The Figure to the right shows the geometric mean time to extinction for 62 species of land-birds that were residents of 16 islands around Britain. A species became extinct on an island when it was no longer observed to nest on the island. The period of observation was reported to have spanned several decades. Species are classified by two binary variables: migratory status (resident or migrant), and size (large or small). Questions of interest are whether migratory status and size are associated with extinction time, after accounting for number of pairs, and whether the effect of size depends on numbers of pairs.

A multiple linear regression model fit to the data is summarized in Table 1. Table 1 shows that the data provide no evidence that the effect of size depends on the pairs ($t_{68} = -0.682$, p-value=.497). Since the model is apparently incorrect (as it contains an useless explanatory variable—the interaction of size and number of pairs), it’s best to not to draw any further inferences from the fitted model. The next stage of analysis is to remove the non-significant interaction term.

The test shown in Table 1 is the hypothesis test used in simple linear regression to test whether a model parameter is zero. For a parameter $\beta \in \{\beta_0, \beta_1, \ldots, \beta_{p-1}\}$, the hypotheses

\begin{align*}
H_0 & : \beta = 0 \\
H_1 & : \beta \neq 0
\end{align*}

The geometric mean extinction time for species $i$ is

$$\bar{y}_{i,\text{geo}} = \left( n^{-1} \sum_{j=1}^{16} t_{ij}^{-1} \right)^{-1}.$$
are $H_0: \beta = 0$ and $H_a: \beta \neq 0$ and the test statistic is the $t$-ratio

$$T = \frac{\hat{\beta}}{\hat{\sigma}(\beta)}.$$  

If $H_0: \beta = 0$ is correct, then $T \sim t_{n-p}$ and p-value $= 2P(T \geq |t|)$, where $t$ is the realized value of the test statistic.

Table 1: Coefficients and standard errors obtained from the multiple linear regression of geometric mean extinction time on number of pairs, species size (small or large), migratory status (migrant or resident) and the interaction of size and number of pairs. $n = 62, R^2 = .657.$

| Variable                  | Coefficient | Std. Error | t-statistic | P(T > |t|) |
|---------------------------|-------------|------------|-------------|---------|
| Intercept                 | .250        | .198       | 1.263       | .2118   |
| Pairs                     | .267        | .0373      | 7.158       | < .0001 |
| Size (large)              | .705        | .201       | 3.509       | .0008   |
| Migrant (true)            | -.425       | .246       | -1.72       | .0899   |
| Pairs×Size(large)         | -.177       | .368       | -1.481      | .632    |

After removing the constructed variable Pairs×Size, the regression model provides convincing evidence that the size of the bird species is associated with extinction time (Table 2). The coefficient associated with the size indicator variable is .652; hence, the estimated difference in extinction time between large and small species, after accounting for number of pairs and migratory status is $\hat{\beta}_{size} = .652$ years. The associated $t$-statistic is $t = 3.91$ (degrees of freedom are $n - p = 62 - 4 = 58$) and the p-value for the two-sided alternative $H_0 : \beta_{migrant} \neq 0$ is p-value $= 2P(T \geq 3.91) = .0002.$

Table 2: Coefficients and standard errors obtained from the multiple linear regression of geometric mean extinction time on number of pairs, species size (small or large) and migratory status (migrant or resident). $n = 62, R^2 = .652.$

| Variable                  | Coefficient | Std. Error | t-statistic | P(T > |t|) |
|---------------------------|-------------|------------|-------------|---------|
| Intercept                 | .283        | .184       | 1.538       | .129    |
| Pairs                     | .265        | .0368      | 7.203       | < .0001 |
| Size (large)              | .652        | .167       | 3.913       | .0002   |
| Migrant (true)            | -.504       | .183       | -2.76       | .0077   |

A comparison of Tables 1 and 2 shows that the main effect coefficients associated with numbers of pairs and size change when the interaction term is dropped from the model. The intercept changes dramatically as well because the model is different. In fact, it is unexpected when model coefficients do not change when an explanatory variable is added or removed.
from a model.

