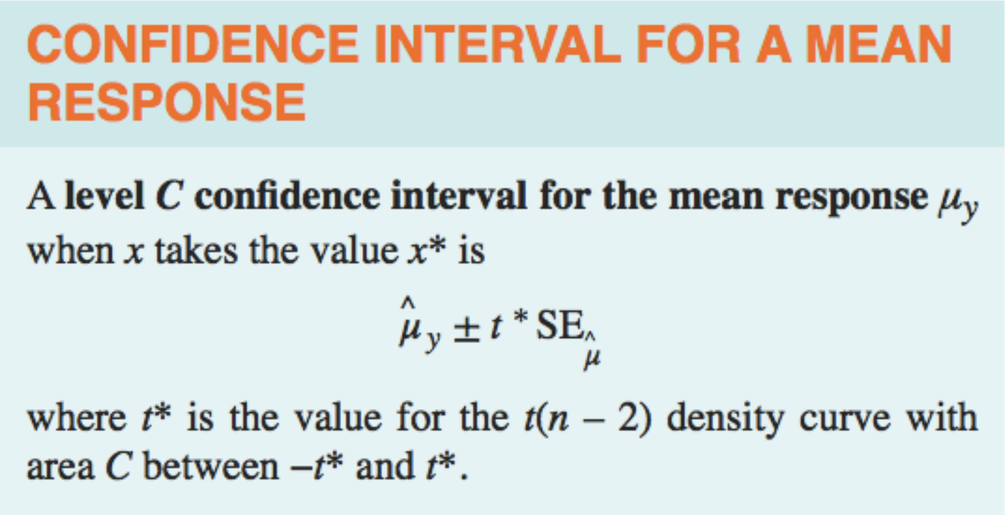
EXAMPLE 10.6

Confidence interval for the slope. A confidence interval for β_1 requires a critical value t^* from the $t(n - 2) = t(98)$ distribution. In Table D, there are entries for 80 and 100 degrees of freedom. The values for these rows are very similar. To be conservative, we will use the larger critical value, for 80 degrees of freedom. Find the confidence level values at the bottom of the table. In the 95% confidence column, the entry for 80 degrees of freedom is $t^* = 1.990$.

To compute the 95% confidence interval for β_1, we combine the estimate of the slope with the margin of error:

$$\begin{aligned} b_1 \pm t * \mathrm{SE}_{b_1} &= -0.655 \pm (1.990)(0.158) \\ &= -0.655 \pm 0.314 \end{aligned}$$

The interval is (–0.969, –0.341). As expected, this is slightly wider than the interval given by software (see Excel output in Figure 10.4). We estimate that, on average, an increase of 1000 steps per day is associated with a decrease in BMI of between 0.341 and 0.969 kg/m^2.

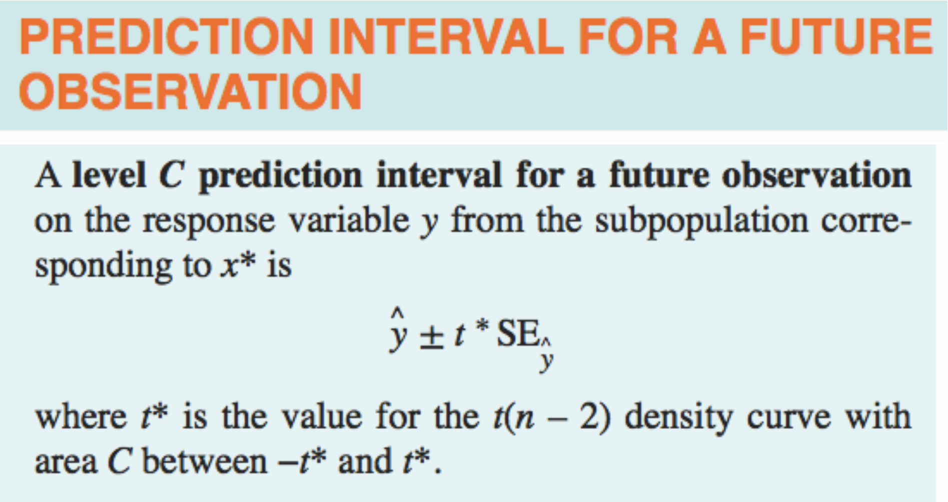


CONFIDENCE INTERVAL FOR A MEAN RESPONSE

A level *C* confidence interval for the mean response μ_y when x takes the value x^* is

$$\hat{\mu}_y \pm t\,{*}\,\mathrm{SE}_{\hat{\mu}}$$

where t^* is the value for the $t(n - 2)$ density curve with area C between $-t^*$ and t^*.

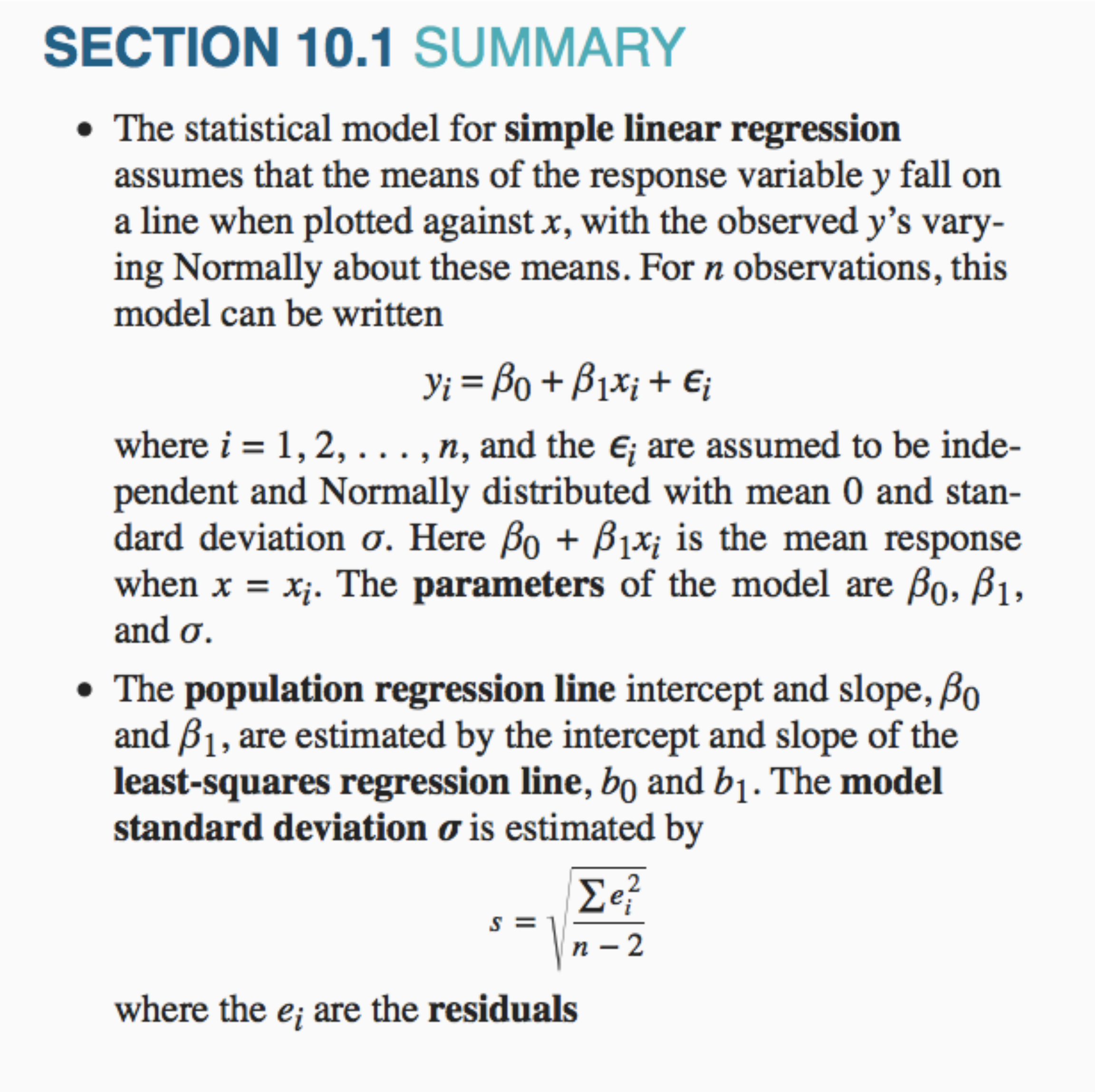


PREDICTION INTERVAL FOR A FUTURE OBSERVATION

A **level C prediction interval for a future observation** on the response variable y from the subpopulation corresponding to x^* is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$$

where t^* is the value for the $t(n - 2)$ density curve with area C between $-t^*$ and t^*.



