Chapter 9: Multiple regression

The simple linear model is extended from a model that describes the mean response \( \mu(y|x) \) as a function of one explanatory variable to \( p - 1 \) explanatory variables, \( x_1, \ldots, x_{p-1} \). The multiple regression model is

\[
\mu(y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}.
\]

There are \( p \) parameters in the linear regression model: \( \beta_0, \beta_1, \ldots, \beta_{p-1}. \)

Meadowfoam is a small herbaceous winter annual adapted to poorly drained wet soils and hence, common in the lowlands of the Pacific Northwest. The seed oil is very stable because it is comprised of long chain fatty acids, and consequently, it’s used in industrial applications, cosmetic and hair-care applications, among others. Seddigh and Jolliff\(^2\) investigated optimal growing conditions for commercial production in a controlled experiment. Using growth chambers, they measured seed production by controlling light intensity (6 levels), and timing of light treatment onset—either at floral induction (late), or 24 days prior to floral induction (early).

Ten seedlings were randomly assigned to each of the 12 = 6 \( \times \) 2 treatments and the average number of flowers per plant was recorded at the end of the experiment. The experiment was replicated so that data consist of \( n = 24 \) observations evenly distributed among the 12 treatments. The experiment and data are summarized in Table 1.

Table 1: Factors, levels and mean number (mean of 10 plants) of flowers observed for each treatment.

<table>
<thead>
<tr>
<th>Timing</th>
<th>150</th>
<th>300</th>
<th>450</th>
<th>600</th>
<th>750</th>
<th>900</th>
</tr>
</thead>
<tbody>
<tr>
<td>late</td>
<td>62.3</td>
<td>44.3</td>
<td>49.6</td>
<td>39.4</td>
<td>31.3</td>
<td>36.8</td>
</tr>
<tr>
<td></td>
<td>77.4</td>
<td>54.2</td>
<td>61.9</td>
<td>45.7</td>
<td>44.9</td>
<td>41.9</td>
</tr>
<tr>
<td>early</td>
<td>77.8</td>
<td>69.1</td>
<td>57</td>
<td>62.9</td>
<td>60.3</td>
<td>52.6</td>
</tr>
<tr>
<td></td>
<td>75.6</td>
<td>78.0</td>
<td>71.1</td>
<td>52.2</td>
<td>45.6</td>
<td>44.4</td>
</tr>
</tbody>
</table>

\(^1\)If we define the constant \( x_0 = 1 \) so that \( x_{0,1} = \cdots = x_{0,n} = 1 \) then we can express the multiple regression model as

\[
\mu(y|x) = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} = \sum_{i=0}^{p-1} x_i \beta_i.
\]

The analysis of these data demands a new model since the previous models only accommodated a single explanatory variable. The one-way analysis of variance is appropriate if the timing of light treatment onset were the only explanatory variable, whereas the simple linear regression models is appropriate if timing is ignored and light intensity is the only explanatory variable. In this situation, however, the effect of the variables are to be analyzed simultaneously. Jumping ahead, the figure below and left shows the data and the fitted regression model. The multiple regression model consists of two parallel responses, in other words, two lines describe the production of flowers. It appears that early onset yields more flowers than late onset, and that flower production declines with increasing light intensity. It also appears that the model fits the data reasonably well.

The Ginzberg data set are observations made on patients hospitalized for depression. Three variables were measured on each patient: depression (Beck self-report depression scale), simplicity: a score of a subject’s need to see the world in black and white, and fatalism score. Both explanatory variables together explain $R^2 = 56.7\%$ of the variation in depression score. The figure above and right plots depression score against fatalism score and shows simplicity score by color-coding the points according to whether the simplicity score was greater than the third quartile $Q_3$, between $Q_1$ and $Q_3$, or less than $Q_1$. 

55
Estimates of \( \mu(y|x) \) are plotted against fatalism score; each line is a set of estimates for a fixed value of simplicity (either \( Q_1 \), \( Q_2 \) or \( Q_3 \)). The fitted model implies that depression score is not an additive combination of fatalism and simplicity since the lines are not parallel. The figure immediately above shows the fitted model as a smooth surface. The surface consists of estimates of the expected depression score given the fatalism and simplicity score.