Occasionally, a one-sided hypothesis test is motivated by a specific scientific question, in which case the output from a standard statistical package can be used, with some care, to obtain the correct p-value. For example, if the hypotheses are

\[ H_0 : \beta_i = \beta \]
\[ H_a : \beta_i > \beta, \]

then the test statistic is

\[ T = \frac{\hat{\beta}_i - \beta}{\sigma(\hat{\beta}_i)}. \]

Restating \( H_0 \) as \( H_0 : \beta_i - \beta = 0 \) implies that the numerator of the \( t \) statistic should be \( \hat{\beta}_i - \beta \) since the expected value of the test statistic must be zero if \( H_0 \) is true so that \( T \sim t_{n-p} \). The p-value is the right-tail area since \( H_a \) can be restated as \( H_a : \beta_i - \beta > 0 \). Specifically, \( p-value = P(T \geq t|H_0) \) where \( t \) is the observed value of the \( t \)-statistic.

**Indicator variables - revisited**

Both migrant status and bird size are represented in the linear regression model as indicator variables. For instance, the indicator variable for migrant status is

\[ x_{migrant} = \begin{cases} 
0, & \text{if the species is a resident} \\
1, & \text{if the species is a migrant} 
\end{cases} \]

The effect of \( x_{migrant} \) is to create a separate intercept for the model; specifically, for migrants, the model surface intercepts the \( y \)-axis at a point .504 years below the intercept point for residents. The reference level is the level for which the indicator variable is 0—in this case, resident is the reference level\(^4\). In some cases, it’s desirable to change the level set internally by R to be the reference level. The command `contrasts(Migrant) = contr.treatment(2, base = 2)` accomplishes the change. The first argument of the command `contr.treatment` specifies the number of levels of the factor and the second argument specifies which level is to be the reference level. The levels and their internal R order can be determined by the function call `levels(Migrant)`. After making the change, the fitted model is shown in Table 3.

\(^4\)Some authors use the terms alias or baseline in place of reference.
Table 3: Coefficients and standard errors obtained from the multiple linear regression of mean extinction time on number of pairs, species size (small or large) and migratory status (migrant or resident). \( n = 62, R^2 = .652 \).

| Variable          | Coefficient | Std. Error | t-statistic | \( P(T > |t|) \) |
|-------------------|-------------|------------|-------------|----------------|
| Intercept         | .283        | .184       | 1.53        | 0.129          |
| Pairs             | .265        | .0368      | 7.20        | < .0001        |
| Size (large)      | .652        | .167       | 3.91        | .0002          |
| Migrant (false)   | .504        | .183       | 2.76        | .0077          |

**Anorexia experiment**

To assess the efficacy two behavioral therapies for the treatment of anorexia, 72 volunteer anorexia patients participated in an experiment in which each was randomly assigned to either a placebo group, family therapy, or cognitive behavioral therapy.\(^5\) Each subject underwent 6 weeks of treatment. Pre- and post-experiment weights were recorded. Previously, I have used the weight difference (post−pre) as a response variable and conducted a one-way analysis of variance. There was convincing evidence (\( F_{1,71} = 7.17, p\text{-value} = .0092 \)) that the therapies resulted in larger mean post-pre weight differences than the placebo therapy. There was some evidence (\( t_{34.2} = 1.93, p\text{-value} = .0616 \)) of a difference between cognitive behavioral and family therapies.

It’s possible that the effect of a particular therapy depends in part on the severity of the disease, and that the mean post-experiment weight would vary with the pre-experiment weight (as those individuals with more severe anorexia will tend to exhibit smaller pre-experiment weights). To accommodate this possibility, post-experiment weight is chosen as the response variable and pre-experiment weight is treated as an explanatory variable.\(^6\)

The first model assumes that the relationship between pre-experiment and post-experiment weight is the same (linear) for each therapy except that the intercept may differ among treatments. I do not view the model as realistic since I think that the subjects in the therapy groups may respond to the treatment and also their weight gain may depend on disease severity as accounted for by pre-experiment weight. Therefore, I expect that there is a different relationship between pre- and post-experiment weights among the groups, and that the slope of the post experiment weights (as a function of pre-experiment weight) will be greater than zero for the therapy groups, and approximately zero for the placebo group. This

---


\(^6\)Pre-experiment is also properly called a **covariate** in this situation in which the main interest lies in a factor (therapy) and pre-experiment weight needs to be accounted for as a source of variation in post-experiment weight.
first, *constrained* model represents a simple but reasonable alternative to a more complicated second model.