SECTION 10.1 SUMMARY

- The statistical model for **simple linear regression** assumes that the means of the response variable y fall on a line when plotted against x, with the observed y's varying Normally about these means. For n observations, this model can be written

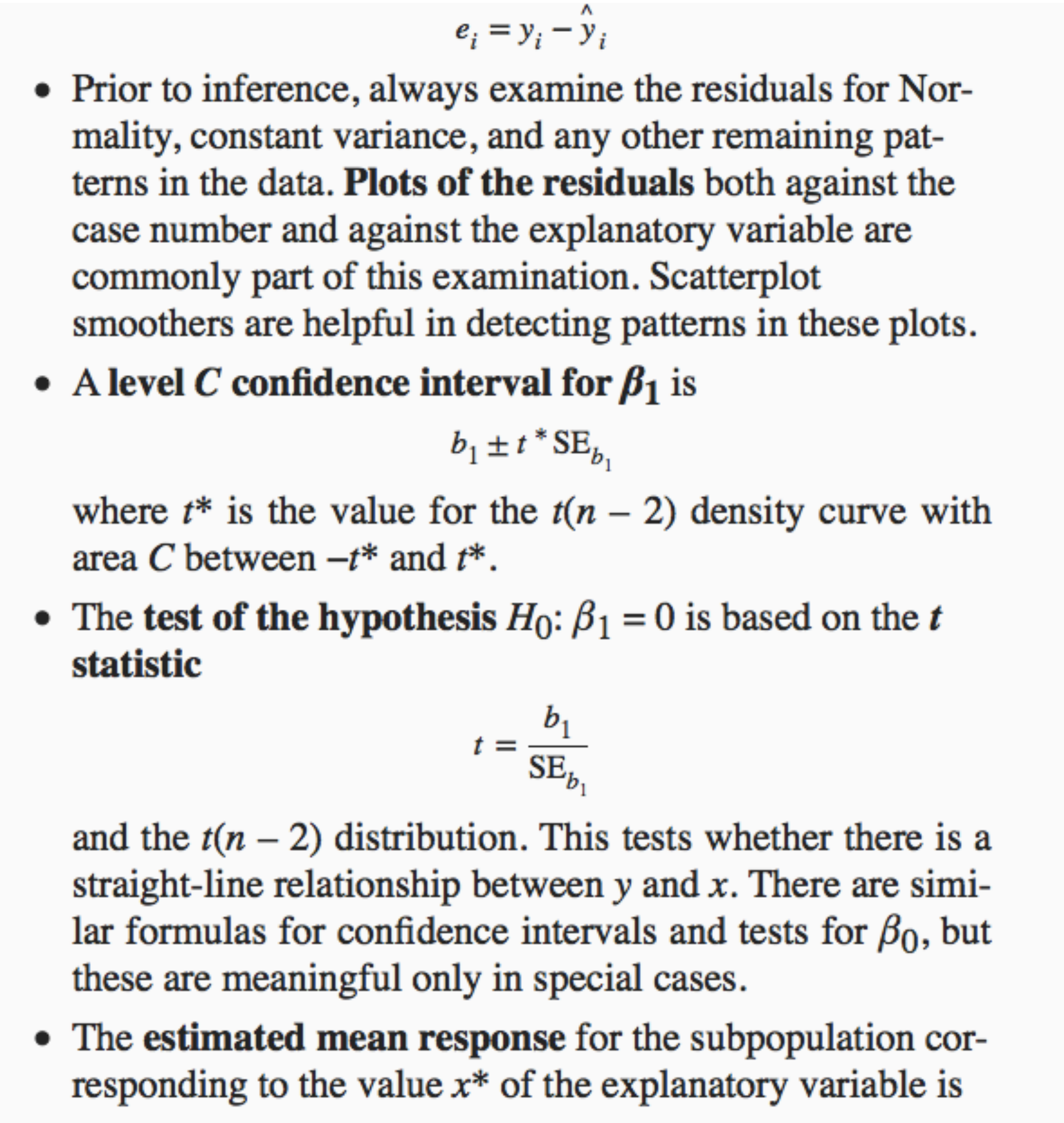
  $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  where $i = 1, 2, \ldots, n$, and the ϵ_i are assumed to be independent and Normally distributed with mean 0 and standard deviation σ. Here $\beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$. The **parameters** of the model are β_0, β_1, and σ.

- The **population regression line** intercept and slope, β_0 and β_1, are estimated by the intercept and slope of the **least-squares regression line,** b_0 and b_1. The **model standard deviation σ** is estimated by

  $$s = \sqrt{\frac{\sum e_i^2}{n - 2}}$$

  where the e_i are the **residuals**



$$e_i = y_i - \hat{y}_i$$

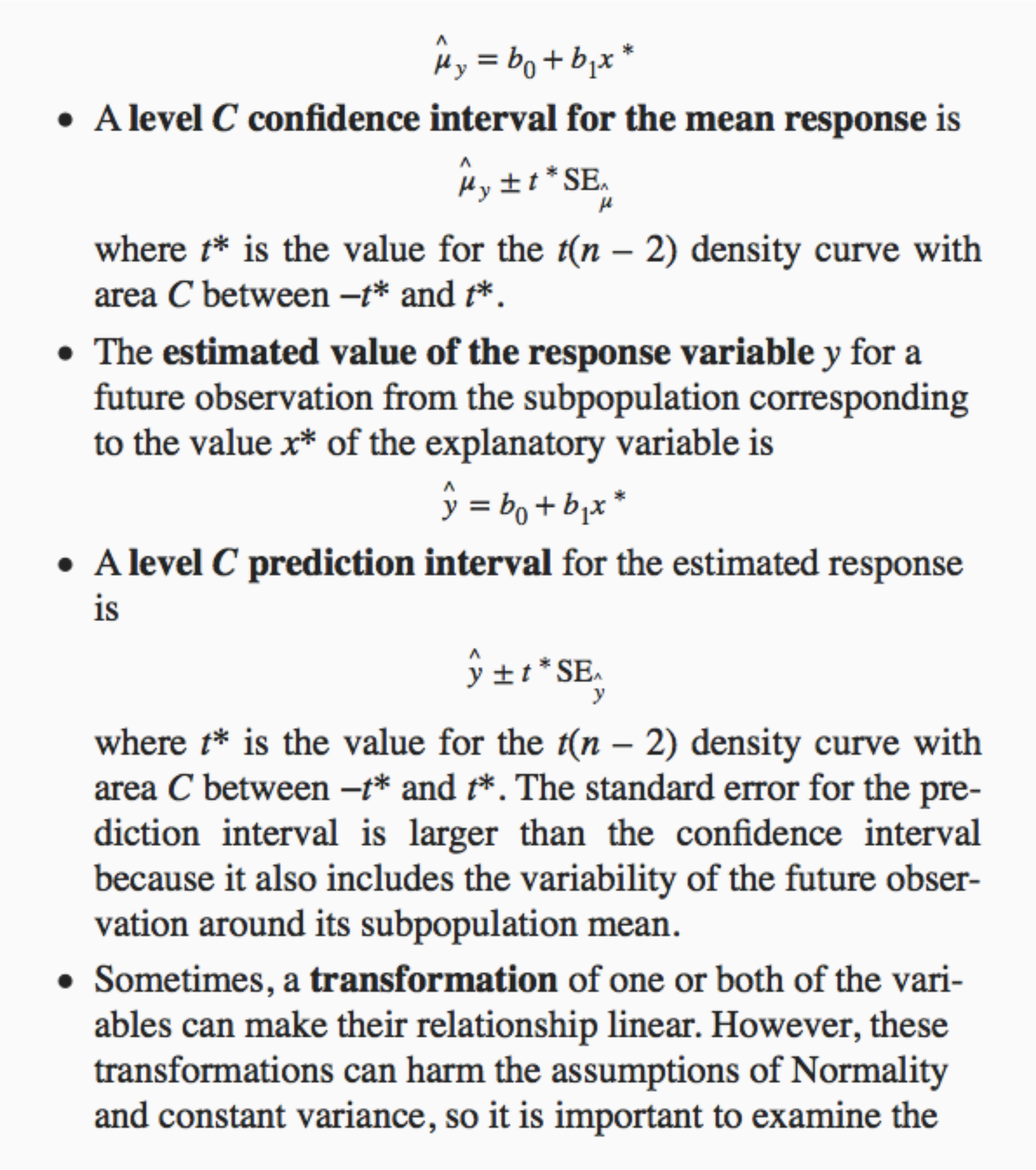
- Prior to inference, always examine the residuals for Normality, constant variance, and any other remaining patterns in the data. **Plots of the residuals** both against the case number and against the explanatory variable are commonly part of this examination. Scatterplot smoothers are helpful in detecting patterns in these plots.
- A **level *C* confidence interval for $\boldsymbol{\beta_1}$** is

  $$b_1 \pm t^* \mathrm{SE}_{b_1}$$

  where t^* is the value for the $t(n - 2)$ density curve with area C between $-t^*$ and t^*.
- The **test of the hypothesis** H_0: $\beta_1 = 0$ is based on the ***t* statistic**

  $$t = \frac{b_1}{\mathrm{SE}_{b_1}}$$

  and the $t(n - 2)$ distribution. This tests whether there is a straight-line relationship between y and x. There are similar formulas for confidence intervals and tests for β_0, but these are meaningful only in special cases.
- The **estimated mean response** for the subpopulation corresponding to the value x^* of the explanatory variable is



$$\hat{\mu}_y = b_0 + b_1 x^*$$

- **A level *C* confidence interval for the mean response** is

$$\hat{\mu}_y \pm t^* \mathrm{SE}_{\hat{\mu}}$$

  where t^* is the value for the $t(n-2)$ density curve with area *C* between $-t^*$ and t^*.