**The multiple regression model**

The regression of \( y \) on \( x_1, x_2, \ldots, x_{p-1} \) is an equation that describes the expected value of \( y \) as a linear combination of \( 1, x_1, x_2, \ldots, x_{p-1} \) where the coefficients are unknown parameters \( \beta_0, \beta_1, \ldots, \beta_{p-1} \):

\[
\mu(y|x_1, x_2, \ldots, x_{p-1}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} = \beta_0 + \sum_{i=1}^{p-1} \beta_i x_i.
\]

There are very few situations in which there is a correct regression model in the sense that there is some particular combination of explanatory variables that describes the mean or expected value of \( y \) better than all other choices of explanatory variables. Every model is wrong, but some models are less wrong than others (which is to say, they more accurately describe the expected value of \( y \)). Some examples of multiple linear regression models are

\[
\begin{align*}
\mu(y|x_1, x_2, \ldots, x_{p-1}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\
\mu(y|x_1, x_2) &= \beta_1 x_1 + \beta_2 x_2 \\
\mu(y|x_1, x_2) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \\
\mu(y|x_1, x_2) &= \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2).
\end{align*}
\]

An example of a nonlinear multiple regression model is \( \mu(y|x_1, x_2) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \).

So far, the multiple linear regression model describes only the expected value of the response variable \( y \). For inferential purposes, the model needs to more completely describe the distribution of the response variable. The full model is very similar to the simple linear regression model. It is

\[
y \overset{\text{iid}}{\sim} N(\mu(y|x_1, \ldots, x_{p-1}), \sigma) \quad \text{where} \quad \mu(y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}.
\]

This model states that the variance of \( y \) is constant, *provided* that the expected value of mean is correctly specified. Ramsey and Schafer express this condition by writing

\[
\text{var}(y|x_1, \ldots, x_{p-1}) = \sigma^2.
\]
An equivalent specification of the model is

\[ y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon \]

where \( \varepsilon \overset{\text{iid}}{\sim} N(0, \sigma) \).

The residuals about the fitted regression model must be realizations of independent random variables \( \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \). A compact model specification relevant to the data is

\[ y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i, \]

where \( \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma), i = 1, 2, \ldots, n \).

**Regression surfaces** Mathematically, a linear model describes a sub-space of \( \mathbb{R}^{p+1} \). For example, a model with two explanatory variables \( x_1 \) and \( x_2 \) defines a plane in three-space, or \( \mathbb{R}^3 \). There are three axes: \( y, x_1 \) and \( x_2 \). The regression plane intersects the \( y \) axis where \( x_1 = x_2 = 0 \). The height of the plane at the intersection is

\[ \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 = \beta_0. \]

The coefficient \( \beta_1 \) is the slope of the plane if \( x_2 \) is fixed at any particular value, and \( \beta_2 \) is the slope of the plane if \( x_1 \) is fixed at any particular value. If more than two explanatory variables are present in the model, then it is difficult to develop a geometrical interpretation. For example, a model with three variables describes a three-dimensional object embedded in 4-space.

**Interpretation of regression coefficients** A coefficient, or parameter estimate for a multiple linear regression model is interpreted as the effect of the associated explanatory variable if all other variables are held fixed.

The effect of a one unit increase of \( x_1 \) on \( \mu(y|x_1, \ldots, x_{p-1}) \), when all other variables are held fixed is

\[ \mu(y|x_1 + 1, x_2 \ldots, x_{p-1}) - \mu(y|x_1, x_2 \ldots, x_{p-1}) = \beta_0 + \beta_1 (x_1 + 1) + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} \]

\[ - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}) = \beta_1. \]

For example, the parameter associated with light intensity in the meadowfoam experiment was estimated to be \( \hat{\beta}_1 = -0.040471 \), and so if timing of onset is held fixed (either early or late), then a one unit (\( \mu\text{mol/m}^2/\text{sec} \)) increase in light intensity unit is estimated to reduce the number of flowers by 0.0405. If light intensity is fixed at any values 150, 300, 450, 600, 750, 900, then it is estimated that there will be \( \hat{\beta}_2 = 12.15 \) more flowers if the timing of light is early compared to late.
The effect of a 1 unit increase in $x_1$ does not depend on the value of either variable. The effect of a simultaneous increase of a one unit change of both $x_1$ and $x_2$ is sum of the coefficients $\beta_1$ and $\beta_2$ since

$$
\mu(y|x_1 + 1, x_2 + 1 \ldots, x_{p-1}) - \mu(y|x_1, x_2 \ldots, x_{p-1}) = \beta_0 + \beta_1(x_1 + 1) + \beta_2(x_2 + 1) + \\
- (\beta_0 + \beta_1 x_1 + \beta_2 x_2) = \beta_1 + \beta_2
$$

For this reason, the model $\mu(y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$ is called an additive model.