The constrained model requires two indicator variables that together identify which therapy as been administered to each subject. It’s logical to assign the control therapy as the reference treatment. Then, the parameter associated with the two indicator variables estimate the difference between the reference therapy (no therapy) and a respective therapy.

For example, the family therapy indicator variable is

\[
x_{\text{fam}} = \begin{cases} 
0, & \text{if the treatment is not family therapy} \\
1, & \text{if the treatment is family therapy}
\end{cases}
\]

Letting \(y\) and \(x_{\text{pre}}\) represent post-experiment weight and pre-experiment weight respectively, the constrained linear model is

\[
\mu(y|x) = \beta_0 + \beta_1 x_{\text{pre}} + \beta_{\text{fam}} x_{\text{fam}} + \beta_{\text{cog}} x_{\text{cog}}.
\]

Thus, the unconstrained model allows for different intercepts for each treatment, but the slope associated with pre-treatment weight is the same for all three treatments.

In contrast, the unconstrained model allows for different intercepts for each treatment (as does the constrained model) and also different slopes for each group. (Possibly) different slopes are introduced by constructing interaction variables between the group indicators \(x_{\text{fam}}\) and \(x_{\text{cog}}\) and pre-experiment weight \(x_{\text{pre}}\). The interaction variable between \(x_{\text{fam}}\) and \(x_{\text{pre}}\) is computed as \(x_{\text{pre} \times \text{fam}} = x_{\text{fam}} \times x_{\text{pre}}\). The unconstrained, or interaction model is

\[
\mu(y|x) = \beta_0 + \beta_1 x_{\text{pre}} + \beta_{\text{fam}} x_{\text{fam}} + \beta_{\text{cog}} x_{\text{cog}} + \beta_{\text{fam} \times \text{pre}} x_{\text{fam}} x_{\text{pre}} + \beta_{\text{cog} \times \text{pre}} x_{\text{cog}} x_{\text{pre}}.
\]

It’s not necessary to construct the interaction variables, or even the indicator variables as the R function call `lm(Post~Pre*Treat)` fits the constrained model by creating the indicator variables internally before the model is fit. The unconstrained model is fit using the function call `lm(Post~Pre*Treat)`. The figures below show the data and the fitted constrained (left) and unconstrained (right) models.

Visually (below), there appears to be substantial differences between the slopes for the therapy groups and the placebo, and little difference in slope among the two therapies. Parameter estimates and tests of significance for the constrained model are shown in Table 4.
Table 4: Coefficients and standard errors obtained from the constrained linear regression of post-experiment weight on pre-experiment weight and therapy. The placebo group is the reference group. $R^2 = .277$, $n = 72$.

| Variable                  | Coefficient | Std. Error | t-statistic | $P(T > |t|)$ |
|---------------------------|-------------|------------|-------------|-------------|
| Intercept                 | 45.6        | 13.2       | 3.45        | .0009       |
| Pre-experiment weight     | .434        | .161       | 2.69        | .0088       |
| Treat (cognitive therapy) | 4.10        | 1.89       | 2.16        | .0339       |
| Treat (family therapy)    | 8.66        | 2.19       | 3.95        | .0002       |

The tests of significance associated with each treatment group are testing whether there is a difference between the placebo group and the specific therapy after accounting for pre-experiment weight. By changing the reference group to either family or cognitive behavioral therapy, a test of significance comparing the two actual therapies can be obtained. Before doing so, it’s best to examine the unconstrained model.

Parameter estimates and tests of significance for the unconstrained model are shown in Table 5. There is abundant evidence that the unconstrained model improves on the constrained model as both interaction variables are statistically significant at the $\alpha = .02$ level. A formal test of significance can be obtained from an extra-sums-of-squares $F$-test, as demonstrated below.
Table 5: Coefficients and standard errors obtained from the unconstrained linear regression of post-experiment weight on pre-experiment weight and therapy. The placebo group is the reference group. $R^2 = .379, n = 72.$

| Variable                  | Coefficient | Std. Error | t-statistic | $P(T > |t|)$ |
|---------------------------|-------------|------------|-------------|-------------|
| Intercept                 | 92.0        | 18.8       | 4.89        | < .0001     |
| Pre-experiment weight     | -.134       | .230       | -.583       | .561        |
| Treat (cognitive therapy) | -76.4       | 28.3       | -2.70       | .0088       |
| Treat (family therapy)    | -77.2       | 33.1       | -2.33       | .0228       |
| Pre×Treat (cognitive)     | .982        | .344       | 2.85        | .0058       |
| Pre×Treat (family)        | 1.04        | .400       | 2.61        | .0112       |