- The **estimated value of the response variable** *y* for a future observation from the subpopulation corresponding to the value x^* of the explanatory variable is

$$\hat{y} = b_0 + b_1 x^*$$

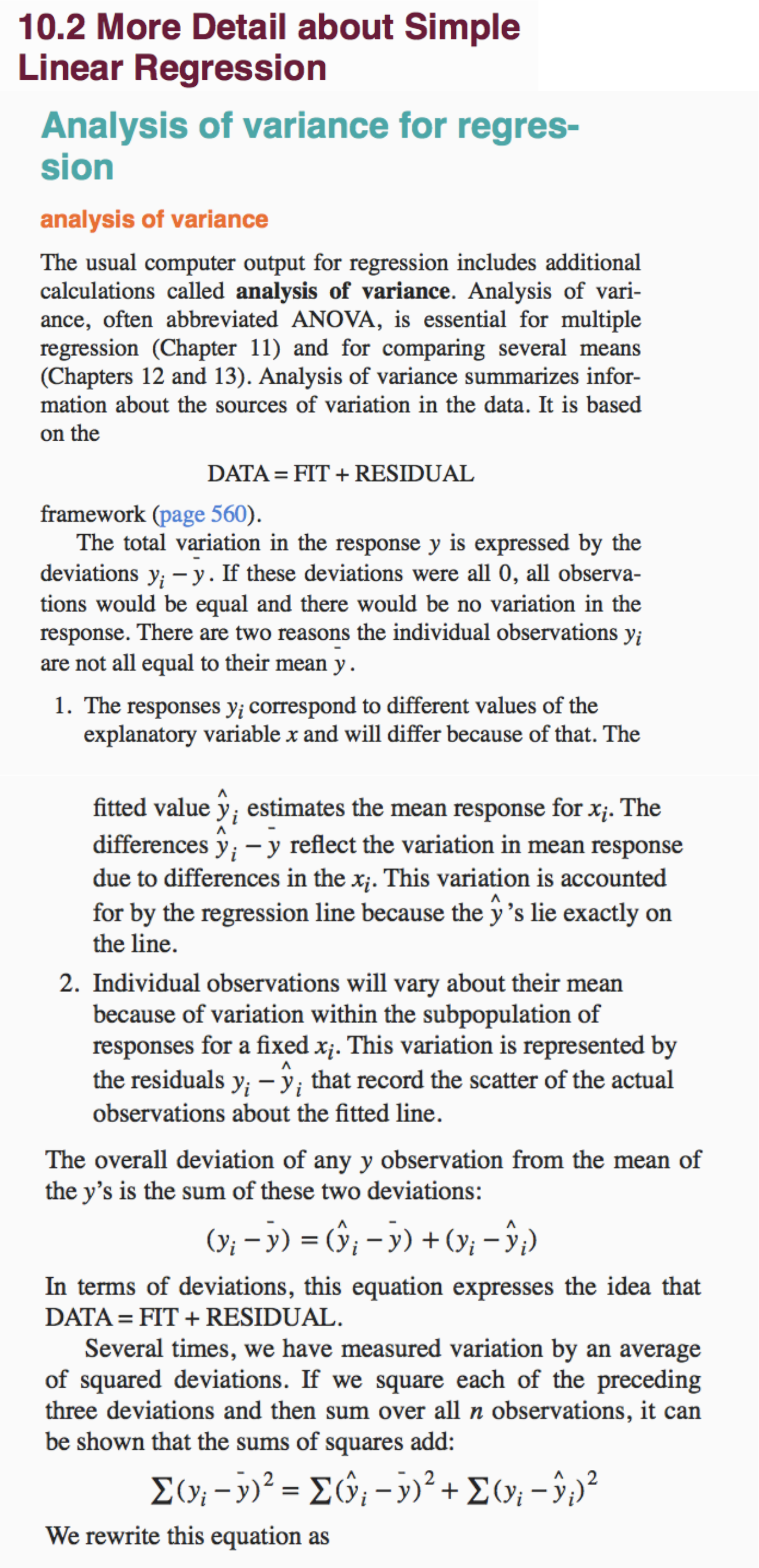
- **A level *C* prediction interval** for the estimated response is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$$

  where t^* is the value for the $t(n-2)$ density curve with area *C* between $-t^*$ and t^*. The standard error for the prediction interval is larger than the confidence interval because it also includes the variability of the future observation around its subpopulation mean.

- Sometimes, a **transformation** of one or both of the variables can make their relationship linear. However, these transformations can harm the assumptions of Normality and constant variance, so it is important to examine the

residuals



10.2 More Detail about Simple Linear Regression

Analysis of variance for regression

analysis of variance

The usual computer output for regression includes additional calculations called **analysis of variance**. Analysis of variance, often abbreviated ANOVA, is essential for multiple regression (Chapter 11) and for comparing several means (Chapters 12 and 13). Analysis of variance summarizes information about the sources of variation in the data. It is based on the

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

framework (page 560).

The total variation in the response y is expressed by the deviations $y_i - \bar{y}$. If these deviations were all 0, all observations would be equal and there would be no variation in the response. There are two reasons the individual observations y_i are not all equal to their mean $\bar{y}$.

1. The responses y_i correspond to different values of the explanatory variable x and will differ because of that. The fitted value $\hat{y}_i$ estimates the mean response for x_i. The differences $\hat{y}_i - \bar{y}$ reflect the variation in mean response due to differences in the x_i. This variation is accounted for by the regression line because the $\hat{y}$'s lie exactly on the line.
2. Individual observations will vary about their mean because of variation within the subpopulation of responses for a fixed x_i. This variation is represented by the residuals $y_i - \hat{y}_i$ that record the scatter of the actual observations about the fitted line.

The overall deviation of any y observation from the mean of the y's is the sum of these two deviations:

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

In terms of deviations, this equation expresses the idea that DATA = FIT + RESIDUAL.

Several times, we have measured variation by an average of squared deviations. If we square each of the preceding three deviations and then sum over all n observations, it can be shown that the sums of squares add:

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

We rewrite this equation as

$$\text{SST} = \text{SSM} + \text{SSE}$$

where

$$\begin{aligned} \text{SST} &= \sum (y_i - \bar{y})^2 \\ \text{SSM} &= \sum (\hat{y}_i - \bar{y})^2 \\ \text{SSE} &= \sum (y_i - \hat{y}_i)^2 \end{aligned}$$

sum of squares

The SS in each abbreviation stands for **sum of squares,** and the T, M, and E stand for total, model, and error, respectively. ("Error" here stands for deviations from the line, which might better be called "residual" or "unexplained variation.") The total variation, as expressed by SST, is the sum of the variation due to the straight-line model (SSM) and the variation due to deviations from this model (SSE). This partition of the variation in the data between two sources is the heart of analysis of variance.

If H_0: $\beta_1 = 0$ were true, there would be no subpopulations, and all of the y's should be viewed as coming from a single population with mean μ_y. The variation of the y's would then be described by the sample variance

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

The numerator in this expression is SST. The denominator is the total degrees of freedom, or simply DFT.

