# Chapter 8
# Linear Regression by Least Squares

## 8.1 Introduction

We usually study more than one variable at a time. When the variables are continuous, and one is clearly a response variable and the others are predictor variables, we usually plot the variables and then attempt to fit a model to the plotted points. With one continuous predictor, the first model we attempt is a straight line; with two or more continuous predictors, we attempt a plane. We plot the model, the residuals from the model, and various diagnostics of the quality of the fit.

In this chapter we are primarily concerned with modeling a straight-line relationship between two variables using $n$ pairs of observations on these variables, a common and fundamental task. One of these variables, conventionally denoted $y$, is a response or output variable. The other variable, often denoted $x$, is known as an explanatory or input or predictor variable. Usually, but not always, it is clear from the context which of the two variables is the response and which is the predictor. For example, if the two variables are personal `income` and `consumption` spending, then `consumption` is the response variable because the amount that is spent depends on how much `income` is available to be spent.

The relationship between $y$ and $x$ is almost never perfectly linear. When the $n$ points are plotted in two dimensions, they appear as a random scatter about some unknown straight line. We model this line as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for } i = 1, \ldots, n \tag{8.1}$$

where

$$\epsilon_i \sim N(0, \sigma^2) \tag{8.2}$$

that is, the $\epsilon_i$ are assumed normally independently distributed with constant mean 0 and common variance $\sigma^2$ [abbreviated as $\epsilon_i \sim NID(0, \sigma^2)$]. In other words, we assume that the response variable is linearly related to the predictor variables, plus

R.M. Heiberger, B. Holland, *Statistical Analysis and Data Display*,
Springer Texts in Statistics, DOI 10.1007/978-1-4939-2122-5_8

a normally distributed random component. Here the intercept $\beta_0$ and slope $\beta_1$ are unknown *regression coefficients* that must be estimated from the data. The variance $\sigma^2$ is a third unknown parameter, introduced along with the assumption of a normally distributed error term, which must also be estimated.

A commonly used procedure for estimating $\beta_0$ and $\beta_1$ is the method of *least squares* because, as we will see in Section 8.3.2, this mathematical criterion leads to simple "closed-form" formulas for the estimates. Under the stated normality assumptions in Equation (8.2) about the residuals $\epsilon_i$ of Model (8.1), the least-squares estimates of the regression coefficients are also the maximum likelihood estimates of these coefficients.

## 8.2 Example—Body Fat Data

### *8.2.1 Study Objectives*

The example is taken from Johnson (1996). A group of subjects is gathered, and various body measurements and an accurate estimate of the percentage of body fat are recorded for each. Then body fat can be fit to the other body measurements using multiple regression, giving, we hope, a useful predictive equation for people similar to the subjects. The various measurements other than body fat recorded on the subjects are, implicitly, ones that are easy to obtain and serve as proxies for body fat, which is not so easily obtained.

Percentage of body fat, age, weight, height, and ten body circumference measurements (e.g., abdomen) are recorded for 252 men. Body fat, a measure of health, is estimated through an underwater weighing technique. Fitting body fat to the other measurements using multiple regression provides a convenient way of estimating body fat for men using only a scale and a measuring tape.

### *8.2.2 Data Description*

We will initially use only 47 observations and only five of the measurements that have been recorded.

`bodyfat`: Percent body fat using Siri's equation, 495/Density − 450

`abdomin`: Abdomen circumference (cm) "at the umbilicus and level with the iliac crest"

`biceps`: Extended biceps circumference (cm) `wrist:` Wrist circumference (cm) “distal to the styloid processes”

`forearm:` Forearm circumference (cm)

### 8.2.3 Data Input

We access the data from `data(fat)` and then look at the data with the scatterplot matrix in Figure 8.1.

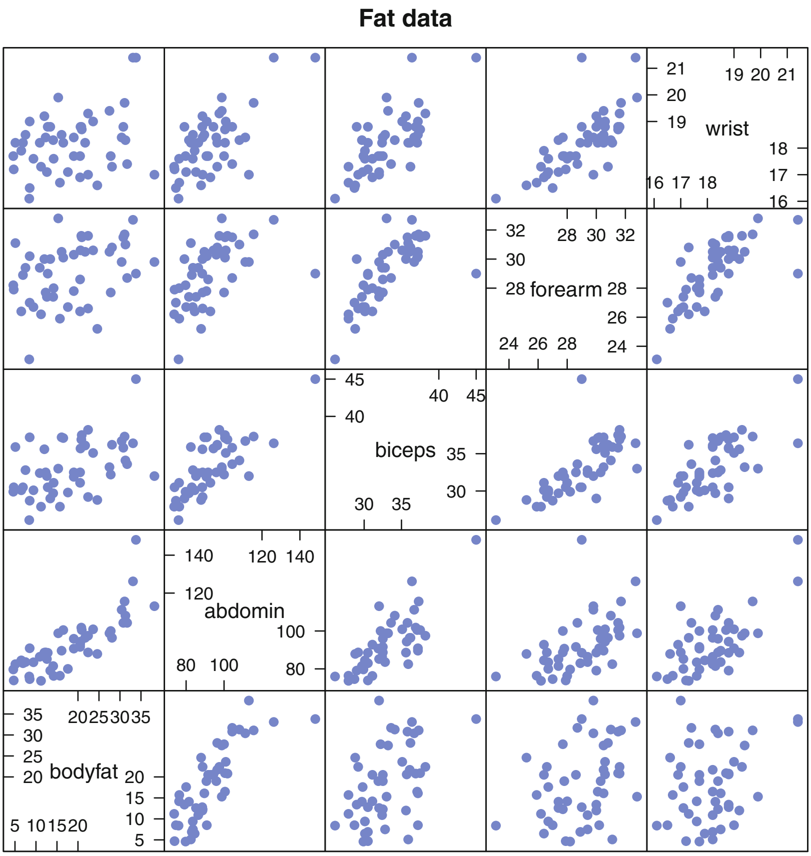


**Fig. 8.1** Body Fat Data

The response variable `bodyfat` is in the bottom row of the plot. We can see that a linear fit makes sense against `abdomin`. A linear relationship between `bodyfat` and the other predictor variables is also visible in the plot, but is weaker. All the predictor variables show correlation with each other.

### *8.2.4 One-X Analysis*

The initial analysis will look at just `bodyfat` and `abdomin`. We will come back to the other variables later. We expand the `bodyfat ~ abdomin` panel of Figure 8.1 in the left column of Figure 8.2 and place two straight lines on the graph in the two rightmost columns. The line in column 3 is visibly not a good fit. It is too shallow and is far above the points in the lower left. The line in column 2, labeled "least-squares fit", is just right. The criterion we use is *least squares*, which means that the sum of the squared differences from the fitted to observed points is to be minimized. The *least-squares* line is the straight line that achieves the minimum.

The top row of Figure 8.2 displays the vertical differences from the fitted to observed points. The bottom row displays the squares of the differences from the fitted to observed points. The least-squares line minimizes the sum of the areas of these squares. It is evident that the sum of the squared areas in column 2 is smaller than the sum of squared areas for the badly fitting line in column 3.

From any of these panels it is apparent that on average, body fat is directly related to abdominal circumference. As will be explained in Section 8.3.5, the least-squares line in Figure 8.2 can be used to predict `bodyfat` from `abdomin`. Note that although it is mathematically correct to say that `abdomin` increases with `bodyfat`, this is a misleading statement because it implies an unlikely direction of causality among these variables.

## 8.3 Simple Linear Regression

### *8.3.1 Algebra*

Figure 8.2 illustrates the least-squares line that best fits `bodyfat` to `abdomin`. Now that we see from the bottom row of the figure that the least-squares line actually does minimize the sum of squares, let us review the mathematics behind the calculation of the least-squares line. The standard notation we use for the least-squares straight line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{8.3}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the *regression coefficients*. We define the residuals by

$$e_i = y_i - \hat{y}_i \tag{8.4}$$

We wish to find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the expression for the sum of squares of the calculated residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i)\Big)^2 \tag{8.5}$$

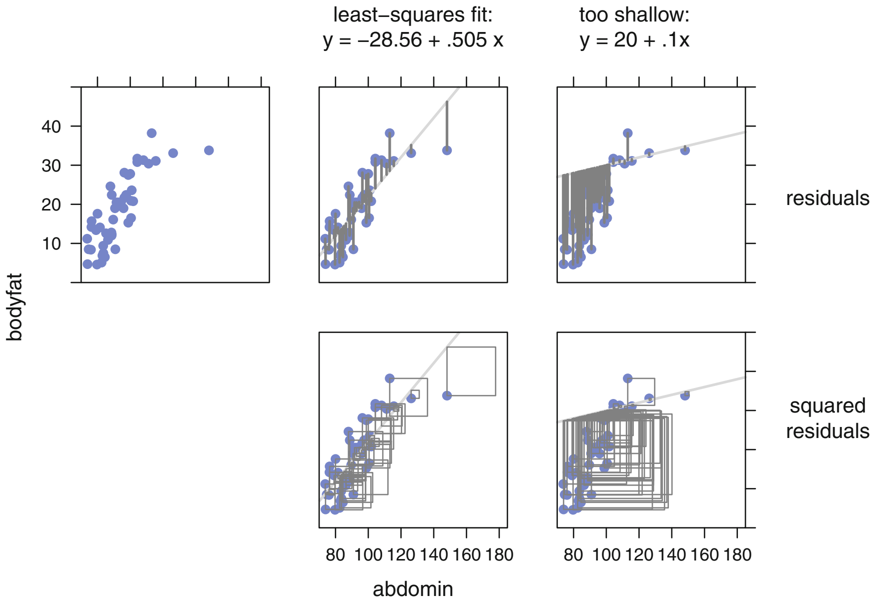


**Fig. 8.2** One $X$-variable and two straight lines. The second column is the least-squares line, the third is too shallow. Row 1 shows the residuals. Row 2 shows the squared residuals. The least-squares line minimizes the sum of the squared residuals.