The parameter estimates (coefficients) from the simple linear regression of depression score on fatalism are shown in Table 2; Table 3 shows the coefficient from the multiple linear regression of depression score on fatalism and simplicity. The coefficients associated with fatalism differ between models, in part because simplicity is correlated with fatalism ($r = .631$).

The interpretation of the parameter estimate associated with fatalism differs between models because the estimated change in mean depression score associated with a one unit change in fatalism is numerically different, and because the parameter estimate from the multiple regression model describes the estimated effect provided simplicity is held fixed.

Table 2: Coefficients and standard errors obtained from the linear regression of weekly numbers of depression score on fatalism. $R^2 = .432$.

| Variable | Coefficient | Std. Error | t-statistic | $P(T > |t|)$ |
|----------|-------------|------------|-------------|-------------|
| Intercept | .3426       | .0941      | 3.64        | .0005       |
| Fatalism  | .6574       | .0842      | 7.80        | < .0001     |

Table 3: Coefficients and standard errors obtained from the linear regression of weekly numbers of depression score on fatalism and simplicity. $R^2 = .519$.

| Variable | Coefficient | Std. Error | t-statistic | $P(T > |t|)$ |
|----------|-------------|------------|-------------|-------------|
| Intercept | .2027       | .0947      | 2.14        | .0354       |
| Fatalism  | .4178       | .1006      | 4.15        | < .0001     |
| Simplicity | .3795     | .1006      | 3.77        | .0003       |

The simple linear regression model estimates the effect of a one unit increase in fatalism to be an increase in depression score of .657. No control of simplicity is assumed. Because fatalism and simplicity are positively correlated, an increase in fatalism is associated with an increase in simplicity (for most or patients), so the effect of a one unit increase in fatalism without fixing simplicity has a greater effect than a one unit increase in fatalism when
fixing simplicity. Analyzing the effect of fatalism with simplicity fixed is unrealistic, and in general, the interpretation of regression coefficients obtained from observational studies is not as clear as in experimental studies.

**Constructed explanatory variables**

The utility of multiple linear regression is greatly expanded by specially constructed explanatory variables. Three types of constructed variables are presented below.

**Power functions of $x$ for curvature**

Galileo conducted an experiment to determine if the horizontal velocity of a moving object is constant.\(^3\)

He rolled a ball off of an incline and measured the horizontal distance traveled. Galileo postulated that, because of the accelerative effect of gravity, the horizontal distance traveled by a ball would follow a parabolic curve; in other words, a second-order polynomial model of the form $y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ where $x$ is the height of the incline and $y$ is the distance traveled. The data, and a second-order polynomial (or quadratic) model fit using multiple linear regression is shown to the right. It appears as if the experiment supports Galileo’s conjecture.

The relationship between distance and height is extended from a linear model to a second-order polynomial model by including an additional variable $x_2$ constructed from the squared distances. Specifically, if $x_{1,i}$ denotes the $i$th distance, then $x_{2,i} = x_{1,i}^2$. For brevity and clarity, the constructed variable $x_2$ usually is denoted as $x^2$. The fitted multiple regression model is

$$\hat{\mu}(y|x) = 199.9 + .708x - .0003x^2.$$ 

Including $x^2$ in the regression model allows the fitted model to bend in one direction. Including both $x^2$ and $x^3$ in the regression model allows the fitted model to bend in two directions, and the inclusion of $k$ powers of $x$ allows the model to bend in $k - 1$ directions.