LOOK BACK

degrees of freedom, p. 40

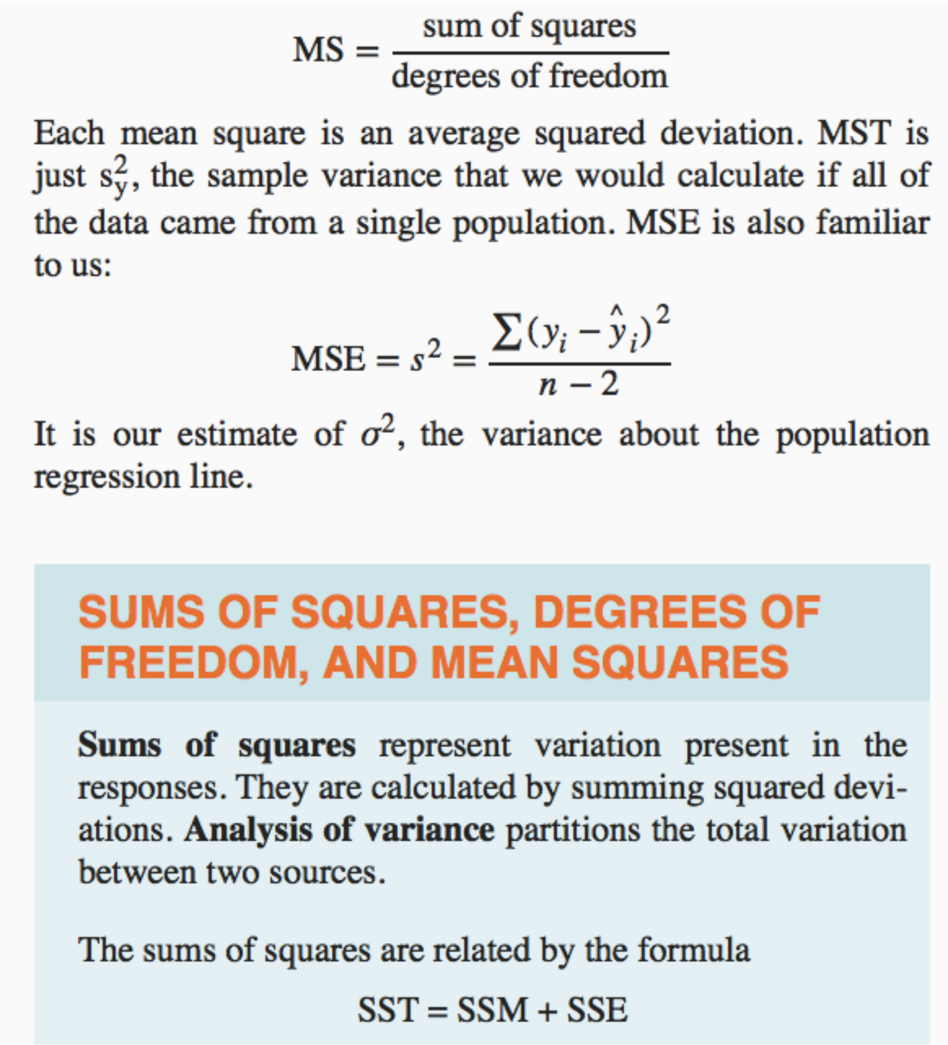
Just as the total sum of squares SST is the sum of SSM and SSE, the total degrees of freedom DFT is the sum of DFM and DFE, the degrees of freedom for the model and for the error:

$$\text{DFT} = \text{DFM} + \text{DFE}$$

The model has one explanatory variable x, so the degrees of freedom for this source are DFM = 1. Because DFT = $n - 1$, this leaves DFE = $n - 2$ as the degrees of freedom for error.

mean square

For each source, the ratio of the sum of squares to the degrees of freedom is called the **mean square**, or simply MS. The general formula for a mean square is



$$\text{MS} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

Each mean square is an average squared deviation. MST is just s_y^2, the sample variance that we would calculate if all of the data came from a single population. MSE is also familiar to us:

$$\text{MSE} = s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2}$$

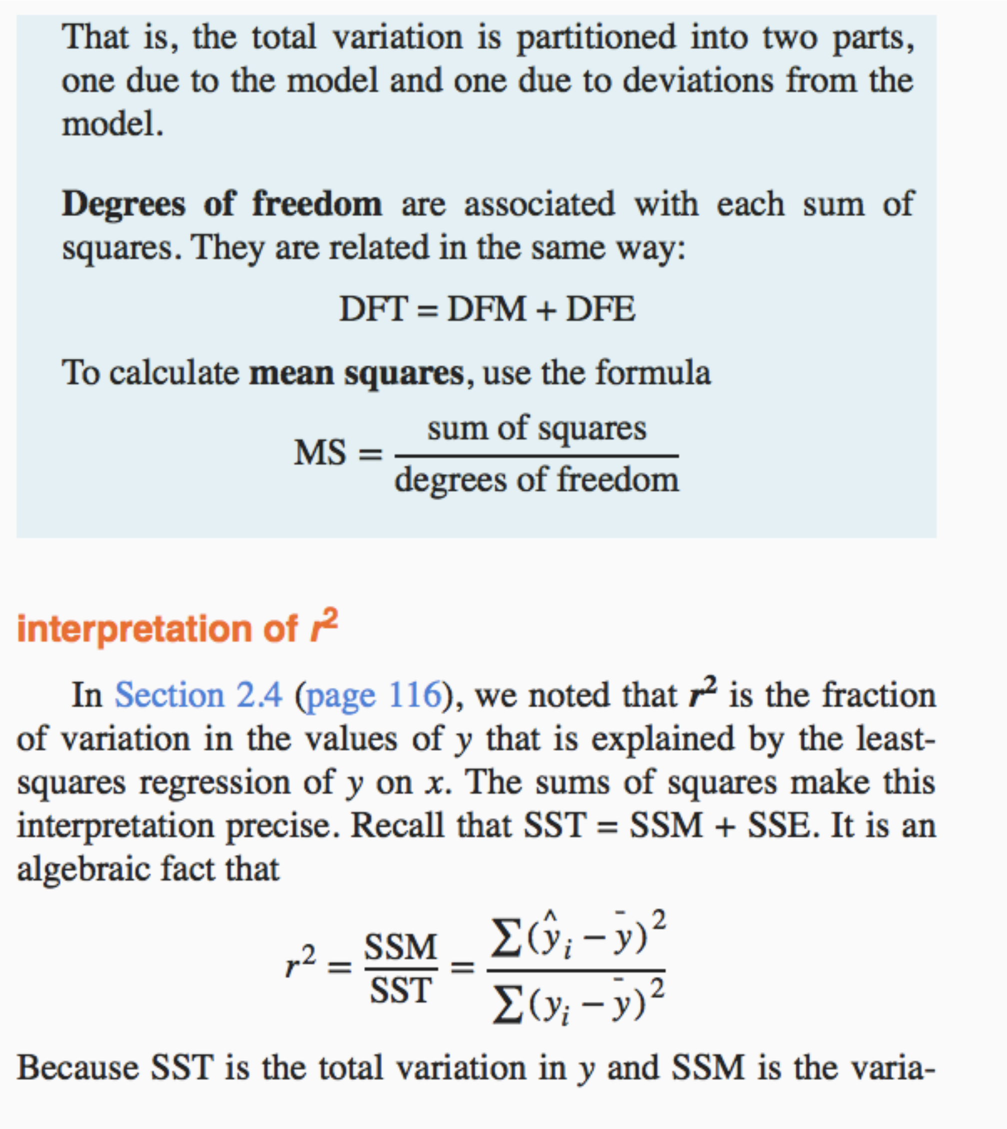
It is our estimate of σ^2, the variance about the population regression line.

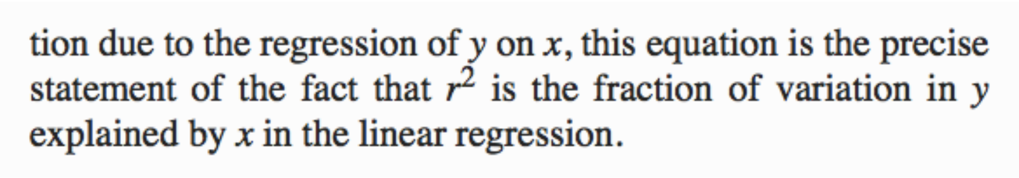
SUMS OF SQUARES, DEGREES OF FREEDOM, AND MEAN SQUARES

Sums of squares represent variation present in the responses. They are calculated by summing squared deviations. **Analysis of variance** partitions the total variation between two sources.

The sums of squares are related by the formula

$$\text{SST} = \text{SSM} + \text{SSE}$$





That is, the total variation is partitioned into two parts, one due to the model and one due to deviations from the model.