We minimize by differentiation with respect to the parameters $\beta_0$ and $\beta_1$, setting the derivatives to 0 (thus getting what are called the *normal equations*)

$$\begin{aligned} \frac{\partial}{\partial\beta_0} \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i)\Big)^2 &= \sum_{i=1}^{n} 2\Big(y_i - (\beta_0 + \beta_1 x_i)\Big)(-1) = 0 \\ \frac{\partial}{\partial\beta_1} \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i)\Big)^2 &= \sum_{i=1}^{n} 2\Big(y_i - (\beta_0 + \beta_1 x_i)\Big)(-x_i) = 0 \end{aligned} \tag{8.6}$$

and then solving simultaneously for the regression coefficients

$$\begin{aligned} \hat\beta_1 &= \frac{\sum (y_i - \bar y)(x_i - \bar x)}{\sum (x_i - \bar x)^2} \\ \hat\beta_0 &= \bar y - \hat\beta_1 \bar x \end{aligned} \tag{8.7}$$

In addition to minimizing the sum of squares of the calculated residuals, $\hat\beta_0$ and $\hat\beta_1$ have the property that the sum of the calculated residuals is zero, i.e.,

$$\sum_{i=1}^{n} e_i = 0 \tag{8.8}$$

We request a proof of this assertion in Exercise 8.9.

For two or more predictor variables, the procedure (equating derivatives to zero) is identical but the algebra is more complex. We postpone details until Section 9.3.

### 8.3.2 Normal Distribution Theory

Under the normality assumption (8.2) for the residuals of Model (8.1), the least-squares estimates are also maximum likelihood estimates. This is true because if the residuals are normally distributed, their likelihood function is maximized when Equation (8.5) is minimized.

In Model (8.1), the unknown population variance of the $\epsilon_i$, $\sigma^2$, is estimated by the sample variance

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \tag{8.9}$$

Because the sample variance is proportional to the residual sum of squares in Equation (8.5), minimizing the sample variance also leads us to the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ in Equations (8.7). The square root $s$ of the sample variance in Equation (8.9), variously termed the *standard error of estimate*, the *standard error*, or the *root mean square error*, indicates the size of a typical vertical deviation of a point from the calculated regression line.

### 8.3.3 Calculations

The results of the statistical analysis are displayed in several tables, primarily the *ANOVA* (analysis of variance) table, the table of regression coefficients, and the table of other statistics shown in Table 8.1. These tables are fundamental to our interpretation of the analysis. The formulas for each number in these tables appear in Tables 8.2, 8.3, and 8.4. As with Tables 6.2 and 6.2, the ANOVA table in Section 8.1 does not include the "Total" line and the interpretation in Table 8.2 does include the "Total" line. R does not print the Total line in its ANOVA tables.

For one-$x$ regression (this example), there is usually only one null and alternative hypothesis of interest:

$$H_0{:}\,\beta_1 = 0 \quad \text{vs} \quad H_1{:}\,\beta_1 \neq 0 \tag{8.10}$$

Both $t = 9.297$ in the table of coefficients and $F = 86.427 = 9.297^2 = t^2$ in the ANOVA table are tests between those hypotheses. The associated $p$-value ($p = .5{\scriptstyle 10}^{-12}$, which we report as $< 0.0001$), is smaller than any reasonable $\alpha$ (the traditional .05 or .01, for example). Therefore, we are justified in rejecting the null

**Table 8.1** ANOVA table and table of regression coefficients for the simple linear regression model with $y$=`bodyfat` and $x$=`abdomin`.

```
> data(fat)

> fat.lm <- lm(bodyfat ~ abdomin, data=fat)

> anova(fat.lm)
Analysis of Variance Table

Response: bodyfat
          Df Sum Sq Mean Sq F value  Pr(>F)
abdomin    1   2440    2440    86.4 4.9e-12 ***
Residuals 45   1271      28
---
Signif. codes:
0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> summary(fat.lm)

Call:
lm(formula = bodyfat ~ abdomin, data = fat)

Residuals:
   Min     1Q Median     3Q    Max
-12.42  -4.11   1.21   3.52   9.65

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.5601     5.1100   -5.59  1.3e-06 ***
abdomin       0.5049     0.0543    9.30  4.9e-12 ***
---
Signif. codes:
0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 5.31 on 45 degrees of freedom
Multiple R-squared:  0.658,Adjusted R-squared:  0.65
F-statistic: 86.4 on 1 and 45 DF,  p-value: 4.85e-12
```

hypothesis in favor of the alternative. Inference on $\beta_0$ frequently makes no sense. In this example, for example, $\beta_0$ is the expected `bodyfat` of an individual having the impossible `abdomin` with zero circumference.

The Total line in the ANOVA table shows the sum of squares and degrees of freedom for the response variable `bodyfat` around its mean. When we divide these two numbers we recognize the formula $\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1) = 80.678$ as Equation (3.6) for the sample variance of the response variable. The goal of the analysis is to *explain* as much of the variance in the response variable as possible with a model

**Table 8.2** Interpretation of items in "ANOVA Table" from Table 8.1. The symbols in the `abdomin` section are subscripted `Reg`, short for "Regression". In this setting, "Regression" refers to the group of all model predictors. In this example there is only one predictor, `abdomin`.

| Name | Notation | Formula | Value in Table 8.1 |
|---|---|---|---|
| Total | | | |
| Sum of Squares | $\mathsf{SS}_{\text{Total}}$ | $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathsf{SS}_{\text{Reg}} + \mathsf{SS}_{\text{Res}}$ | 3711.199 |
| Degrees of Freedom | $\mathsf{df}_{\text{Total}}$ | $n - 1$ | 46 |
| Variance about Mean | | $\mathsf{SS}_{\text{Total}} / \mathsf{df}_{\text{Total}}$ | 80.678 |
| Residual | | | |
| Sum of Squares | $\mathsf{SS}_{\text{Res}}$ | $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ | 1270.699 |
| Degrees of Freedom | $\mathsf{df}_{\text{Res}}$ | $n - 2$ | 45 |
| Mean Square | $\mathsf{MS}_{\text{Res}}$ | $\hat{\sigma}^2 = s^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$ | 28.238 |
| abdomin | | | |
| Sum of Squares | $\mathsf{SS}_{\text{Reg}}$ | $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ | 2440.500 |
| Degrees of Freedom | $\mathsf{df}_{\text{Reg}}$ | number of predictor variables | 1 |
| Mean Square | $\mathsf{MS}_{\text{Reg}}$ | variability in $\hat{y}$ attributable to $\hat{\beta}_1$ <br> $\left(\dfrac{\text{abdomin Sum of Squares}}{\text{abdomin Degrees of Freedom}}\right)$ | 2440.500 |
| $F$-Value | $F_{\text{Reg}}$ | $\left(\dfrac{\text{abdomin Mean Square}}{\text{Residual Mean Square}}\right)$ | 86.427 |
| Pr(> $F$) | $p_{\text{Reg}}$ | $P(F_{1,45} > 86.427) = 1 - \mathcal{F}_{1,45}(86.427)$ | < 0.0001 |

that relates the response to the predictors. When we have explained the variance, the *residual* (or leftover) mean square $s^2$ is much smaller than the sample variance of the response variable.