The model summarized in Table 5 is somewhat confusing since the intercept is large and the coefficients associated with the indicators of treatment $x_{cog}$ and $x_{fam}$ are negative and large in magnitude. It’s possible improve the situation by centering the pre-experiment weights by subtracting the mean from each response observation, i.e., computing

$$
x_{\text{pre centered}} = x_{\text{pre}} - \bar{x}_{\text{pre}}
= x_{\text{pre}} - 82.408.
$$

After fitting the model using the centered pre-experiment weights\(^7\), the fitted coefficients associated with $x_{cog}$ and $x_{fam}$ estimate the difference in mean post-experiment weight between the placebo group and the two therapy groups when pre-experiment weight is fixed at $\bar{x}_{\text{pre}} = 82.408$ lb. Table 6 shows the model after centering pre-experiment weights.

Table 6: Coefficients and standard errors obtained from the unconstrained linear regression of post-experiment weight on centered pre-experiment weight and therapy. The placebo group is the reference group. $R^2 = .379, n = 72.$

| Variable                  | Coefficient | Std. Error | t-statistic | $P(T > |t|)$ |
|---------------------------|-------------|------------|-------------|-------------|
| Intercept                 | 81.0        | 1.302      | 62.2        | < .0001     |
| Pre-experiment weight     | -.134       | .230       | -.583       | .561        |
| Treat (cognitive therapy) | 4.46        | 1.78       | 2.50        | .0149       |
| Treat (family therapy)    | 8.75        | 2.07       | 4.22        | < .0001     |
| Pre×Treat (cognitive)     | .982        | .344       | 2.85        | .0058       |
| Pre×Treat (family)        | 1.04        | .400       | 2.61        | .0112       |

Now $\hat{\beta}_{fam} = 8.75$ (Table 6) is the estimated difference between the mean post-experiment weight of patients receiving family therapy and the mean post-experiment weight of patients receiving the placebo treatment given that the patients’ pre-experiment weight is

\(^7\)The analysis based on this model is sometimes called an analysis of covariance (ANCOVA).
When pre-experiment weight is not centered (Table 5), the comparison of treatment effects is made at the point $x_{pre} = 0$ lb (a nonsensical value).

The coefficients associated with $x_{cog}$ and $x_{fam}$ in Table 6 are easier to interpret, but the tests of significance have changed. The $t$-statistics are different because in the first case, the statistic compares the differences in mean post-experiment weight when pre-experiment weight is zero (Table 5), and in the second case, when pre-experiment weight is $\bar{x}_{pre} = 82.408$ lb.

The $t$-statistics are potentially misleading because their value (and p-value) depend on how pre-experiment is measured. The extra-sums-of-squares $F$-statistic described below does not suffer from this problem.

Generally, when two or more explanatory variables interact with each other variable in a model, analysts do not

1. Interpret the coefficients of either variable since they depend on the value of the other (interacting) variable (as illustrated above).

2. Assess the significance of a main effect. The term main effect refers to the non-interacting version of the variable. In the anorexia example, $x_{fam}$, $x_{cog}$ and $x_{pre}$ are main effects. Rather than attempting to assess the significance of a main effect, the correct objective is to assess the significance of the variable. When interaction is present, an interacting variable is represented in the model by multiple parameters (those that are associated with the main and the interaction variables). Assessing the significance of the variable is usually accomplished by an extra-sums-of-squares $F$-test that compares the fit of the full model with both main and interaction effects to the fit of a reduced model containing only the main effect variables.

The extra-sums-of-squares $F$-test

The extra-sums-of-squares $F$-test is used to formally compare the fit of two competing models when one model is a constrained version of the another. It was presented in Chapter 6 within the context of the one-way analysis of variance.\(^8\)

- The objective is to compare the residual error of a constrained version of the model containing $p_{\text{constrained}}$ model parameters to the residual error of an unconstrained version

\(^8\)The one-way analysis of variance is actually a special case of multiple linear regression where all variables in the full model are indicator variables of treatment groups arising from a single factor.
of the model with \( p_{\text{unconstrained}} \) model parameters. Let

\[ r = p_{\text{unconstrained}} - p_{\text{constrained}} > 0. \]

- The parameters \( \beta_i, \beta_{i+1}, \ldots, \beta_{i+r} \) are free in the unconstrained model are \textit{constrained to be zero} in the constrained model. The estimates of \( \beta_i, \beta_{i+1}, \ldots, \beta_{i+r} \) are determined wholly by the data.