---

There is little point in attempting to interpret individual coefficients of a polynomial model since the effect of a one unit change in \( x \) depends on \( x \). For example, the effect of a one unit change in \( x \) is

\[
\beta_1(x + 1) + \beta_2(x + 1)^2 - (\beta_1 x + \beta_2 x^2) = \beta_1 + \beta_2(2x + 1).
\]

The situation is even worse if more polynomial terms are included in the model. Since the interpretive value of the model is effectively lost by when a second- or higher-order polynomial is fit, polynomial models should only be used when there is some scientific motivation or if the model is to be used primarily for prediction (so that there is little interest in interpreting the coefficients).\(^4\)

**Indicator variables**

Indicator variables are used when the population (and data) are believed to originate from several groups or sub-populations. In the first applications discussed herein, the mean \( \mu(y|x) \) is believed to follow the same model for each group except that each group has its own intercept (the slope is the same for each group). The meadowfoam flower production data illustrates the situation (recall the first Figure above). The two groups are plants that received the light treatment at floral induction, and at 24 days prior to floral induction.

To set up a model that has a common slope for both groups and a separate intercept for each group, an indicator variable is defined according to

\[
x_{\text{early}} = \begin{cases} 
0, & \text{if onset of light treatment was late} \\
1, & \text{if onset of light treatment was early}
\end{cases}
\]

The variable \( x_{\text{light}} \) is light intensity and the model is

\[
\mu(y|x) = \beta_0 + \beta_1 x_{\text{light}} + \beta_2 x_{\text{early}}
\]

If \( y_i \) was observed when the onset of light treatment was late, then the model is

\[
\mu(y_i|x_{\text{light},i}, x_{\text{early},i}) = \beta_0 + \beta_1 x_{\text{light},i} + \beta_2 \times 0 \\
= \beta_0 + \beta_1 x_{\text{light},i}.
\]

On the other hand, if \( y_i \) was observed when the onset of light treatment was early, then the model is

\[
\mu(y_i|x_{\text{light},i}, x_{\text{early},i}) = \beta_0 + \beta_1 x_{\text{light},i} + \beta_2 \times 1 \\
= (\beta_0 + \beta_2) + \beta_1 x_{\text{light},i} \\
= \beta_0^* + \beta_1 x_{\text{light},i}
\]

---

\(^4\)Improving the fit of the model is not generally an acceptable reason for including higher-order terms.
where $\beta_0 = \beta_0 + \beta_2$. If $\beta_2 = 0$, then $\beta_0 = \beta_0$ and the timing of onset of light treatment has no effect on the mean number of flowers. Parameter estimates and tests of significance are shown in the following table.

Table 4: Summary of the fitted model showing parameter estimates and tests of significance obtained from the meadowfoam data set. The model is $\mu(y|x) = \beta_0 + \beta_1 x_{\text{light}} + \beta_2 x_{\text{early}}$. In addition, $\hat{\sigma} = 6.441$, df = 21, $R^2 = 0.799$.

| Parameter Estimate | Std. Error | $t$  | $Pr(T>|t|)$ |
|--------------------|------------|------|-------------|
| $\beta_0$          | 71.30      | 3.274| 21.8        | $\leq .0001$ |
| $\beta_1$          | -.0405     | .0051| -7.88       | $\leq .0001$ |
| $\beta_2$          | 12.16      | 2.630| 4.62        | .0001        |

Table 4 shows strong evidence of a difference in numbers of flowers per plant attributable to timing of onset ($t_{21} = 4.62$, p-value = .0001), and that numbers of flowers decreases as light level increases ($t_{21} = -7.88$, p-value $\leq .0001$). Specifically, initiating light treatment 24 days before the onset of floral induction is estimated to increase the expected number of flowers per plant by 12.16, and a 95% confidence interval for the (true) difference is [6.70, 17.61].

It has been assumed, without justification, that the effect of light on flower production is linear. The experimental design provides the opportunity to test whether the linear model fit is comparable to a less-constrained alternative model that specifies a separate mean level for each of the 6 light levels. Light is treated as a factor in this model. The model is set up using indicator variables of the form

$$
x_{300} = \begin{cases} 
0, & \text{if light intensity } \neq 300 \\
1, & \text{if light intensity } = 300 \\
\vdots
\end{cases}$$

$$
x_{900} = \begin{cases} 
0, & \text{if light intensity } \neq 900 \\
1, & \text{if light intensity } = 900 \\
\end{cases}$$

It is not necessary to use an indicator variable for the first level of light (150) since this level is identified by all other indicator variables being 0 in value.