Degrees of freedom are associated with each sum of squares. They are related in the same way:

$$\text{DFT} = \text{DFM} + \text{DFE}$$

To calculate **mean squares**, use the formula

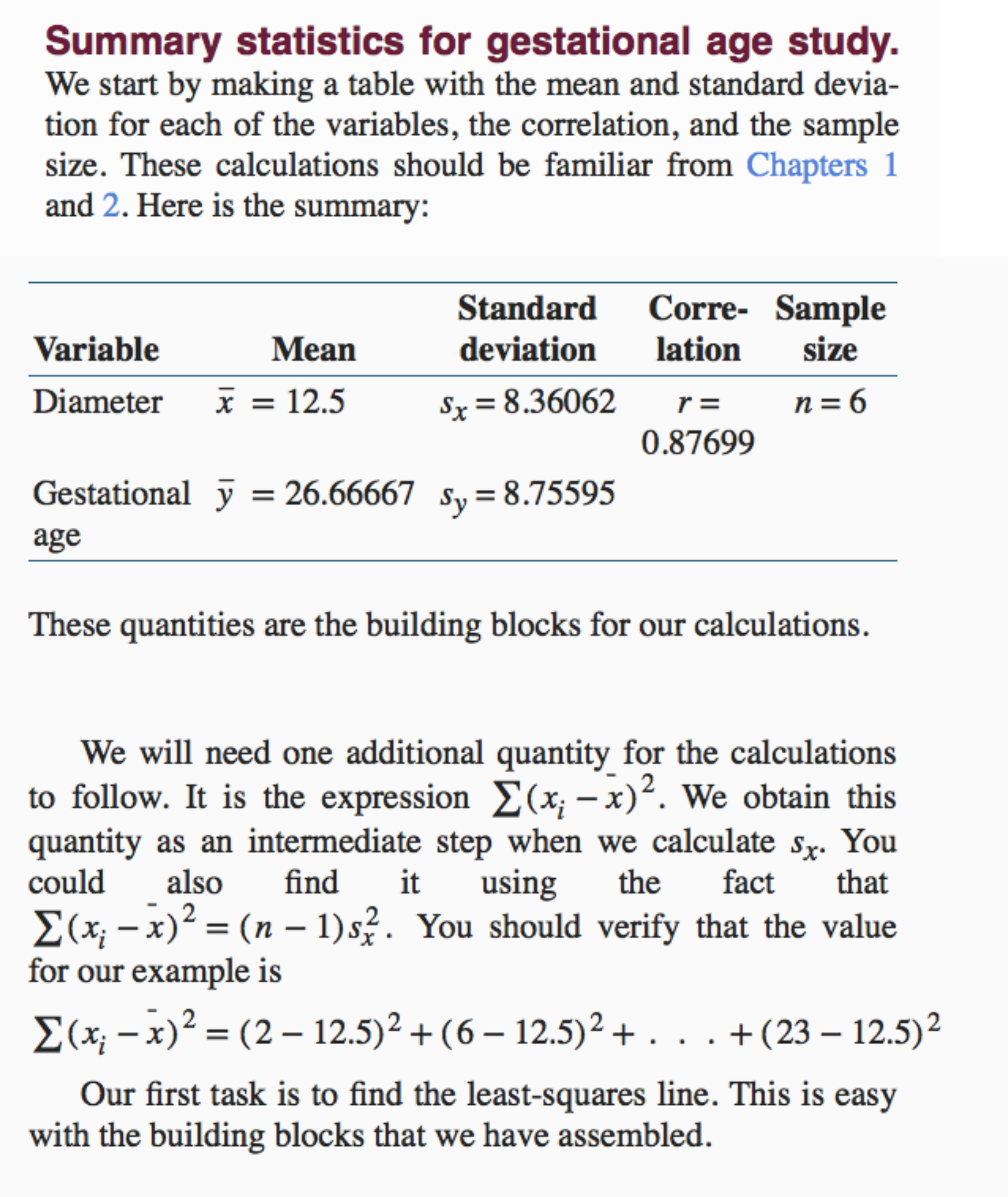
$$\text{MS} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

interpretation of r^2

In Section 2.4 (page 116), we noted that r^2 is the fraction of variation in the values of y that is explained by the least-squares regression of y on x. The sums of squares make this interpretation precise. Recall that SST = SSM + SSE. It is an algebraic fact that

$$r^2 = \frac{\text{SSM}}{\text{SST}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Because SST is the total variation in y and SSM is the variation due to the regression of y on x, this equation is the precise statement of the fact that r^2 is the fraction of variation in y explained by x in the linear regression.



Summary statistics for gestational age study.

We start by making a table with the mean and standard deviation for each of the variables, the correlation, and the sample size. These calculations should be familiar from Chapters 1 and 2. Here is the summary:

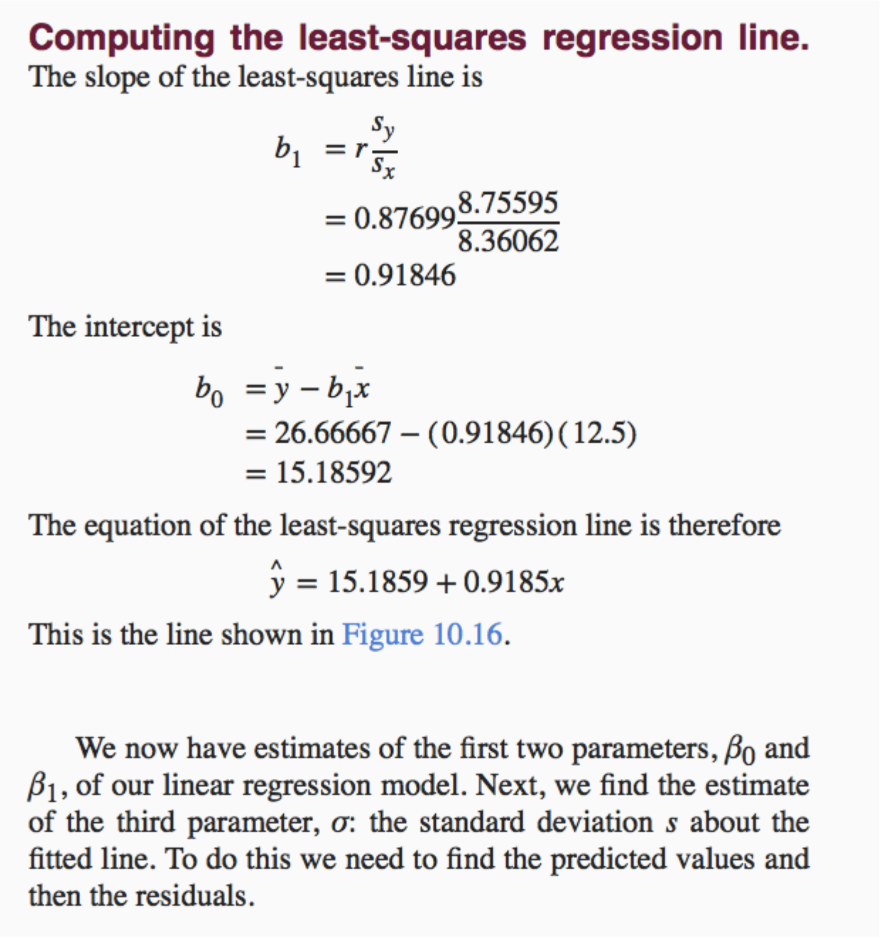
Variable	Mean	Standard deviation	Correlation	Sample size
Diameter	$\bar{x} = 12.5$	$s_x = 8.36062$	$r =$ 0.87699	$n = 6$
Gestational age	$\bar{y} = 26.66667$	$s_y = 8.75595$		

These quantities are the building blocks for our calculations.

We will need one additional quantity for the calculations to follow. It is the expression $\sum(x_i - \bar{x})^2$. We obtain this quantity as an intermediate step when we calculate s_x. You could also find it using the fact that $\sum(x_i - \bar{x})^2 = (n-1)s_x^2$. You should verify that the value for our example is

$$\sum(x_i - \bar{x})^2 = (2 - 12.5)^2 + (6 - 12.5)^2 + \ . \ . \ . + (23 - 12.5)^2$$

Our first task is to find the least-squares line. This is easy with the building blocks that we have assembled.