The *coefficient of determination*, also known as *Multiple* $R^2$, usually accompanies ANOVA tables. This measure, generally denoted $R^2$, is the proportion of variation in the response variable that is accounted for by the predictor variable(s). It is desirable that $R^2$ be as close to 1 as possible. Models with $R^2$ considerably below 1 may be acceptable in some disciplines. The defining formula for $R^2$ is

$$R^2 = \frac{\mathsf{SS}_{\text{Reg}}}{\mathsf{SS}_{\text{Total}}} \tag{8.11}$$

In regression models with only one predictor, an alternative notation is $r^2$. This notation is motivated by the fact that $r^2$ is the square of the sample correlation

**Table 8.3** Interpretation of items in “Table of Regression Coefficients” from Table 8.1.

| Name | Notation | Formula | Value in Table 8.1 |
|---|---|---|---|
| (Intercept) | | | |
| Value | $\hat{\beta}_0$ | $\bar{y} - \hat{\beta}_1 \bar{x}$ | –28.560 |
| Standard Error | $\hat{\sigma}_{\beta_0}$ | $\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$ | 5.110 |
| $t$-value | $t_{\beta_0}$ | $\frac{\hat{\beta}_0}{\hat{\sigma}_{\beta_0}}$ | –5.589 |
| Pr(> \|$t$\|) | $p_{\hat{\beta}_0}$ | $P(\, t_{45} > \lvert -5.589 \rvert \,)$ | < 0.0001 |
| abdomin | | | |
| Value | $\hat{\beta}_1$ | $\frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$ | 0.505 |
| Standard Error | $\hat{\sigma}_{\beta_1}$ | $\hat{\sigma} / \sqrt{\sum(x_i - \bar{x})^2}$ | 0.054 |
| $t$-value | $t_{\beta_1}$ | $\hat{\beta}_1 / \hat{\sigma}_{\beta_1}$ | 9.297 |
| Pr(> \|$t$\|) | $p_{\hat{\beta}_1}$ | $P(t_{45} > \lvert 9.297 \rvert)$ | < 0.0001 |

coefficient $r$ between the response and predictor variable. $r$ is the usual estimate of the population correlation coefficient defined and interpreted in Equation (3.14). A formula for the sample correlation $r$ is

$$r = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(x_i - \bar{x})}} \tag{8.12}$$

It can be shown that $-1 \le r \le 1$. If $r = \pm 1$, then $x$ and $y$ are perfectly linearly related, directly so if $r = 1$ and inversely so if $r = -1$. The arithmetic sign of $r$ matches the arithmetic sign of $\hat{\beta}_1$.

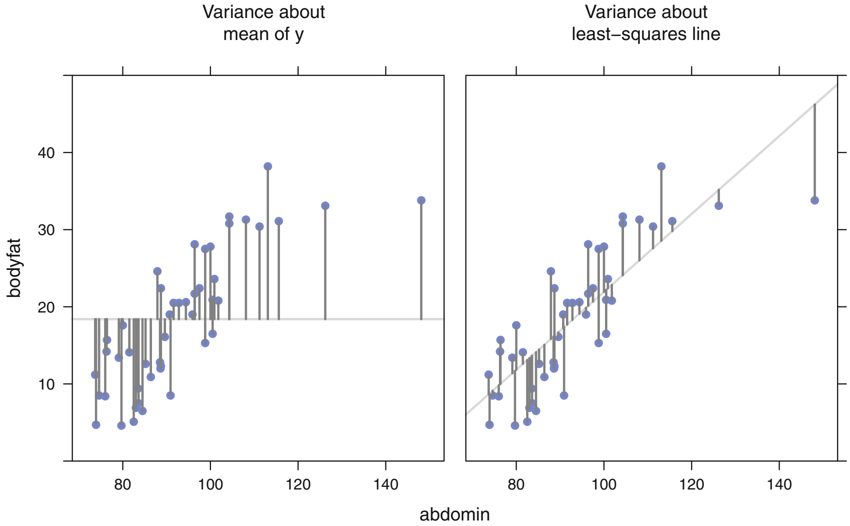
In the present body fat example, we find $r = 0.811$ and $r^2 = 0.658$. This value of $r$ is consistent with the moderately strong positive linear relationship between `bodyfat` and `abdomin` in the least-squares fit shown in Figure 8.2. Continuing with this example, the estimated response variance ignoring the predictor is 80.678 and the estimated response variance paying attention to the predictor `abdomin`, the Residuals Mean Square, is 28.238. Graphically, we see in Figure 8.3 that the variance estimate 80.678 about the mean belongs to Figure 8.3a and the variance estimate 28.238 about the regression line belongs to Figure 8.3b.

**Table 8.4** Interpretation of additional items, some of which are shown in Table 8.1.

| Name | Notation | Formula | Value based on Table 8.1 |
|---|---|---|---|
| Coefficient of Determination | | | |
| Multiple $R^2$ | $R^2$ | $\left(\dfrac{\text{abdomin Sum of Squares}}{\text{Total Sum of Squares}}\right)$ | 0.6576 |
| | p | Number of predictor $x$ variables in the model | 1. |
| | | in the model | |
| Adjusted $R^2$ | $R^2_{\text{adj}}$ | $1 - \left(\dfrac{n-1}{n-p-1}\right)(1-R^2)$ | 0.6500 |
| Dependent Mean | $\bar{Y}$ | $\dfrac{\sum Y_i}{n}$ | 18.3957 |
| Residual Standard Error | $\hat{\sigma} = s$ | $\sqrt{s^2}$ | 5.3139 |
| Coefficient of Variation | $cv$ | $s/\bar{Y}$ | 28.8867 |

While these two estimates of response variance are intuitive, they are not actually the statistically correct numbers to compare because they are not independent. The Total Sum of Squares is the sum of the Residuals Sum of Squares and the `abdomin` Sum of Squares. These two components of the Total Sum of Squares are independent and are therefore the base for the correct quantities to compare. The `abdomin` mean square is an unbiased estimate of $\sigma^2$ if $H_0$ is true but an overestimate of $\sigma^2$ if $H_0$ is false. The Residuals Mean Square is unbiased for $\sigma^2$ in either case. Therefore, the ratio of these two mean squares will tend to be close to 1 if $H_0$ is true but greater than 1 otherwise. With the assumption of independent normally distributed $\epsilon_i$, the ratio, given as the $F$-Value = 86.427 in the table, follows a (central) $F$ distribution with 1 and 45 degrees of freedom if $H_0$ is true, but not otherwise. Appeal to this distribution tells us whether the ratio is significantly greater than 1. When the observed Pr(> $F$) value in the table (in this case < 0.0001) is small, we interpret that as evidence that $H_0$ is false.

The formal statement of the test is: Under the null hypothesis that $\beta_1 = 0$ (that is, that information about $x$=`abdomin` gives no information about $y$=`bodyfat`), the probability of observing an $F$-value as large as the one we actually saw (in this case 86.427) is very small (in this case the probability is less than 0.0001). This very small $p$-value (assuming $H_0$ is true) is very strong evidence that $H_0$ is not true, that is, it is evidence that $\beta_1 \neq 0$. We will therefore act as if $H_0$ is false and take further actions as if the relationship of the fitted regression model actually explains what is going on.



a. $\sum(y - \bar{y})^2 = \text{SS}_{\text{Total}} = 3711.199$ b. $\sum(y - \hat{y})^2 = \text{SS}_{\text{Residual}} = 1270.699$

**Fig. 8.3** Variance about mean and about least-squares line.

The estimate $\hat{\beta}_1$ from Equation (8.7) can be rewritten as a weighted sum of $y_i$-values or of single-point slopes $\hat{\beta}_{1i} = (y_i - \bar{y})/(x_i - \bar{x})$

$$\hat{\beta}_1 = \sum_i (y_i - \bar{y}) \left( \frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \right) \tag{8.13}$$

$$= \sum_i \left( \frac{y_i - \bar{y}}{x_i - \bar{x}} \right) \left( \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right) \tag{8.14}$$

Figure 8.4 illustrates equation 8.14 with the R command

```
demo("betaWeightedAverage", ask=FALSE).
```

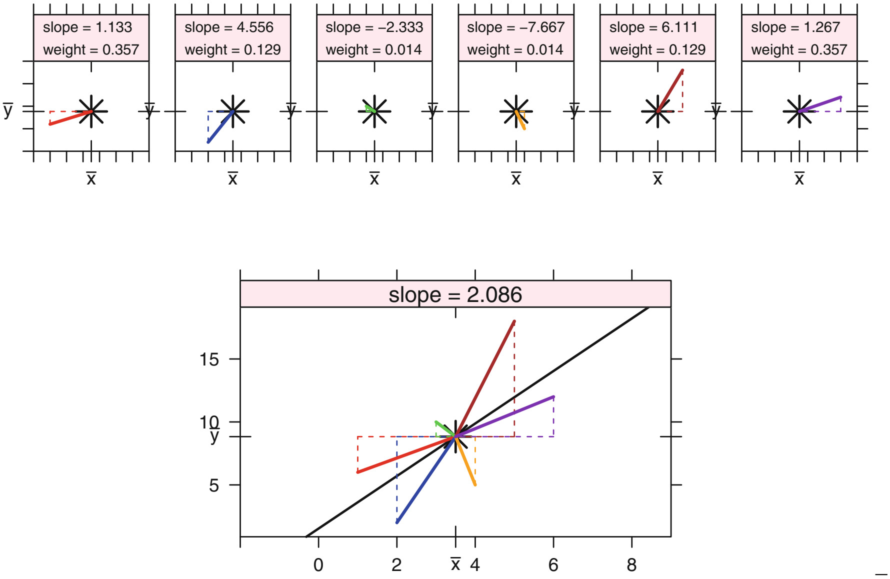
The variance of $\hat{\beta}_1$

$$\sigma^2_{\hat{\beta}_1} = \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \tag{8.15}$$

is constructed from the sum in Equation (8.13) with formulas based on Equation (3.9) (see Exercise 8.7). The sample estimate of the standard error of $\hat{\beta}_1$ is