The residual sums-of-squares for the constrained model will be no smaller than the residual sums-of-squares for the unconstrained model, i.e., \( \text{SSR}_{\text{constrained}} \geq \text{SSR}_{\text{unconstrained}} \). The magnitude of the difference reflects the importance of the \( r \) additional parameters in the unconstrained model.

- The denominator of the \( F \)-statistic uses the unconstrained model estimator of \( \sigma^2 \). It is

\[ \hat{\sigma}^2_{\text{unconstrained}} = \frac{\text{SSR}_{\text{unconstrained}}}{n - p_{\text{unconstrained}}}. \]

\( \hat{\sigma}^2_{\text{unconstrained}} \) is the estimated residual variance associated with the unconstrained model (in Chapter 6 this model was also called the full model).

- The hypotheses are

\[ H_0 : \beta_i = \beta_{i+1} = \cdots = \beta_{i+r} = 0 \quad \text{versus} \quad H_a : \text{at least one of } \beta_i, \beta_{i+1}, \ldots, \beta_{i+r} \text{ is not } 0. \]

- The test statistic is

\[ F = \frac{(\text{SSR}_{\text{constrained}} - \text{SSR}_{\text{unconstrained}})/r}{\hat{\sigma}^2_{\text{unconstrained}}} = \frac{\text{MS}_{\text{lack-of-fit}}}{\hat{\sigma}^2_{\text{unconstrained}}}, \]

where \( r \) is the difference in number of model parameter between the constrained and unconstrained models, and \( \text{MS}_{\text{lack-of-fit}} = (\text{SSR}_{\text{constrained}} - \text{SSR}_{\text{unconstrained}})/r \). If \( H_0 \) is correct, then

\[ F \sim F_{r, n - p_{\text{unconstrained}}} \]

As always with \( F \)-statistics,

\[ p\text{-value} = P(F \geq f|H_0), \]

where \( f \) is the observed value of the test statistic.
Returning to the anorexia experiment, the test of significance for interaction between pre-
experiment weight and treatment compares the fit of the models

constrained: \( \mu(y|x) = \beta_0 + \beta_1 x_{\text{pre}} + \beta_{\text{fam}} x_{\text{fam}} + \beta_{\text{cog}} x_{\text{cog}} \)

with

unconstrained: \( \mu(y|x) = \beta_0 + \beta_1 x_{\text{pre}} + \beta_{\text{fam}} x_{\text{fam}} + \beta_{\text{cog}} x_{\text{cog}} + \beta_{\text{fam}\times\text{pre}} x_{\text{fam}} x_{\text{pre}} + \beta_{\text{cog}\times\text{pre}} x_{\text{cog}} x_{\text{pre}}. \)

Table 7 provides the sums-of-squares and details of the \( F \)-test. From the table, \( \text{SSR}_{\text{constrained}} = 3311.3, \text{SSR}_{\text{unconstrained}} = 2844.8, \ r = 2 \) and \( \hat{\sigma}^2_{\text{unconstrained}} = 43.10. \) Thus,

\[
F = \frac{(\text{SSR}_{\text{constrained}} - \text{SSR}_{\text{unconstrained}})/r}{\hat{\sigma}^2_{\text{unconstrained}}} = \frac{(3311.3 - 2844.8)/2}{43.10} = \frac{233.25}{43.10} = 5.411.
\]

There is convincing evidence that the unconstrained model fit is better than the fit of the
constrained model since \( p \)-value = \( P(F_{2,66} \geq 5.411) = .0067. \) Thus, there is convincing
evidence that the relationship between pre- and post-experiment weight differs among the
three treatment groups. The results of the test as summarized below the table.\(^9\)

<table>
<thead>
<tr>
<th>Source of Residual Sum variation</th>
<th>of squares</th>
<th>d.f.</th>
<th>Mean square</th>
<th>( F )-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constrained model</td>
<td>3311.3</td>
<td>68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack-of-fit</td>
<td>466.5</td>
<td>2</td>
<td>233.24</td>
<td>5.411</td>
<td>.0067</td>
</tr>
<tr>
<td>Unconstrained model</td>
<td>2844.8</td>
<td>66</td>
<td>43.10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Pre-experiment weight interacts with treatment.