Computing the least-squares regression line. The slope of the least-squares line is

$$\begin{aligned} b_1 &= r\frac{s_y}{s_x} \\ &= 0.87699\frac{8.75595}{8.36062} \\ &= 0.91846 \end{aligned}$$

The intercept is

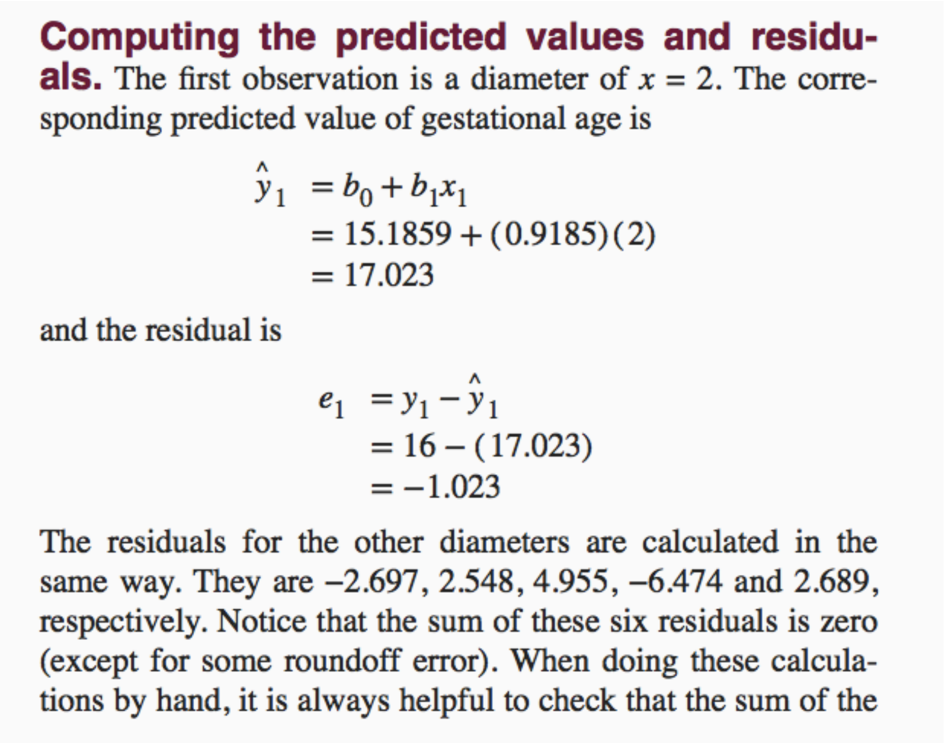
$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{x} \\ &= 26.66667 - (0.91846)(12.5) \\ &= 15.18592 \end{aligned}$$

The equation of the least-squares regression line is therefore

$$\hat{y} = 15.1859 + 0.9185x$$

This is the line shown in Figure 10.16.

We now have estimates of the first two parameters, β_0 and β_1, of our linear regression model. Next, we find the estimate of the third parameter, σ: the standard deviation s about the fitted line. To do this we need to find the predicted values and then the residuals.



Computing the predicted values and residuals. The first observation is a diameter of $x = 2$. The corresponding predicted value of gestational age is

$$\begin{aligned} \hat{y}_1 &= b_0 + b_1x_1 \\ &= 15.1859 + (0.9185)(2) \\ &= 17.023 \end{aligned}$$

and the residual is

$$\begin{aligned} e_1 &= y_1 - \hat{y}_1 \\ &= 16 - (17.023) \\ &= -1.023 \end{aligned}$$

The residuals for the other diameters are calculated in the same way. They are –2.697, 2.548, 4.955, –6.474 and 2.689, respectively. Notice that the sum of these six residuals is zero (except for some roundoff error). When doing these calculations by hand, it is always helpful to check that the sum of the residuals is zero.

EXAMPLE 10.20

Computing s^2. The estimate of σ^2 is s^2, the sum of the squares of the residuals divided by $n - 2$. The estimated standard deviation about the line is the square root of this quantity.

$$\begin{aligned} s^2 &= \frac{\sum e_i^2}{n-2} \\ &= \frac{(-1.023)^2 + (-2.697)^2 + \ .\ .\ . + (2.689)^2}{4} \\ &= 22.127 \end{aligned}$$

So the estimate of the standard deviation about the line is

$$s = \sqrt{22.12702} = 4.704$$

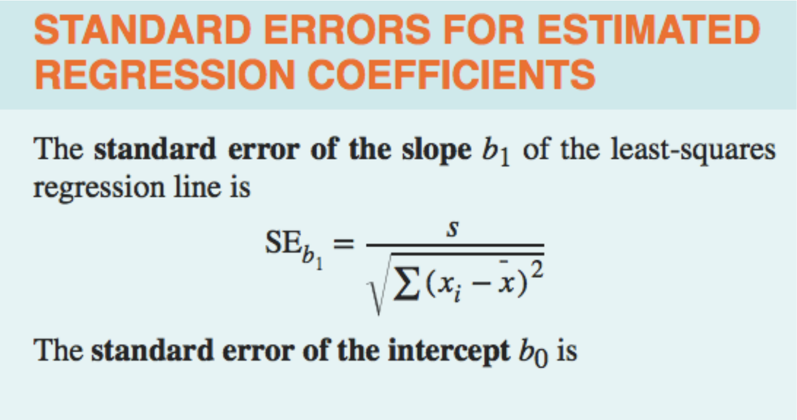
Inference for slope and intercept Confidence intervals and significance tests for the slope β_1 and intercept β_0 of the population regression line make use of the estimates b_1 and b_0 and their standard errors. Some algebra and the rules for variances establishes that the standard deviation of b_1 is

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Similarly, the standard deviation of b_0 is

$$\sigma_{b_0} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

To estimate these standard deviations, we need only replace σ by its estimate s.

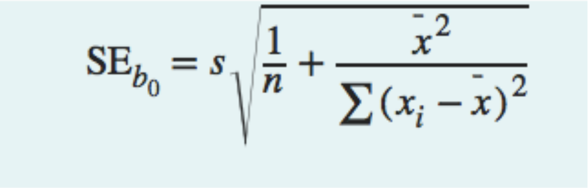


STANDARD ERRORS FOR ESTIMATED REGRESSION COEFFICIENTS

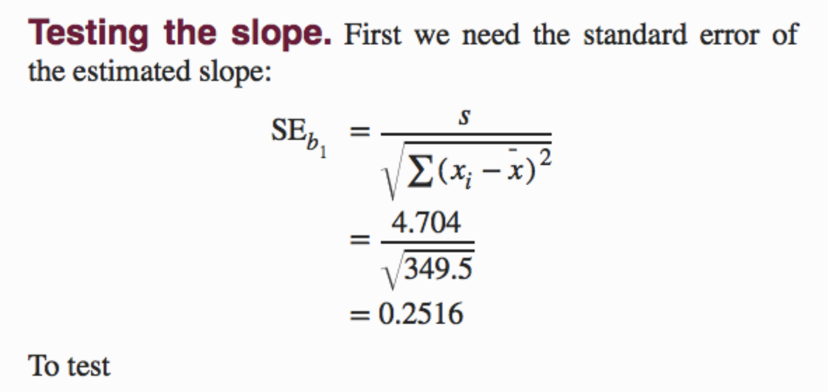
The **standard error of the slope** b_1 of the least-squares regression line is

$$\text{SE}_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

The **standard error of the intercept** b_0 is



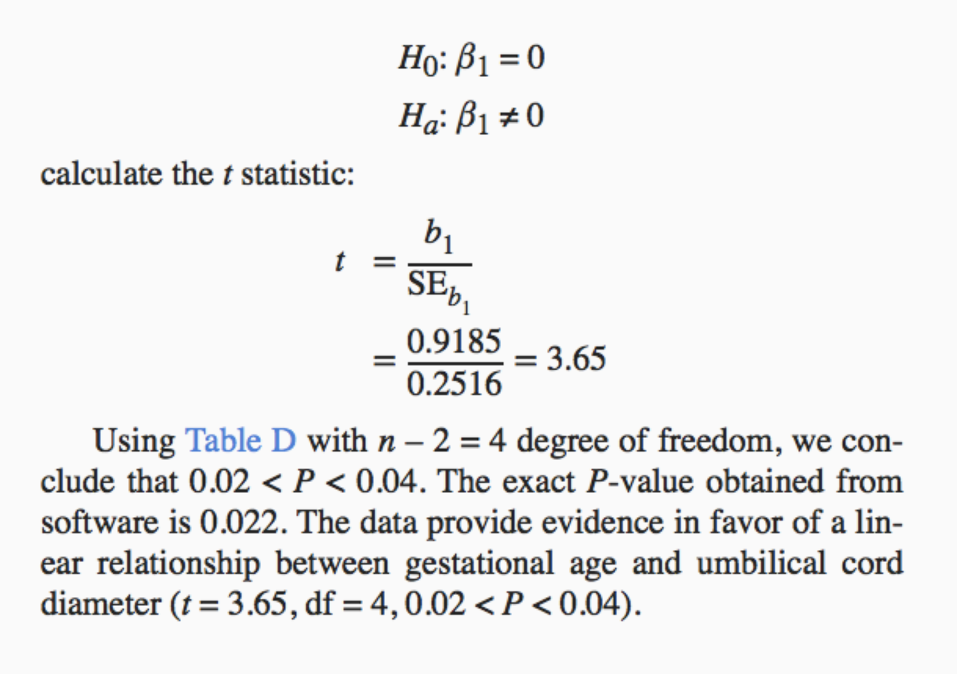
$$\text{SE}_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$