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}} \tag{8.16}$$

Under $H_0$, and with the assumption of independent normally distributed $\epsilon_i$, the $t$-ratio $t_{\hat{\beta}_1} = \hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1}$ has a $t_{45}$ distribution allowing us to use the $t$ table in our tests. It



```
> bWA
   color x  y
1    red 1  6
2   blue 2  2
3  green 3 10
4 orange 4  5
5  brown 5 18
6 purple 6 12
```

**Fig. 8.4** Equation 8.14 shows that the slope $\hat{\beta}_1$ can be written as the weighted sum of the single-point slopes $\hat{\beta}_{1i} = (y_i - \bar{y})/(x_i - \bar{x})$. The top set of panels shows the set of single-point slopes. The bottom panel shows all six single-point slopes and the regression line whose slope is the weighted sum of the individual slopes. The dataset for this example is displayed.

follows from this that a $100(1 - \alpha)\%$ confidence interval on $\beta_1$ is

$$\hat{\beta}_1 \pm t_{df,\frac{\alpha}{2}}\ \hat{\sigma}_{\hat{\beta}_1}$$

where $df = \mathsf{df}_{\text{Res}}$ degrees of freedom.

Similarly, we can show (see Exercise 8.8)

$$\sigma^2_{\hat{\beta}_0} = \text{var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right) \tag{8.17}$$

### 8.3.4 Residual Mean Square in Regression Printout

The *residual mean square* is also called the *error mean square*. It is called *residual* because it is the variability in the response variable left over after fitting the model. It is called *error* because it is a measure of the difference between the model and the data. We prefer the term “residual” and discourage the term “error” because the term “error” suggests a mistake, and that is not the intent of this component of the analysis. Nevertheless, on occasion we use the term “error” as a synonym for “residual” to match the continued use by SAS of “Error Mean Square” rather than our preferred “Residual Mean Square”. See Table 8.5 for a comparison of several notations. See Tables 8.5 and 8.6 for illustrations of how the fitted values and the residuals are related to the various sums of squares used in the ANOVA table. These tables show the linear and quadratic identities introduced in Section 6.A.

### 8.3.5 New Observations

One of the uses of a fitted regression equation is to make inferences about new observations. A new observation $y_0$ at $x_0$ has the model

$$y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0 = \mu_0 + \epsilon_0$$

where

- $y_0$ is a single unobserved value
- $x_0$ is a the value of the predictor $x$ at the new observation
- $\beta_0$ and $\beta_1$ are the regression coefficients.

The concepts that we introduce here extend, almost without change, to the multiple regression setting of Chapter 9. We therefore preview the slightly more elaborate notation of Chapter 9. The model in Equation (8.1) can be rewritten in matrix notation as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{8.18}$$

$$\underset{n\times 1}{Y} = \underset{n\times(1+p)}{X} \; \underset{(1+p)\times 1}{\beta} + \underset{n\times 1}{\epsilon}$$

We restrict $p = 1$ in Chapter 8. More generally, beginning in Chapter 9, $p$ is a positive integer.

**Table 8.5** Residual Mean Square in Regression Printout. The "Residual Mean Square" and "Error Mean Square" are two names for the same concept. Note that the (Std Err Residual)$_i$ is different for each $i$. It is smallest for $x$ values closest to $\bar{x}$ and increases as the $x$ values move away from $\bar{x}$. This is the reason that the confidence bounds for the regression line (see Figure 8.5) show curvature.

For each observation $i$ the standard regression printout shows

$$\begin{array}{ccccc}
\widehat{\text{var}}(\hat{\mu}_i) & + & \widehat{\text{var}}(e_i) & = \widehat{\text{var}}(y_i) = & \hat{\sigma}^2 \\
h_i\hat{\sigma}^2 & + & (1-h_i)\hat{\sigma}^2 & = & \hat{\sigma}^2 \\
(\text{Std Err Predict})_i^2 & + & (\text{Std Err Residual})_i^2 & \multicolumn{2}{l}{= \text{Residual Mean Square}} \\
 & & & \multicolumn{2}{l}{= \text{Error Mean Square}}
\end{array}$$

```
> h <- hat(model.matrix(fat.lm))

> pred <- predict(fat.lm, se.fit=TRUE)

> res <- resid(fat.lm)

> sigma.hat.square <- anova(fat.lm)["Residuals", "Mean Sq"]

> fat.predvalues <-
+ data.frame("y=bodyfat"=fat$bodyfat,  "x=abdomin"=fat$abdomin,
+            h=h,                       mu.hat=pred$fit,
+            e=res,                     var.mu.hat=h*sigma.hat.square,
+            var.resid=(1-h)*sigma.hat.square,
+            sigma.hat.square=sigma.hat.square,
+            se.fit=sqrt(h*sigma.hat.square),
+            se.resid=sqrt((1-h)*sigma.hat.square))

> fat.predvalues[1:3, 1:7]
  y.bodyfat x.abdomin       h mu.hat      e var.mu.hat var.resid
1      12.6      85.2 0.02762  14.46 -1.860     0.7800     27.46
2       6.9      83.0 0.03171  13.35 -6.450     0.8954     27.34
3      24.6      87.9 0.02399  15.82  8.776     0.6773     27.56

> ## fat.predvalues
>
> ## linear identity
> all.equal(rowSums(fat.predvalues[,c("mu.hat", "e")]),
+           fat$bodyfat,
+           check.names=FALSE)
[1] TRUE

> ## quadratic identity
> (SSqReg <- sum((fat.predvalues$mu.hat - mean(fat$bodyfat))^2))
[1] 2440

> (SSqRes <- sum(res^2))
[1] 1271

> (SSqTot <- sum((fat$bodyfat - mean(fat$bodyfat))^2))
[1] 3711

> all.equal(SSqReg + SSqRes, SSqTot)
[1] TRUE
```

**Table 8.6** We show the linear identity $y_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x}) + e_i$ and the quadratic identity $\sum(y_i - \bar{y})^2 = \sum(\beta_1 x_i)^2 + \sum \epsilon_i^2$ for least squares regression. The linear identity is the partitioning of the column of $y_i$ into columns for the grand mean, the product of the regression coefficient and the difference of $x_i$ from $\bar{x}$, and the column of residuals $e_i$. The quadratic identity is the arithmetic behind the sums of squares in the ANOVA table.

$$\begin{aligned} y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i && \text{for } i = 1, \ldots, n \quad \text{from Equation (8.1)} \\ &= (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i + e_i && \\ &= \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) + e_i && \text{linear identity} \end{aligned}$$

