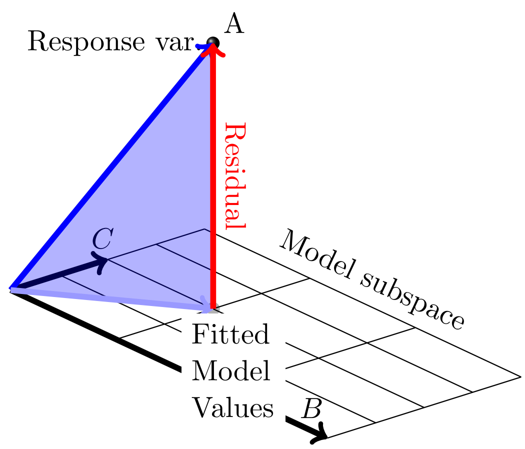

Figure 9.7: Projecting A onto the subspace defined by a set of two model vectors, B and C. The model triangle is shaded.

R^2 for models has a similar interpretation. Consider the model A $\sim$ B+C. Since there are two explanatory vectors involved, there is an ambiguity: which is the angle to consider: the angle A to B or the angle A to C?

Figure 9.7 shows the situation. Since there are two explanatory vectors, the response is projected down onto the space that holds both of them, the **model subspace**. There is still a vector of fitted model values and a residual vector. The three vectors taken together — response variable, fitted model values, and residual — form a right triangle. The angle between the response variable and the fitted model values is the one of interest. R is the cosine of that angle.

Because the vectors B and C could be oriented in any direction relative to one another, there's no sense in worrying about whether the angle is acute (less than 90°) or obtuse. For this reason, R^2 is used — there's no meaning in saying that R is negative.

9.6 Computational Technique

The coefficient of determination, R^2, compares the variation in the response variable to the variation in the fitted model value. It can be calculated as a ratio of variances:

DATA FILE
swim100m.csv

```
> swim = fetchData("swim100m.csv")
> mod = lm(time ~ year + sex, data=swim)
> with(swim, var(fitted(mod)) / var(time))
[1] 0.844
```

The **regression report** is a standard way of summarizing models. Such a report is produced by most statistical software packages and used in many fields. The first part of the table contains the coefficients — labeled "Estimate" — along with other information that will be introduced starting in Chapter 12. The R^2 statistic is a standard part of the report; look at the second line from the bottom.

```
> summary(mod)
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 555.7168    33.7999   16.44  < 2e-16 ***
year         -0.2515     0.0173  -14.52  < 2e-16 ***
sexM         -9.7980     1.0129   -9.67  8.8e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.98 on 59 degrees of freedom
Multiple R-squared: 0.844,       Adjusted R-squared: 0.839
F-statistic:  160 on 2 and 59 DF,  p-value: <2e-16
```

Occasionally, you may be interested in the correlation coefficient r between two quantities. You can, of course, compute r by fitting a model, finding R^2, and taking a square root.

```
> mod2 = lm(time ~ year, data=swim)
> summary(mod2)
```

The summary report (not shown here — you can do the calculation yourself!) gives $R^2 = 0.5965$. This corresponds to r of

```
> sqrt(0.5965)
[1] 0.772
```

The `cor()` function computes this directly:

```
> with(swim, cor(time, year))
[1] -0.772
```

Note that the negative sign on r indicates that record swim `time` decreases as `year` increases. This information about the direction of change is contained in the sign of the coefficient from the model. The magnitude of the coefficient tells how fast the `time` is changing (with units of seconds per year). The correlation coefficient (like R^2) is without units.

Keep in mind that the correlation coefficient r summarizes only the simple linear model A $\sim$ B where B is quantitative. But the coefficient of determination, R^2, summarizes any model; it is much more useful. If you want to see the direction of change, look at the sign of the coefficient.

Reading Questions

- How does R^2 summarize the extent to which a model has captured variability?