Testing the slope. First we need the standard error of the estimated slope:

$$\begin{aligned}\text{SE}_{b_1} &= \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}\\ &= \frac{4.704}{\sqrt{349.5}}\\ &= 0.2516\end{aligned}$$

To test

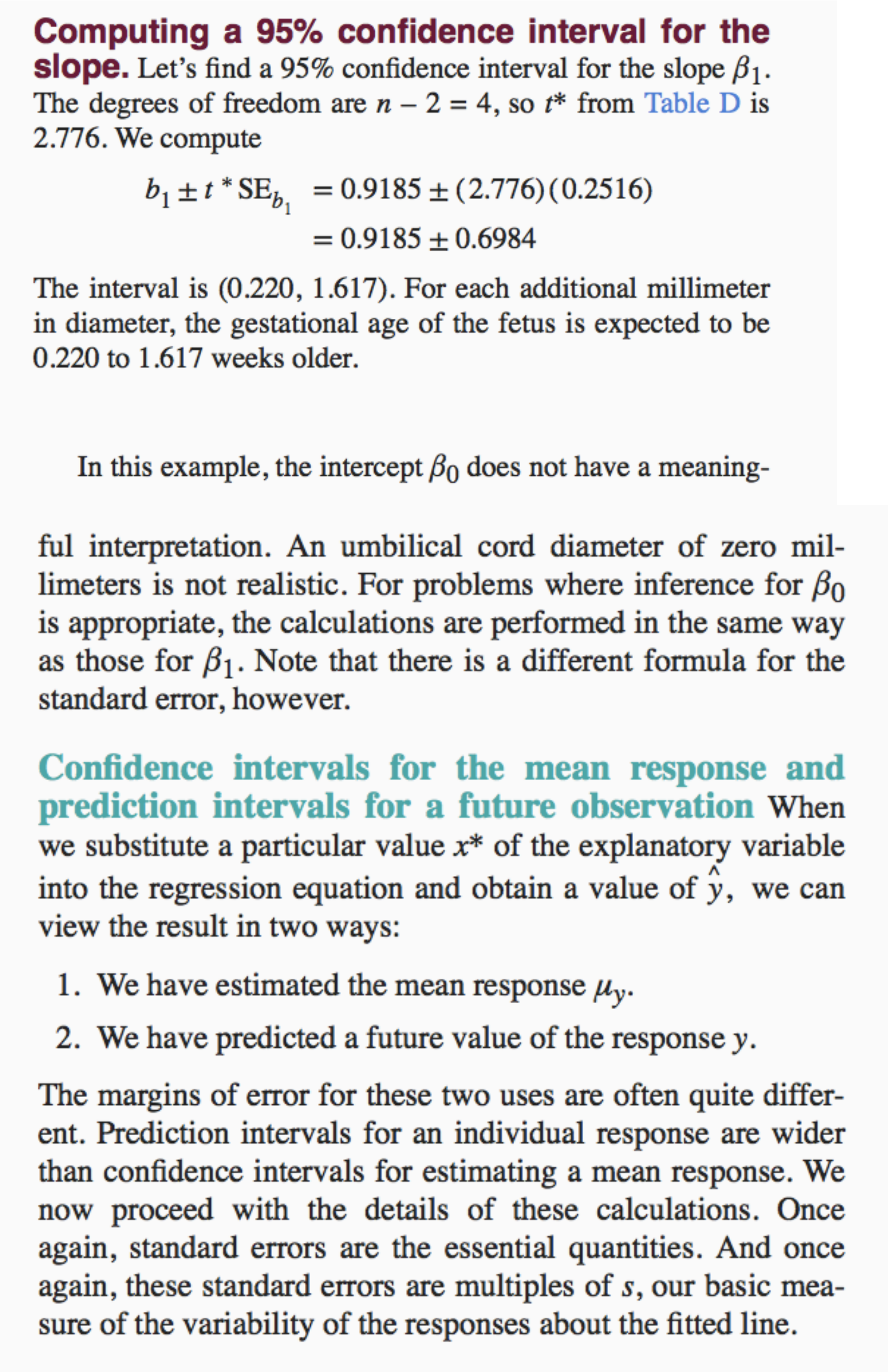


$$H_0\colon \beta_1 = 0$$
$$H_a\colon \beta_1 \neq 0$$

calculate the t statistic:

$$\begin{aligned}t &= \frac{b_1}{\text{SE}_{b_1}}\\ &= \frac{0.9185}{0.2516} = 3.65\end{aligned}$$

Using Table D with $n - 2 = 4$ degree of freedom, we conclude that $0.02 < P < 0.04$. The exact P-value obtained from software is 0.022. The data provide evidence in favor of a linear relationship between gestational age and umbilical cord diameter ($t = 3.65$, df $= 4$, $0.02 < P < 0.04$).



Computing a 95% confidence interval for the slope. Let's find a 95% confidence interval for the slope β_1. The degrees of freedom are $n - 2 = 4$, so t^* from Table D is 2.776. We compute

$$\begin{aligned} b_1 \pm t^* \text{SE}_{b_1} &= 0.9185 \pm (2.776)(0.2516) \\ &= 0.9185 \pm 0.6984 \end{aligned}$$

The interval is (0.220, 1.617). For each additional millimeter in diameter, the gestational age of the fetus is expected to be 0.220 to 1.617 weeks older.

In this example, the intercept β_0 does not have a meaningful interpretation. An umbilical cord diameter of zero millimeters is not realistic. For problems where inference for β_0 is appropriate, the calculations are performed in the same way as those for β_1. Note that there is a different formula for the standard error, however.

Confidence intervals for the mean response and prediction intervals for a future observation When we substitute a particular value x^* of the explanatory variable into the regression equation and obtain a value of $\hat{y}$, we can view the result in two ways:

1. We have estimated the mean response μ_y.
2. We have predicted a future value of the response y.

The margins of error for these two uses are often quite different. Prediction intervals for an individual response are wider than confidence intervals for estimating a mean response. We now proceed with the details of these calculations. Once again, standard errors are the essential quantities. And once again, these standard errors are multiples of s, our basic measure of the variability of the responses about the fitted line.

STANDARD ERRORS FOR $\hat{\mu}$ AND $\hat{Y}$

The standard error of $\hat{\mu}$ is

$$\mathrm{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})2}{\sum(x_i - \bar{x})^2}}$$

The standard error for predicting an individual response $\hat{y}$ is

$$\mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})2}{\sum(x_i - \bar{x})^2}}$$

Computing a confidence interval for μ. Let's find a 95% confidence interval for the average gestational age when the umbilical cord diameter is 10 millimeters. The estimated mean age is

$$\begin{aligned}\hat{\mu} &= b_0 + b_1 x\\ &= 15.1859 + (0.9185)(10)\\ &= 24.371\end{aligned}$$

The standard error is

$$\begin{aligned}\mathrm{SE}_{\hat{\mu}} &= s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\\ &= 4.704\sqrt{\frac{1}{6} + \frac{(10.0 - 12.5)^2}{349.5}}\\ &= 2.021\end{aligned}$$