| $i$ | $y_i$ | $\bar{y}$ | $\beta_1(x_i - \bar{x})$ | $e_i$ |
|---|---|---|---|---|
| 1 | 12.6 | 18.4 | −3.935 | −1.860 |
| 2 | 6.9 | 18.4 | −5.046 | −6.450 |
| 3 | 24.6 | 18.4 | −2.572 | 8.776 |
| 4 | 10.9 | 18.4 | −3.329 | −4.166 |
| 5 | 27.8 | 18.4 | 3.538 | 5.866 |
| 6 | 20.6 | 18.4 | 0.710 | 1.494 |
| 7 | 19.0 | 18.4 | −1.158 | 1.762 |
| 8 | 12.8 | 18.4 | −2.269 | −3.327 |
| 9 | 5.1 | 18.4 | −5.299 | −7.997 |
| 10 | 12.0 | 18.4 | −2.218 | −4.177 |
| 11 | 7.5 | 18.4 | −4.743 | −6.153 |
| 12 | 8.5 | 18.4 | −1.057 | −8.839 |
| 13 | 20.5 | 18.4 | −0.704 | 2.808 |
| 14 | 20.8 | 18.4 | 4.447 | −2.042 |
| 15 | 21.7 | 18.4 | 1.720 | 1.584 |
| 16 | 20.5 | 18.4 | −0.098 | 2.202 |
| 17 | 28.1 | 18.4 | 1.720 | 7.984 |
| 18 | 22.4 | 18.4 | 2.275 | 1.729 |
| 19 | 16.1 | 18.4 | −1.714 | −0.582 |
| 20 | 16.5 | 18.4 | 3.790 | −5.686 |
| 21 | 19.0 | 18.4 | 1.468 | −0.863 |
| 22 | 15.3 | 18.4 | 2.932 | −6.028 |
| 23 | 15.7 | 18.4 | −8.379 | 5.683 |
| 24 | 17.6 | 18.4 | −6.561 | 5.765 |
| 25 | 14.2 | 18.4 | −8.429 | 4.233 |
| 26 | 4.6 | 18.4 | −6.712 | −7.083 |
| 27 | 8.5 | 18.4 | −9.288 | −0.608 |
| 28 | 22.4 | 18.4 | −2.168 | 6.172 |
| 29 | 4.7 | 18.4 | −9.641 | −4.055 |
| 30 | 9.4 | 18.4 | −4.794 | −4.202 |
| 31 | 12.3 | 18.4 | −2.168 | −3.928 |
| 32 | 6.5 | 18.4 | −4.289 | −7.607 |
| 33 | 13.4 | 18.4 | −7.015 | 2.020 |
| 34 | 20.9 | 18.4 | 3.790 | −1.286 |
| 35 | 31.1 | 18.4 | 11.415 | 1.289 |
| 36 | 38.2 | 18.4 | 10.152 | 9.652 |
| 37 | 23.6 | 18.4 | 3.992 | 1.212 |
| 38 | 27.5 | 18.4 | 2.932 | 6.172 |
| 39 | 33.8 | 18.4 | 27.825 | −12.421 |
| 40 | 31.3 | 18.4 | 7.628 | 5.276 |
| 41 | 33.1 | 18.4 | 16.767 | −2.063 |
| 42 | 31.7 | 18.4 | 5.709 | 7.595 |
| 43 | 30.4 | 18.4 | 9.193 | 2.811 |
| 44 | 30.8 | 18.4 | 5.709 | 6.695 |
| 45 | 8.4 | 18.4 | −8.581 | −1.415 |
| 46 | 14.1 | 18.4 | −5.804 | 1.508 |
| 47 | 11.2 | 18.4 | −9.742 | 2.546 |
| $\sum$ columns$_i^2$ | 19616 | 15905 | 2440 | 1271 |

$$\begin{aligned} \text{Total Sum of Squares} &= \sum y_i^2 - \sum \bar{y}^2 \\ &= \sum (y_i - \bar{y})^2 \\ &= 19616 - 15905 \\ &= 3711 \\ &= \sum (\beta_1 x_i)^2 + \sum \epsilon_i^2 \\ &= 2440 + 1271 \quad \text{quadratic identity} \end{aligned}$$

In the extended notation, a new observation $y_0$ at $x_{0+}$ has the model

$$y_0 = x_{0+}\beta + \epsilon_0 = \mu_0 + \epsilon_0$$

where

- $y_0$ is a single unobserved value
- $x_{0+}$ is a $1 \times (1 + p)$ row of predictors [$(1\ x_0)$ in Chapter 8]
- $\beta$ is a $(1 + p)$-vector of regression coefficients [$(\beta_0\ \beta_1)'$ in Chapter 8].

There are two related questions to ask about the new observation:

1. Estimate the parameter $\mu_0 = E(y_0) = x_{0+}\beta$.
2. Predict a specific observation $y_0 = \mu_0 + \epsilon_0$.

Estimation intervals for new $\mu_0$ and prediction intervals for new $y_0$ based on a new value $x_{0+}$ depend on the quantity $h_0$ defined as

$$h_0 = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{8.19}$$

The formula for $h_0$ is similar to the leverage formula for $h_i$ to be introduced in Equations (9.14) or (9.15), where the new value $x_{0+}$ replaces one of the observed values $X_{i+}$. The notation $i$ specifically means one of the original $n$ observations and the notation 0 means an additional observation that need not be one of the original ones. Equation (8.19) is specifically for simple linear regression ($p = 1$). The more complex formula in Equations (9.14) or (9.15) is needed when $p > 1$.

Answering the questions requires information about estimated variances:

1. Estimate the
   a. parameter $\mu_0 = E(y_0) = x_{0+}\beta$ with
   b. estimator $\hat{\mu}_0 = x_{0+}\hat{\beta}$,
   c. variance of the estimator $\text{var}(\hat{\mu}_0) = h_0\sigma^2$, and
   d. estimated variance of the estimator $\widehat{\text{var}}(\hat{\mu}_0) = h_0\hat{\sigma}^2$.
2. Predict
   a. a specific observation $y_0 = \mu_0 + \epsilon_0$ with
   b. predictor $\hat{y}_0 = \hat{\mu}_0 = x_{0+}\hat{\beta}$ (the same as the parameter estimate),

c. variance of the predictor $\text{var}(\hat{y}_0) = \text{var}(\hat{\mu}_0 + \epsilon_0) = \text{var}(\hat{\mu}_0) + \text{var}(\epsilon_0)$, and

d. estimated variance of the predictor $\widehat{\text{var}}(\hat{y}_0) = \widehat{\text{var}}(\hat{\mu}_0) + \widehat{\text{var}}(\epsilon_0) = h_0\hat{\sigma}^2 + \hat{\sigma}^2 = \hat{\sigma}^2(h_0 + 1)$.

In the special case that $x_{0+} = x_{i+}$ (one of the observed points), we have

$$\widehat{\text{var}}(\hat{y}_i) = (1 + h_i)\hat{\sigma}^2 = \hat{\sigma}^2 + \widehat{\text{var}}(\hat{\mu}_i)$$

Note that the (standard error)$^2$ for prediction $\hat{\sigma}^2(h_0 + 1)$ is larger than the (standard error)$^2$ for estimation $\hat{\sigma}^2 h_0$. A prediction interval for individual observations $\hat{y}_0$ estimates the range of observations that we might see. A confidence interval for the estimated mean of the new observations estimates the center point of the predicted range.

Most regression programs print the standard error for estimation of the mean: $\hat{\sigma}\sqrt{h_0}$, the confidence interval for estimating $\mu_0 = E(y_0|x)$: $\hat{y}_0 \pm t_{df,\frac{\alpha}{2}}\hat{\sigma}\sqrt{h_0}$, [also shown in Equation (9.24)], and the prediction interval for a new observation $(y_0|x)$: $\hat{y}_0 \pm t_{df,\frac{\alpha}{2}}\ \hat{\sigma}\sqrt{1 + h_0}$ [also shown in Equation (9.25)]. These items are discussed in detail in Section 9.9.

The commands that construct the confidence and prediction intervals in R, and their interpretation, are shown in Table 8.7. To see the standard error for prediction of a new observation, we must manually do the arithmetic

$$\hat{\sigma}^2 h_0 + \hat{\sigma}^2 = (1 + h_0)\hat{\sigma}^2 \tag{8.20}$$

The two questions about a new observation are actually familiar questions in a new guise. They are the same questions addressed in Section 3.6 about the location parameter $\mu$ of a sample from a single variable. We elaborate on the comparison in Table 8.8.

In both the confidence interval and the prediction interval of the regression problem in Table 8.8, the magnitude of (Standard Deviation)$^2$ increases as the new value $x$ moves further from the mean $\bar{x}$ of the existing $x_i$'s. This indicates that we have more confidence in a prediction for an $x$ in the vicinity of the $x_i$'s of the existing data than in an $x$ far from the $x_i$'s of the existing data. The lesson is that extrapolations of the fitted regression relationship for remote values of $x$ are likely to be unreliable.

Confidence and prediction intervals for a particular new observation at $x_0$ are shown in Table 8.7. These intervals can be extended to confidence and prediction *bands* by letting $x_0$ vary over the entire range of $x$. Figure 8.5 illustrates such 95% bands for `fat.lm`, the modeling of `bodyfat` as a function of `abdomin`, displayed in Table 8.1. The 0.95 probability statement applies to each particular value of $x = x_0$. It does not apply to statements that the bands enclose the infinite set of all possible means or predictions as $x$ varies over its range.