2. Mean post-experiment weight varies with pre-experiment weight and the relationship
differs among therapies.

3. Therapy affects mean post-experiment weight and the effect of a particular therapy
varies with pre-experiment weight. The form of the relationship differs among thera-
pies.

\(^9\)The statements differ according to whether pre-experiment weight or therapy is emphasized.
R has a convenient function for computing the extra-sums-of-squares $F$-test. To use it, fit and save both models as objects. Then pass the fitted models to the function `anova`. The sequence of function calls used to compute the test above is

\[
\begin{align*}
\text{lm.1} &= \text{lm}(\text{Post} \sim \text{Pre} * \text{Treat}) \\
\text{lm.2} &= \text{lm}(\text{Post} \sim \text{Pre} + \text{Treat}) \\
\text{anova(lm.2, lm.1)}
\end{align*}
\]

The left-most argument is the unconstrained model.

The $F$-test for overall significance

Most statistical packages compute the $F$-test for overall significance, an extra-sums-of-squares $F$-test that compares the fitted (unconstrained or full) model with $p$ parameters to the single mean model (or constrained or reduced model) with one parameter and no explanatory variables, i.e., $\mu(y|x) = \beta_0$. If any of the $p$ parameters is linearly associated with the response variable, the test statistic is likely to be statistically significant. Consequently, the test is usually uninteresting and is usually not reported or commented on.

Prostate-specific antigen

Prostate-specific antigen (PSA) is a protein produced by the cells of the prostate gland. PSA is present in small quantities in the blood of men with healthy prostates, but is often elevated in the presence of prostate cancer and in other prostate disorders. A blood test to measure PSA is considered the most effective test currently available for the early detection of prostate cancer. However, the effectiveness of PSA as a screening technique has been questioned. Not only is PSA found in the blood serum of healthy men, it is also found in the blood of healthy women but more commonly in the blood of women with breast, lung, renal, or uterine cancer.

Some insight into the predictive value of PSA may be gleaned from data collected from $n = 125$ men who received a radical prostatectomy.\textsuperscript{10} The Figure above shows the sample

distribution of PSA. According to the American Cancer Society guidelines, men with a PSA level greater than 10.0 ng/ml are at an increased risk for prostate cancer (more than a 67% chance compared to men with PSA levels less than 10.0 ng/ml). Levels between 4.0 ng/ml and 10.0 ng/ml may indicate prostate cancer (about a 25% chance), or other noncancerous conditions.

The variables measured in the study are shown in the table below. The objective of the multiple regression analysis is to investigate the extent to which PSA is related to the variables in the table. These variables are understood to be related to, or reflections of prostate cancer. If PSA is to be considered a good predictor of prostate cancer, then these variables ought to explain most of the observed variation in PSA.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpsa</td>
<td>log PSA</td>
</tr>
<tr>
<td>lcavol</td>
<td>log cancer volume</td>
</tr>
<tr>
<td>lweight</td>
<td>log prostate weight</td>
</tr>
<tr>
<td>age</td>
<td>age (years)</td>
</tr>
<tr>
<td>lbph</td>
<td>log of the amount of benign prostatic hyperplasia</td>
</tr>
<tr>
<td>svi</td>
<td>seminal vesicle invasion (\text{binary}^a)</td>
</tr>
<tr>
<td>lcp</td>
<td>log of capsular penetration (b)</td>
</tr>
<tr>
<td>gleason</td>
<td>numeric score (c)</td>
</tr>
<tr>
<td>pgg45</td>
<td>percent of Gleason score 4 or 5 (d)</td>
</tr>
</tbody>
</table>

\(a\) Indicates seminal vesicle invasion by prostate cancer. Patient prognosis is poor when seminal vesical invasion is found.

\(b\) The prostatic capsule is a membrane that surrounds and encloses the prostate gland.

\(c\) A Gleason score is assigned by a pathologist to a tissue sample based upon its microscopic appearance. Cancers with a higher Gleason score are more aggressive and have a worse prognosis.