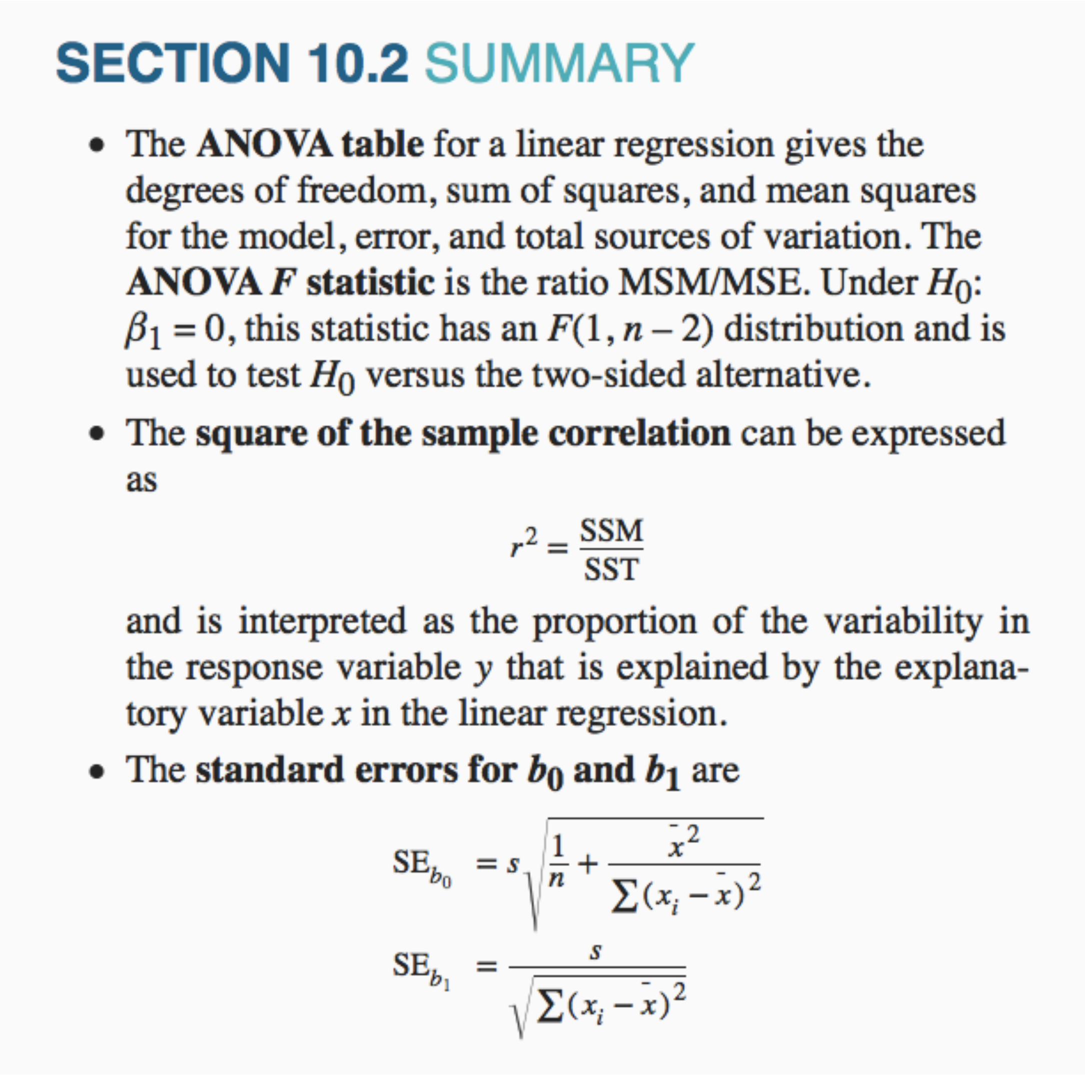
To find the 95% confidence interval, we compute

$$\begin{aligned}\hat{\mu} \pm t^* \mathrm{SE}_{\hat{\mu}} &= 24.371 \pm (2.776)(2.021)\\ &= 24.371 \pm 5.610\end{aligned}$$

The interval is 18.761 to 29.981 weeks of age. This is a pretty wide interval given gestation lasts for about 40 weeks.

Calculations for the prediction intervals are similar. The only difference is the use of the formula for $\mathrm{SE}_{\hat{y}}$ in place of $\mathrm{SE}_{\hat{\mu}}$. This results in a much wider interval. In fact, the interval is slightly more than 28 weeks in width. Even though a linear relationship was found statistically significant, it does

not appear umbilical cord diameter is a precise predictor of gestational age.



SECTION 10.2 SUMMARY

- The **ANOVA table** for a linear regression gives the degrees of freedom, sum of squares, and mean squares for the model, error, and total sources of variation. The **ANOVA *F* statistic** is the ratio MSM/MSE. Under H_0: $\beta_1 = 0$, this statistic has an $F(1, n-2)$ distribution and is used to test H_0 versus the two-sided alternative.
- The **square of the sample correlation** can be expressed as

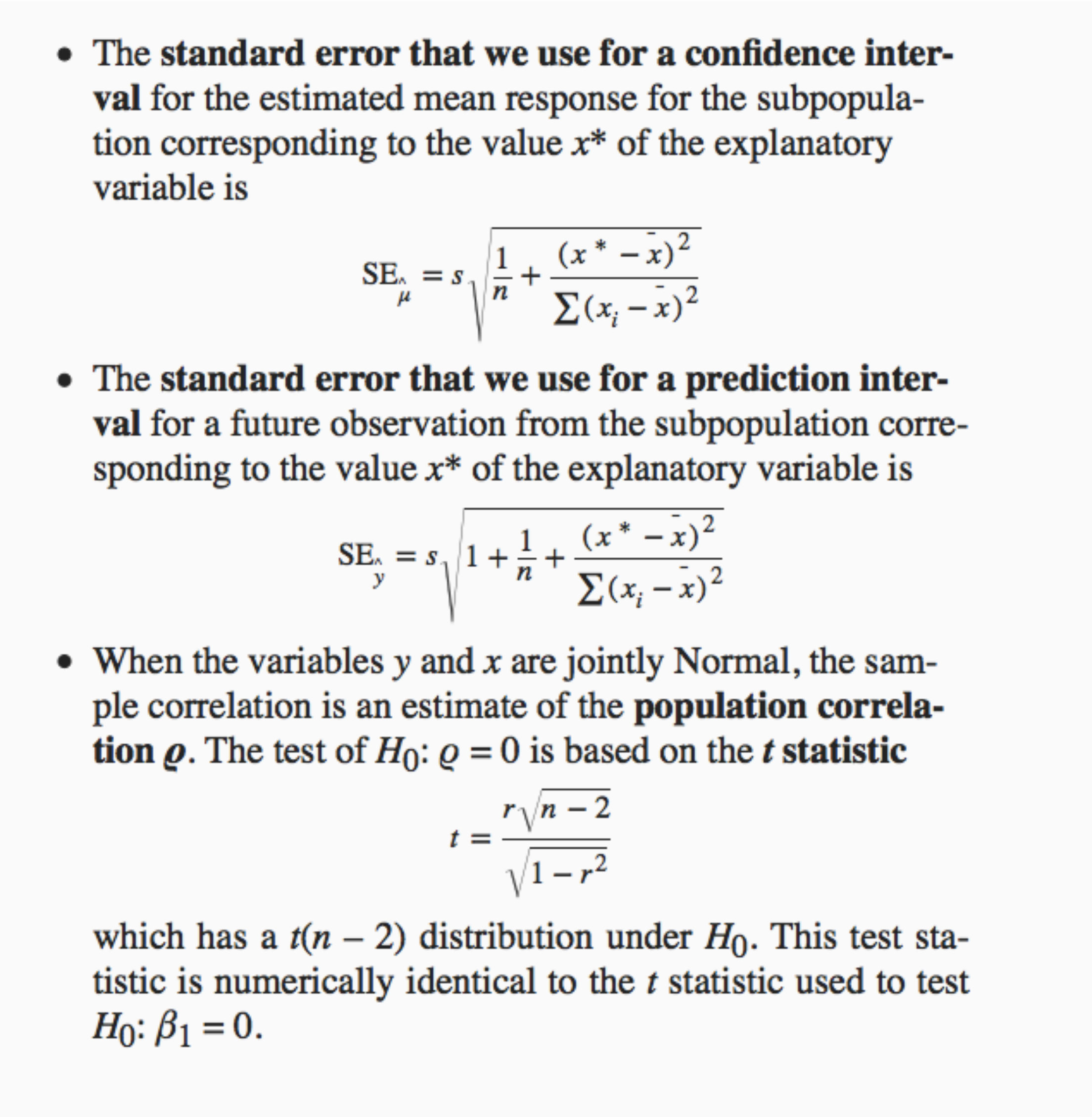
$$r^2 = \frac{\text{SSM}}{\text{SST}}$$

and is interpreted as the proportion of the variability in the response variable y that is explained by the explanatory variable x in the linear regression.

- The **standard errors for b_0 and b_1** are

$$\text{SE}_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

$$\text{SE}_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$



- The **standard error that we use for a confidence interval** for the estimated mean response for the subpopulation corresponding to the value *x** of the explanatory variable is

$$\mathrm{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

- The **standard error that we use for a prediction interval** for a future observation from the subpopulation corresponding to the value *x** of the explanatory variable is

$$\mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

- When the variables *y* and *x* are jointly Normal, the sample correlation is an estimate of the **population correlation ϱ**. The test of H_0: $\varrho = 0$ is based on the ***t* statistic**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which has a $t(n - 2)$ distribution under H_0. This test statistic is numerically identical to the *t* statistic used to test H_0: $\beta_1 = 0$.