**Table 8.7** Construction of the confidence and prediction intervals for new observations in R. See also the discussion surrounding Equations (9.24) and (9.25).

```
> old.data <-
+     data.frame(y=rnorm(50), x1=rnorm(50), x2=rnorm(50), x3=rnorm(50))

> example.lm <- lm(y ~ x1 + x2 + x3, data=old.data)

> (example.coef <- coef(example.lm))
(Intercept)          x1          x2          x3
   -0.09670     0.11571    -0.12581    -0.09652

> (new.data <- data.frame(x1=3, x2=2, x3=45))
  x1 x2 x3
1  3  2 45

> predict(example.lm, newdata=new.data, se.fit=TRUE,
+         interval="confidence")
$fit
     fit    lwr   upr
1 -4.344 -18.03 9.337

$se.fit
[1] 6.797

$df
[1] 46

$residual.scale
[1] 0.9492


> predict(example.lm, newdata=new.data, se.fit=TRUE,
+         interval="prediction")
$fit
     fit    lwr  upr
1 -4.344 -18.16 9.47

$se.fit
[1] 6.797

$df
[1] 46

$residual.scale
[1] 0.9492


> c(1, data.matrix(new.data)) %*% example.coef
       [,1]
[1,] -4.344
```

**Table 8.8** Comparison of confidence and prediction intervals in the one-sample problem ($t$-test) and in the regression problem.

| | One Sample | Regression |
|---|---|---|
| Model Parameters: | | |
| Model | $y = \mu_Y + \epsilon$ | $y_x = \beta_0 + \beta_1 x + \epsilon$ |
| Parameter | $\mu_Y$ | $\mu_{YX} = \beta_0 + \beta_1 x$ |
| Variance of $\epsilon$ | $\text{var}(\epsilon) = \sigma^2_Y$ | $\text{var}(\epsilon) = \sigma^2_{YX}$ |
| Sample Statistics: | | |
| Estimate | $\hat{\mu}_Y = \bar{y}$ | $\hat{\mu}_{yx} = b_0 + b_1 x$<br>$\hat{y}_i = b_0 + b_1 x_i$ |
| Variance | $s^2_Y = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$ | $s^2_{YX} = \sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-2)$ |

Estimate Parameter:

(Standard Deviation)$^2$ for Confidence Interval Estimate

| | |
|---|---|
| What is the average height $\mu_Y$ of everyone? | What is the average height $\mu_{YX}$ of those people who are $x = 10$ years old? |
| $s^2_{\hat{\mu}_Y} = s^2_{\bar{Y}} = \frac{s^2_Y}{n} = s^2_Y\left(\frac{1}{n}\right)$ | $s^2_{\hat{\mu}_{yx}} = s^2_{YX} h_x = s^2_{YX}\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$ |

Prediction Interval:

(Standard Deviation)$^2$ for Prediction Interval for an Individual Response

| | |
|---|---|
| How tall is the next person? | How tall is the next 10-year-old? |
| $\hat{y} = \hat{\mu}_Y + \epsilon = \bar{y} + \epsilon$ | $\hat{y}_x = \hat{\mu}_{yx} + \epsilon = (b_0 + b_1 x) + \epsilon$ |
| $s^2_{\hat{y}} = \frac{s^2_Y}{n} + s^2_Y = s^2_Y\left(\frac{1}{n} + 1\right)$ | $s^2_{\hat{y}_x} = s^2_{YX} h_x + s^2_{YX} = s^2_{YX}(1 + h_x)$<br>$= s^2_{YX}\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$ |

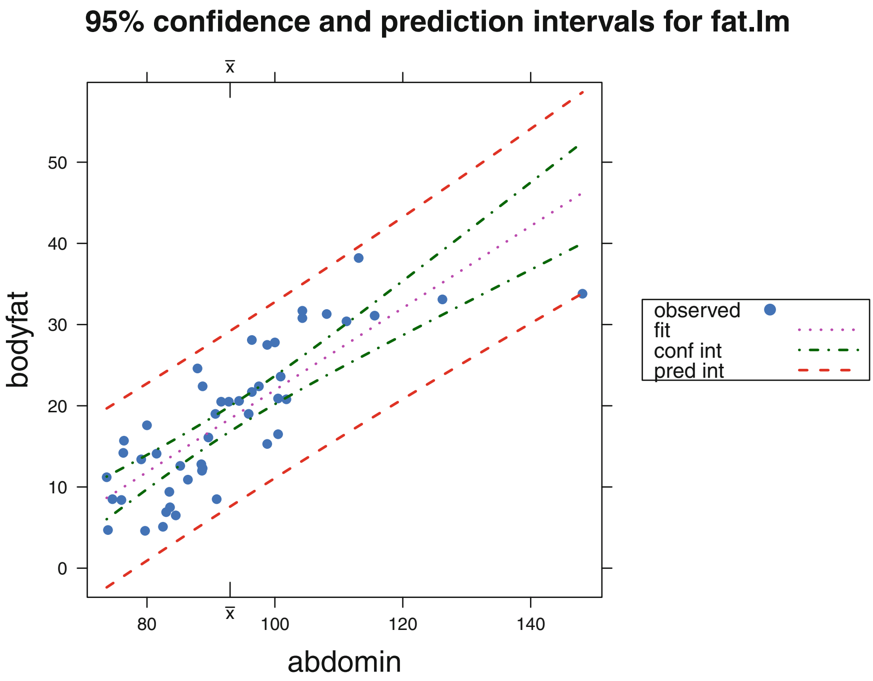


**Fig. 8.5** Confidence and prediction bands for modeling `bodyfat ~ abdomin`, body fat data. The widths of these bands are minimized at $x = \bar{x}$ because $h_0$ is minimized at $x = \bar{x}$.

## 8.4 Diagnostics

There are two steps to a statistical analysis. The first step is to construct a model and estimate its parameters. Sections 8.3.2 and 9.2 discuss estimation of the parameters of linear models with one and two predictor variables. The second step is to study the quality of the fit of the data to that model and determine if the model adequately describes the data. This section introduces the diagnostics. They are investigated more thoroughly in Section 11.3.

The choice of diagnostic techniques is connected directly to the model and assumptions. If the assumption (8.2) that the error terms $\epsilon_i$ are normally independently distributed with constant mean 0 and variance $\sigma^2$ is valid, then the residuals $e_i = (y_i - \hat{y}_i)$ will be approximately normally distributed. More precisely, the $n$ values $e_i$ will behave exactly like $n$ numbers independently chosen from the normal distribution and subjected to $p + 1$ linear constraints. In the simplest case, when $p = 0$ (one-sample $t$-test in Chapter 5), the residuals $e_i$ behave like $n$ independent normals centered on their observed mean $\bar{x}$. For simple linear regression ($p = 1$), the residuals behave like $n$ independent normals vertically centered on a straight line specified by the two estimated parameters $\hat{\beta}_1$ and $\hat{\beta}_0$.

The diagnostic techniques are various procedures for looking at approximately normal numbers and seeing if they display any systematic behavior. If we see systematic behavior, then we conclude that the model did not capture all the interesting features of the data. We iterate the analysis steps by trying to model the systematic behavior we just detected and then looking at the residuals from the newer model.

Figure 8.6 shows several diagnostic plots from the simple regression model of Section 8.1. These are our versions of standard plots of the Fitted Values and Residuals from the regression analysis. The first three panels are based on the output of the R statement `plot(fat.lm)` (using the `plot.lm` method in the **stats** package). The fourth is based on an S-Plus plot. All four as displayed here were drawn with the statement `lmplot(fat.lm)` using the `lmplot` function in the **HH** package.

(We show plots from `plot(fat.lm)` in Figure 11.18. We prefer the **lattice**-based appearance of our first three plots to the **base** graphics of `plot(fat.lm)`. We believe the fourth panel of Figure 11.18 (enlarged in Figure 11.19) can't be described until Chapter 11. We believe the fourth panel of Figure 8.6 is highly informative and wish that R had included it as part of their standard display.)

We discuss each panel in turn, with the numbering sequence $\binom{13}{24}$.

1. Panels 1 and 2 are coordinated. Panel 1 is a plot of the Residuals $e = y - \hat{y}$ against the Fitted Values $\hat{y}$ along with a horizontal line at $e = 0$. The horizontal line corresponds to the least-squares fit of the Residuals against the Fitted Values. There is, by construction, no linear effect in this panel. There may be quadratic (or higher-order polynomial) effects visible. The marginal distribution of the Fitted Values $\hat{y}$ may show patterns that need further investigation. When there is only one $x$-variable, as in the example in Figure 8.6, the Fitted Values are a linear transformation of the $x$-variable. In this example, we see that the $x$-value of the point with the largest absolute residual is noticeably larger than any of the other $x$-values.

2. Panel 2 plots $\sqrt{|e|} = \sqrt{|\text{Residuals}|}$ against the Fitted Values $\hat{y}$. It shows much of the same information as Panel 1. The absolute value folds the negative residuals onto the positive direction in order to emphasize magnitude of departure from the model at the expense of not showing direction. The square root transformation brings in the larger residuals and spreads out the smaller ones. See the discussion of the ladder of powers in Section 4.9 for more information on the effects of transformations. In this display we chose to retain the original directionality by choice of plotting symbol and color.