The separate means model is constructed by entering all of these variables in the model and removing $x_{\text{light}}$, the quantitative light variable. The separate means model is

$$
\mu(y|x_{\text{early}}, x_{300}, \ldots, x_{900}) = \beta_0 + \beta_1 x_{\text{early}} + \beta_2 x_{300} + \cdots + \beta_6 x_{900}.
$$

Table 5 below shows the indicator variables for the linear regression of flower production on timing and treating light as a factor. The R function call `model.matrix(lm(y~A+B))` will
produce all indicator variables used in the regression of $y$ on $A$ and $B$ regardless of whether the explanatory variables are factors or quantitative variables. Thus, it is not necessary to construct indicator variables for a factor—R will construct them automatically when a factor is specified as an explanatory variable in a linear regression model.

In Table 5, there are one fewer indicator variables than levels for the factors. Normally, the fewest (but sufficient) number of indicator variables are used, and that number is $I - 1$, where $I$ is the number of levels of the factor.

Table 5: Parameters and variables for the meadowfoam regression analysis treating light as a factor.

<table>
<thead>
<tr>
<th>Timing</th>
<th>Light</th>
<th>Constant (1)</th>
<th>$x_{early}$</th>
<th>$x_{300}$</th>
<th>$x_{450}$</th>
<th>$x_{600}$</th>
<th>$x_{750}$</th>
<th>$x_{900}$</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late</td>
<td>150</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\beta_0$</td>
</tr>
<tr>
<td>Late</td>
<td>300</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\beta_0 + \beta_2$</td>
</tr>
<tr>
<td>Late</td>
<td>450</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\beta_0 + \beta_3$</td>
</tr>
<tr>
<td>Late</td>
<td>600</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$\beta_0 + \beta_4$</td>
</tr>
<tr>
<td>Late</td>
<td>750</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$\beta_0 + \beta_5$</td>
</tr>
<tr>
<td>Late</td>
<td>900</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>$\beta_0 + \beta_6$</td>
</tr>
<tr>
<td>Early</td>
<td>150</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\beta_0 + \beta_1$</td>
</tr>
<tr>
<td>Early</td>
<td>300</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\beta_0 + \beta_1 + \beta_2$</td>
</tr>
<tr>
<td>Early</td>
<td>450</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$\beta_0 + \beta_1 + \beta_3$</td>
</tr>
<tr>
<td>Early</td>
<td>600</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$\beta_0 + \beta_1 + \beta_4$</td>
</tr>
<tr>
<td>Early</td>
<td>750</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>$\beta_0 + \beta_1 + \beta_5$</td>
</tr>
<tr>
<td>Early</td>
<td>900</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\beta_0 + \beta_1 + \beta_6$</td>
</tr>
</tbody>
</table>

The meaning of each parameter can be deduced:

$\beta_0 = \mu(y|x)$ when timing is late and light intensity is 150

$\beta_1 = \mu(y|early) - \mu(y|late)$ when light intensity is fixed

$\beta_2 = \mu(y|light intensity = 300) - \mu(y|light intensity = 150)$ when timing is fixed

$\vdots$

$\beta_6 = \mu(y|light intensity = 900) - \mu(y|light intensity = 150)$ when timing is fixed

The parameter estimates are shown in Table 6.
Table 6: Coefficients and standard errors obtained from the linear regression of light (treated as a factor) and timing. The model is \( \mu(y|x_{\text{early}}, x_{300}, \ldots, x_{900}) = \beta_0 + \beta_1 x_{\text{early}} + \beta_2 x_{300} + \cdots + \beta_6 x_{900} \).

| Parameter | Estimate | Std. Error | t-statistic | \( P(T > |t|) \) |
|-----------|----------|------------|-------------|----------------|
| \( \beta_0 \) | 67.20    | 3.629      | 18.5        | < .0001        |
| \( \beta_1 \) | 12.16    | 2.743      | 4.43        | .0003          |
| \( \beta_2 \) | -9.125   | 4.751      | -1.92       | .0717          |
| \( \beta_3 \) | -13.37   | 4.751      | -2.81       | .0119          |
| \( \beta_4 \) | -23.22   | 4.751      | -4.89       | .0001          |
| \( \beta_5 \) | -27.75   | 4.751      | -5.84       | < .0001        |
| \( \beta_6 \) | -29.35   | 4.751      | -6.18       | < .0001        |

The figure (right) shows the estimates \( \hat{\mu}(y|x) \) from the separate means model. The observations are identified by small points and the separate means estimates are identified by larger points. The estimates for each group closely follow a line, and so it’s reasonable to expect a lack-of-fit test to show no evidence that the linear model fails to fit the data.