\(d\) A Gleason score 4 or 5 is assigned if the tissue has few or no recognizable glands. Many cells are invading the surrounding tissue, and there are often just sheets of cells throughout the surrounding tissue.
Preliminary analysis showed that a transformation of PSA to the log scale improved linearity of the relationships with the explanatory variables.

For interpretability, the remaining variables (all explanatory) are standardized. The parameter estimates $\hat{\beta}_1, \ldots, \hat{\beta}_8$ are then roughly comparable measures of the effect of the associated variable on log PSA. A precise comparison is not really possible because of correlations and associations among the variables as evidenced by the Figure (right). However, a one unit change in a standardized variable corresponds to a comparison at $x$ and $x + s_x$. If log PSA can be predicted well from these alternative indicators and measures of prostate cancer progression measured on excised tissue, then log PSA can be argued to be an informative measure of cancer risk or progression.

No attempt is made to introduce biological realism into the models; instead, the objective is limited to determining the extent of observed variation in log PSA that can be explained by the pathologically derived measures of cancer.

The figure above, showing scatterplots of the response and those explanatory variables found to be statistically significant predictors at the $\alpha = .05$ significance level, reveals substantial association among variables. One observation was identified as a recording error and has been removed from the data set. Table 9 shows the model after eliminating all explanatory variables not found to be significant at the $\alpha = .1$ level. I've adopted a test-wise $\alpha$ level of $.1$ for testing since the consequences of incorrectly including a variable for predictive purposes are less severe than when the objective is to identify a defensible scientific model. If an explanatory variable is included in a model when it is not an informative predictor of the response variable mean, then the associated parameter estimate tends to be small in magnitude and does not substantively affect predictions compared to a model without the suspect variable.\textsuperscript{11}

\textsuperscript{11}If the objective were to develop a scientifically realistic model of the factors affecting PSA, then incorrectly identifying a variable as being important to explaining variation in PSA is a fairly significant mistake.
Table 9: Coefficients and standard errors obtained from the linear regression of log PSA on log cancer volume, log prostate weight and seminal vesicle invasion. $R^2 = .649$, $n = 95$.

| Variable | Coefficient | Std. Error | $t$-statistic | $P(T > |t|)$ |
|----------|-------------|------------|---------------|-------------|
| Intercept | -1.206      | .660       | -1.83         | .071        |
| log cancer volume | .524        | .0746      | 7.02          | < .0001     |
| log prostate weight | .781        | .186       | 4.19          | < .0001     |
| seminal vesicle invasion (yes) | 4.27        | 1.92       | 2.22          | .0292       |
| log prostate weight $\times$ seminal vesicle invasion | -.963       | .512       | -1.88         | .0632       |

The residual plot (right) shows no evidence of lack of fit or nonconstant variance. A normal quantile-quantile plot (not shown) indicates that the residuals do not deviate from a normal distribution. The model shown above reveals that log cancer volume, log prostate weight, and seminal vesicle invasion and the interaction between log prostate weight, and seminal vesicle are important variables explaining variation in log PSA. The variables together explain 64.9% of the observed variation in log PSA, and so can be said to explain the majority of the variation in log PSA. Since interpretation and model simplicity are secondary to the principal objectives, a richer model containing all two-interaction variables as explanatory variables was fit. One interaction variable was retained: log prostate weight $\times$ seminal vesical invasion (previously identified as important). A brief cross-validation study of prediction error was conducted since the focus is on prediction.

Cross-validation

Cross-validation\(^\text{12}\) is a method of estimating the accuracy of a predictive function that avoids optimistic bias induced by over-fitting. Over-fitting is the tendency of a prediction function to predict a response $y_i$ more accurately when the data tuple $(y_i, x_{i,1}, \ldots, x_{i,p-1})$ belongs to sample used to fit the model (often called the training sample in the lexicon of statistical learning) than when the data tuple does not belong to the training sample. Cross-validation algorithms hold each observation tuple out of the training sample when the response $y_i$.

is predicted. Accuracy is estimated by comparing the predictions to the observed values. Cross-validation estimators of accuracy are essentially unbiased if the training sample is representative of the population of interest, a condition that demands that the training sample has been collected via probability sampling, or some other method such as systematic sampling that is not likely to generate a sample for which the responses can be predicted with greater accuracy that the responses belonging to the full population.