3. Panel 3 is a normal probability plot with the Residuals on the vertical axis and the normal quantiles on the horizontal axis. The diagonal line has the standard deviation $s$ for its slope. When the residuals are approximately normal, the points will be close to the diagonal line. Asymmetries in the residuals will be visible. Short tails in the distribution of the residuals will be visible as an "S"-shaped

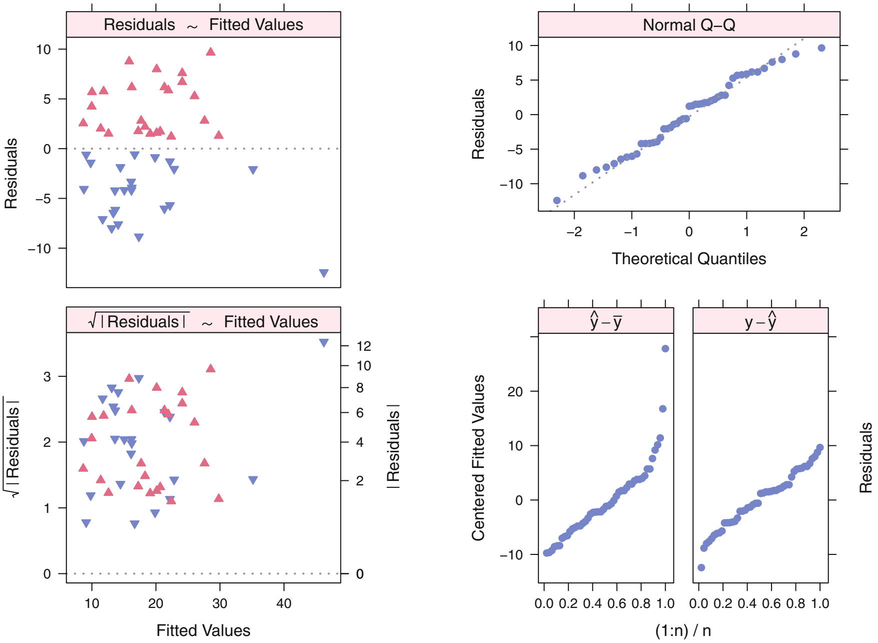


**Fig. 8.6** Diagnostics for `lm(bodyfat ~ abdomin, data=fat)`. Diagnostic Plots of the Residuals and Fitted Values from a regression analysis. See Section 8.4 for an extensive discussion of each of the four panels in this display. On the left we show two views of the Residuals plotted against the Fitted Values, with the Residuals themselves on the top, and the square root of the absolute values of the Residuals on the bottom. On the top right, we show the QQ plot of the Residuals against the Normal quantiles. On the right bottom, we show the *r-f spread plot*—a two-panel display of the transposed empirical distributions of the Centered Fitted Values and of the Residuals (see Section 8.5).

display, and long tails in the distribution of the residuals (seen as vertical outliers in panels 1 and 2) will be visible as a mirror-image “Ƨ” shape. See Section 5.8 for further discussion of probability plots.

4. Panel 4 is subdivided into two transposed empirical distributions. The left panel shows the Centered Fitted Values $\hat{y} - \bar{y}$ and the right panel shows the Residuals $y - \hat{y}$. The relative vertical ranges of these two panels gives some information on the multiple correlation coefficient $R^2$. We develop the construction and interpretation of panel 4 in Section 8.5 and Figure 8.7.

## 8.5 ECDF of Centered Fitted Values and Residuals

The ECDF plot of Centered Fitted Values and Residuals is the *r-f spread plot* defined by Cleveland (1993). The empirical distribution of $S(x)$ is defined in Section 5.7 as the fraction of the data that is less than or equal to $x$. The empirical distribution is defined analogously to the cumulative distribution $F(x) = P(X \le x)$ of a theoretical distribution.

We discuss each of the panels of Figure 8.7.

a. The plot of the cumulative distribution is a plot of $F(x)$ against $x$.

b. The empirical cumulative distribution of an observed set of data is a plot of proportion($X \le x$) against $x$. If there are $n$ observations in the dataset, we plot $i/n$ against $x_{[i]}$. We use the convention here that subscripts in square brackets mean that the data have been sorted. For example, let us look at the fitted values $\hat{y}$ and residuals $e = y - \hat{y}$ from the regression analysis in Table 8.1. The left side of Figure 8.7b is the cumulative distribution of the fitted values. The right side is the cumulative distribution of the residuals. Note that these plots are on very different scales for the abscissa and therefore cannot easily be compared visually.

c. We construct Figure 8.7c by making two adjustments to Figure 8.7b. First, we center the fitted values on their mean. Second, we plot both graphs on the same abscissa scale by forcing them to have the same $x$-axis constructed as the range of the union of their individual abscissas.

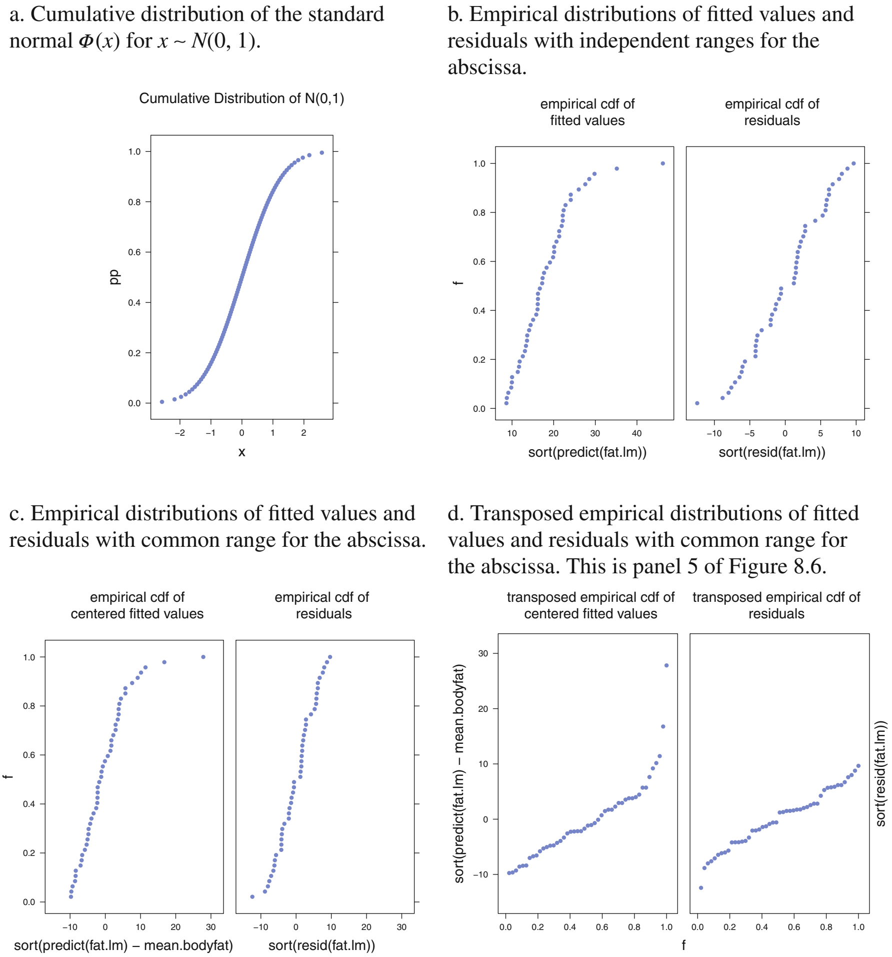
d. Figure 8.7d is the transpose of the pair of graphs in Figure 8.7c. We interchange the axes, putting the proportions on the abscissa and the data (centered fitted values in the left panel and residuals in the right panel) on the ordinate. We therefore force the $y$-axes to have a common limits. S-Plus uses Figure 8.7d as the fifth diagnostic plot of their analog of Figure 8.6. The vertical axis now uses the same $y$ units as panels 1 and 3 of Figure 8.6.

If our model explains the data well, then we would anticipate that the residuals have less variability than the fitted values.

The multiple correlation $R^2$ can be written as

$$R^2 = \frac{\mathsf{SS}_{\text{Reg}}}{\mathsf{SS}_{\text{Total}}} = \frac{\mathsf{SS}_{\text{Reg}}}{\mathsf{SS}_{\text{Reg}} + \mathsf{SS}_{\text{Res}}} \tag{8.21}$$

We can use the squared range of the fitted values as a surrogate for the $\mathsf{SS}_{\text{Reg}}$ and the squared range of the residuals as a surrogate for the $\mathsf{SS}_{\text{Res}}$. This leads to the interpretation of panel 4 of Figure 8.6 as an indicator of $R^2$. We show a series of illustrations of this interpretation in Figure 8.8. If the ranges of the $\hat{y} - \bar{y}$ and $y - \hat{y}$ panels are similar, then $R^2 \approx \frac{1}{2}$. If the range of the fitted values is larger, then the $R^2$ is closer to 1, and if the range of the fitted values is smaller, then the $R^2$ is closer to 0.



a. Cumulative distribution of the standard normal $\Phi(x)$ for $x \sim N(0, 1)$.

b. Empirical distributions of fitted values and residuals with independent ranges for the abscissa.

c. Empirical distributions of fitted values and residuals with common range for the abscissa.

d. Transposed empirical distributions of fitted values and residuals with common range for the abscissa. This is panel 5 of Figure 8.6.