**Interaction**

Two explanatory variables \( x_1 \) and \( x_2 \) interact if the effect of \( x_1 \) on \( \mu(y|x) \) depends on the level of \( x_2 \).\(^5\) There is convincing evidence that fatalism interacts with simplicity (Table 7).

Table 7: Coefficients and standard errors obtained from the linear regression of depression score on fatalism and simplicity and the interaction of fatalism and simplicity. \( R^2 = .567 \).

| Variable           | Coefficient | Std. Error | t-statistic | \( P(T > |t|) \) |
|--------------------|-------------|------------|-------------|----------------|
| Intercept          | 0.296       | .192       | -1.544      | .127           |
| Fatalism           | .826        | .169       | 4.90        | < .0001        |
| Simplicity         | .937        | .212       | 4.42        | < .0001        |
| Fatalism \( \times \) Simplicity | -.404       | .137       | -2.95       | .00421         |

\(^5\)If the effect of \( x_1 \) on \( \mu(y|x) \) depends on the level of \( x_2 \), then the effect of \( x_2 \) on \( \mu(y|x) \) depends on the level of \( x_1 \).
The interaction variable is denoted as Fatalism × Simplicity since it is constructed by multiplying fatalism and simplicity:

\[ x_{i, \text{Fatalism} \times \text{Simplicity}} = x_{i, \text{Fatalism}} \times x_{i, \text{Simplicity}}, \quad i = 1, 2, \ldots, n. \]

For brevity, let \( x_1 \) denote fatalism, \( x_2 \) denote simplicity and \( x_3 = x_1 \times x_2 \) denote the interaction of fatalism and simplicity. The effect of fatalism (when interaction is present) is measured, as before, by the change in depression score associated with a one unit increase in fatalism. The model estimate of the effect is

\[
\mu(y|x_1 + 1, x_2) - \mu(y|x_1, x_2) = \hat{\beta}_0 + \hat{\beta}_1(x_1 + 1) + \hat{\beta}_2x_2 + \hat{\beta}_3(x_1 + 1)x_2
\]

\[= (\hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \hat{\beta}_3x_1x_2)\]

\[= \hat{\beta}_1 + \hat{\beta}_3x_2.\]

A one unit change in fatalism changes depression score by \( \triangle y = .826 - .404 \times x_{\text{Simplicity}} \) since \( \hat{\beta}_1 = .826 \) and \( \hat{\beta}_3 = -.404 \). If the value of simplicity is \( Q_1 = .647 \), then a one unit increase in fatalism is estimated to change depression score by \( \triangle y = .826 - .404 \times .647 = .565 \). If the value of simplicity is \( Q_3 = 1.269 \), then a one unit increase in fatalism is estimated to change depression score by .313.

Interaction is usually regarded as a nuisance since the interpretation of coefficients is more involved with interaction present than when it is not in the model. A formal test of significance for a interaction parameter (e.g., \( \hat{\beta}_3 \)) is sometimes desirable to reduce concerns that the fitted model is overly simplistic. Testing all possible interactions when there are many variables (particularly in observational studies) is usually unwise because of the problem of controlling the family-wise Type I error rate. Generally, there should be good reasons for introducing interaction variables into a model.

A question that often arises in problems that involve a quantitative variable and one or more factors is whether completely separate regression models are appropriate for each group (i.e., each level). Suppose, in the meadowfoam analysis, that light is treated as a quantitative variable. Then, separate regression lines can be created by constructing an interaction variable between light and each indicator variable of a factor level. As there are only two levels of the timing factor (early and late), there is only one indicator variable. The model is

\[
\mu(y|x) = \beta_0 + \beta_1x_{\text{light}} + \beta_2x_{\text{early}} + \beta_3x_{\text{interaction}}
\]

where \( x_{i, \text{interaction}} = x_{i, \text{light}} \times x_{i, \text{early}}, \quad i = 1, 2, \ldots, n, \) and

\[
x_{\text{early}} = \begin{cases} 
0, & \text{if onset of light treatment was late} \\
1, & \text{if onset of light treatment was early.} 
\end{cases}
\]