A $k$-fold cross-validation algorithm randomly partitions a data set as $k$ disjoint subsets each containing approximately $n/k$ observations. Each subset serves as a test set for estimating the accuracy of linear regression model (or another prediction function) constructed from all observations excluded from the test set. After fitting the linear regression model, it is used to predict each of the left-out observations. Each observation is predicted once. Since the partitioning of the data set is accomplished by a random process, the process often is repeated a number of times (usually between 20 and 200 times).

For this problem, I set $k = 10$ and carried out six replications of the 10-fold cross-validation algorithm. The Figure above shows the predictions plotted against the observations for each replication. The association between cross-validation predictions and the observed responses is moderate ($r = .778$). The cross-validation coefficient of determination is $R^2 = .606 = .778^2$ not very large, and I conclude that there are other variables not identified in this study that affect log PSA, and as a diagnostic tool, it should not be used alone.

The value of $R^2$ reported by R was .649, which over-estimates the proportion of observed variation in log PSA explained by the regression model by approximately $100(.649−.606)% = 4.5\%$. The difference is attributed to over-fitting bias.

**Adjusted $R^2$**

The adjusted $R^2$ is an measure of goodness of fit closely related to the coefficient of determination. The adjusted $R^2$ suffers from less bias that $R^2$. Bias in $R^2$ is only a concern when $p$ is large relative to $n$, say whenever $30p > n$. When $p$ is large relative to $n$, then the coefficient of determination tends to optimistically represent the goodness of fit of the model.
to the data. The adjusted $R^2$ is

$$\text{adjusted } R^2 = \frac{\hat{\sigma}_{\text{total}}^2 - \hat{\sigma}_{\text{LM}}^2}{\hat{\sigma}_{\text{total}}^2}$$

where $\hat{\sigma}_{\text{total}}^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2/(n - 1)$ and $\hat{\sigma}_{\text{LM}}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n - p)$ is the residual mean square. The coefficient of determination is computed from the same sums-of-squares but without scaling by $n - 1$ and $n - p$. Scaling by the degrees of freedom reduces bias but unfortunately the interpretation of $R^2$ as the fraction of the observed variation in the response variable explained by the regression does not extend to the adjusted $R^2$. For the prostate data, the adjusted $R^2$ and $R^2$ are .634 and .649, respectively. Both estimates are optimistically biased.

The matrix approach to multiple linear regression

The statistical theory and computational methods supporting multiple linear regression are matrix-based and, providing that the reader is familiar with linear algebra, remarkably simple. For the benefit of the reader that is comfortable with linear algebra, a brief discussion is presented.

The following notation is used: a vector is lower-case and bold-face and a matrix upper-case and bold-face. The principal vectors and matrices are

$$\beta_{p \times 1} = (\beta_0, \beta_1, \cdots, \beta_{p-1})^T$$

$$y_{n \times 1} = (y_1, y_2, \cdots, y_n)^T$$

$$X_{n \times p} = \begin{pmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\
1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p-1}
\end{pmatrix}$$

As with simple linear regression, the estimator $\hat{\beta}$ is the least squares estimator; that is, it minimizes the sum of the squared residuals

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1x_{i,1} - \cdots - \hat{\beta}_{p-1}x_{i,p-1})^2 = (y - X\beta)^T(y - X\beta).$$

Differentiating the expression of the right with respect to $\beta$ and setting the matrix of partial derivatives equal to the zero-vector yields the normal equations

$$X^TX\beta = X^Ty.$$
Solving for $\beta$ gives the estimator

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$ 

If $p = 2$ so that there is a single explanatory variable, then the estimators of $\beta_0$ and $\beta_1$ are the same as before. The estimator of the variance of $\beta_i$ is $\hat{\sigma}^2_{LM}(\hat{\beta}_i)$ is the $i + 1$ diagonal element of the estimated variance-covariance matrix $\hat{\sigma}^2(X^T X)^{-1}$ where $\hat{\sigma}^2_{LM} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p)$. The standard error of $\hat{\beta}_i$ is $\hat{\sigma}_{LM}(\hat{\beta}_i) = \sqrt{\hat{\sigma}^2_{LM}(\hat{\beta}_i)}$. An accessible discussion of the matrix approach to linear regression can be found in *Applied Regression Analysis and Generalized Linear Models, Second Edition*, John Fox, Sage, 2008 as well as many other texts.