Cumulative Distribution of N(0,1)
pp
x
empirical cdf of fitted values
empirical cdf of residuals
f
sort(predict(fat.lm))
sort(resid(fat.lm))
empirical cdf of centered fitted values
empirical cdf of residuals
sort(predict(fat.lm) − mean.bodyfat)
sort(resid(fat.lm))
transposed empirical cdf of centered fitted values
transposed empirical cdf of residuals
sort(predict(fat.lm) − mean.bodyfat)
sort(resid(fat.lm))
f

**Fig. 8.7** Explanation of panel 4 of Figure 8.6. Panels a,b,c are empirical distribution plots and panel d is the transposed empirical distribution plot of the fitted values and residuals from the linear regression `fat.lm <- lm(bodyfat ~ abdomin, data=fat)`. Please see the discussion in the text of Section 8.5 for more detailed description of the panels in this figure.

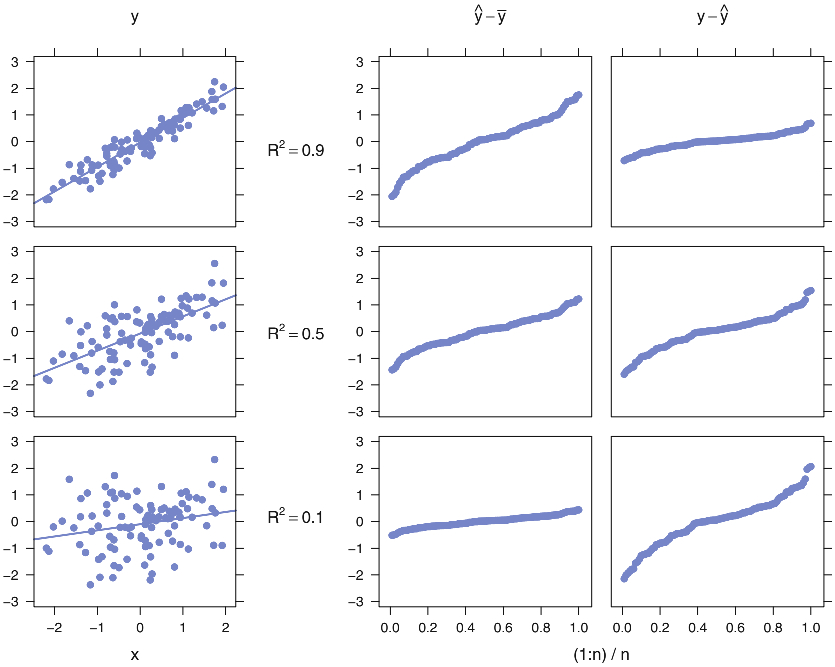


**Fig. 8.8** There are three columns $y$, $\hat{y} - \bar{y}$, and $y - \hat{y}$. The rows of the $y$ column shows a plot of $y$ against $x$ along with the fitted regression line for each of three levels of $R^2$ (.1, .5, .9). The $\hat{y} - \bar{y}$ and $y - \hat{y}$ columns show the transposed ECDF of the Fitted Values and Residuals for those situations. For $R^2 = .1$, the Residuals $y - \hat{y}$ has a wider range than the Centered Fitted Values $\hat{y} - \bar{y}$. For $R^2 = .5$, the two ranges are equal. For $R^2 = .9$, the Residuals $y - \hat{y}$ has a narrower range than the Centered Fitted Values $\hat{y} - \bar{y}$.

## 8.6 Graphics

The figures in this chapter represent several different types of plots.

Figure 8.1 is a scatterplot matrix, constructed in R with `splom()`.

Figures 8.2 and 8.3 use `regrresidplot`, a function in the **HH** package in R. Our function `panel.residSquare`, used by `regrresidplot`, constructs the squares that represent the squared residuals with real squares on the plotting surface. The heights of the squares are in $y$-coordinates. The widths of the squares are the same number of inches (or cm) on the plotting surface as the heights. Each of the figures has been placed into a lattice structure which enforces the same $x$- and $y$-ranges for comparability. Our function `regrresidplot` is based on the explanation of least-squares regression in Smith and Gonick (1993).

Figure 8.5 is a scatterplot drawn with **HH** function `ci.plot` with superimposed lines for the fitted regression line and the confidence and prediction intervals.

Figures 8.6, 8.7, and 8.8 use functions in the **HH** package that are based on the R function `plot.lm` to display the standard plots of Residuals and Fitted Values from a regression analysis. The ECDF plots of Centered Fitted Values and Residuals are drawn by the **HH** function `diagplot5new`, which is based on the S-Plus function `rfplot`, which in turn is based on a plot by Cleveland (1993).

## 8.7 Exercises

**8.1.** Hand et al. (1994) report on a study by Lea (1965) that investigated the relationship between mean annual `temperature` (degrees F) in regions of Britain, Norway, and Sweden, and the rate of `mortality` from a type of breast cancer in women. The data are accessed as `data(breast)`.

a. Plot the data. Does it appear that the relationship can be adequately modeled by a linear function?
b. Estimate the regression line and add this to your plot.
c. Calculate and interpret $R^2$.
d. Calculate and interpret the standard error of estimate.
e. Interpret the estimated slope coefficient in terms of the variables `mortality` and `temperature`.
f. Find a 95% confidence interval on the population slope coefficient.
g. Find a 95% prediction interval for a region having mean annual temperature 45.
h. One of these 16 data points is unusual compared to the others. Describe how.

**8.2.** Shaw (1942), later in Mosteller and Tukey (1977), shows the `level` of Lake Victoria Nyanza relative to a standard level and the number of `sunspots` in each of 20 consecutive years. The data are accessed as `data(lake)`. Use linear regression to model the lake level as a function of the number of sunspots in the same year.

**8.3.** Does muscle mass decrease with age? The `age` in years and muscle `mass` were obtained from 16 women. The data come from Neter et al. (1996) and are accessed as `data(muscle)`.

a. Plot `mass` vs `age` and overlay the fitted regression line.
b. Interpret the slope coefficient in terms of the model variables.
c. Predict with 90% confidence the muscle mass of a 66-year-old woman.
d. Interpret the calculated standard error of estimate.
e. Interpret $R^2$ in terms of the model variables.

**8.4.** The dataset `data(girlht)` contains the heights (in cm) at ages 2, 9, and 18 of 70 girls born in Berkeley, California in 1928 or 1929. The variables are named `h2`, `h9`, and `h18`, respectively. The data come from a larger file of physical information on these girls in Cook and Weisberg (1999).

a. Regress `h18` on `h9` and also `h18` on `h2`.

b. Discuss the comparative strengths of these two regression relationships.

c. Interpret the slope coefficients of both regressions.

**8.5.** We would expect that the price of a diamond ring would be closely related to the size of the diamond the ring contains. Chu (1996) presents data on the `price` (Singapore dollars) of ladies' diamond rings and the number of `carats` in the ring's diamond. The data are accessed as `data(diamond)`.

a. Regress `price` on `carats`.

b. Notice that the estimated intercept coefficient is significantly less than 0. Therefore, this model is questionable, although the range of the predictor variables excludes 0. Instead fit a model without an intercept term.

c. Compare the goodness of fits of the two models. Which is preferable?

**8.6.** The data `data(income)`, from Bureau of the Census (2001), contains year 2000 data on the percentage of `college` graduates and per capita personal `income` for each of the 50 states and District of Columbia. Regress `income` on `college`. Interpret the meaning of $R^2$ for these data. Discuss which states have unusually low or high per capita income in relation to their percentage of college graduates.

**8.7.** Prove Equation (8.15)

$$\sigma^2_{\hat\beta_1} = \text{var}(\hat\beta_1) = \frac{\sigma^2}{\sum(x_i - \bar x)^2}$$

The proof is primarily algebraic manipulation. Rewrite (8.13) as a weighted sum of the independent $y_i$, that is as

$$\hat\beta_1 = \sum(y_i - \bar y)\left(\frac{(x_i - \bar x)}{\sum(x_i - \bar x)^2}\right) = \sum y_i k_i \tag{8.22}$$

then write

$$\text{var}(\hat\beta_1) = \sigma^2 \sum k_i^2 \tag{8.23}$$

and simplify.

**8.8.** Prove Equation (8.17) that the variance of the estimate of the intercept $\hat\beta_0$ has variance

$$\sigma^2_{\hat\beta_0} = \text{var}(\hat\beta_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)$$

**8.9.** Algebraically prove the assertion in Equation (8.8) that in simple regression, the sum of the calculated residuals is zero.

**8.10.** In Figure 8.2 we construct the actual squares of the residuals and show that the sum of the areas of the squared residuals is smallest for the least-squares line. We do the construction in the simplest way, placing the other three sides on the side that is already there representing the residual. Other possibilities are

a. Place the left–right center of the square on the residual line. Use the function `panel.residSquare` as the model for your function.

b. Place a circle (a real circle in inches of graph surface) on the points. Base your function on the functions `panel.residSquare` and the descriptions of the R `points` function (`?points`). The value `pch=1` provides a circle. You can use the `cex` argument to control the size of the circles.

   Option 1: Keep the existing residual line and center the circle on the observed point.

   Option 2: Use the existing residual line as the diameter of the circle.