\(^6\)A test for evidence of interaction can be viewed as a lack-of-fit (or goodness-of-fit) test for the model without interaction.
For observations obtained when timing is late, the model is

\[
\mu(y|x) = \beta_0 + \beta_1 x_{\text{light}} + \beta_2 x_{\text{early}} + \beta_3 x_{\text{light}} \times x_{\text{early}}
\]

\[
= \beta_0 + \beta_1 x_{\text{light}} + \beta_2 \times 0 + \beta_3 x_{\text{light}} \times 0
\]

\[
= \beta_0 + \beta_1 x_{\text{light}}.
\]

For observations obtained when timing is early, the model is

\[
\mu(y|x) = \beta_0 + \beta_1 x_{\text{light}} + \beta_2 x_{\text{early}} + \beta_3 x_{\text{light}} \times x_{\text{early}}
\]

\[
= \beta_0 + \beta_1 x_{\text{light}} + \beta_2 \times 1 + \beta_3 x_{\text{light}} \times 1
\]

\[
= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_{\text{light}}.
\]

With this set-up, \(\beta_0\) is determined entirely by those observations made when timing is late and \(\beta_2\) is the difference between the intercept that would be computed using only the observations made when timing is early and \(\beta_0\). Consequently, \(\beta_0 + \beta_2\) is the intercept that would be computed if only the late observations were regressed on light. A parallel set of interpretations are obtained for the slope.

Table 8: Coefficients and standard errors obtained from the linear regression of number of flowers on light and timing and the interaction of light and timing. \(R^2 = .799\).

| Variable                  | Coefficient | Std. Error | t-statistic | \(P(T > |t|)\) |
|---------------------------|-------------|------------|-------------|---------------|
| Intercept                 | 71.6        | 4.34       | 16.5        | .127          |
| Timing (Early)            | 11.5        | 6.14       | 1.87        | .0753         |
| Light                     | -.041       | .00743     | -5.52       | < .0001       |
| Timing (Early)×Light      | .00121      | .0105      | .115        | .910          |

From Table 8, the fitted intercept for the late timing group is 71.6 and, for the early group, it is 71.6 + 11.5 = 83.1, and the fitted slope for the early group is \(-.041 + .00121 = -.040\). In effect, separate regression equations have been fit for both groups. The advantage of using all the data at once to construct separate regressions is that formal tests of significance comparing the groups are immediately obtained from the regression output. Furthermore, the standard errors associated with the parameter estimates tend to be smaller than when obtained from separate regressions. Table 8 shows that there is no evidence that the effect of light depends on the timing (\(t = .115\), p-value = .910).

At this point in the analysis, it is appropriate to discard the interaction model and adopt the simpler model shown in Table 4.
A strategy for data analysis

With several, or perhaps many possible explanatory variables that might be included in a multiple regression model, it becomes more difficult to arrive at a model that adequately represents the population or process of interest. There are usually several competing objectives. The preeminent objectives are accuracy and simplicity of the fitted model, though problem-specific objectives are sometimes identifiable. Variable selection is the topic of Chapter 12, but a brief outline of a strategy is useful at this time.

1. **Preliminaries** Identify the questions of interest. Review the study design as it relates to the data (e.g., observational study versus designed experiment). Define the population of interest, and determine the scope of inference.

2. **Initial investigations into the data** The objective is to gain some insight into the data and identify areas wherein the data may not conform to conditions necessary for or assumed by the linear regression model. At this stage, most of the analysis is graphical in nature; in particular, plot the response variable against each of the explanatory variables are of central interest. Watch for outliers and observations with large leverage.

3. **Formulate an inferential model** To the extent possible, cast the questions of interest in terms of the model parameters. For example, a linear effect of light on flower production is expressed by a linear model in which light is treated as a quantitative explanatory variable.

4. **Evaluate the model** The objective is to determine if the model fits the data, and so a lack-of-fit test should be carried out if there are sufficient numbers of replicate \(x\)-values. A richer model containing quadratic terms (e.g., \(x^2\)) and/or interaction terms can be used to assess goodness-of-fit. Examine residuals and remove variables that are not important. If the model is not sufficient, then it may be helpful to conduct further investigations of the data (step 2.).

5. **Draw inferences** Assuming that the model evaluation process has yielded an adequate model, carry out inferential methods aimed at answering the questions of interest (e.g., hypothesis tests involving the parameters and confidence intervals).

6. **Presentation** Present the results in a form appropriate for the audience and with language that the audience is comfortable with.