8 ANOVA for multiple linear regression

8.1 Example

An estimate is required of the percentage yield of petroleum spirit from crude oil, based upon certain rough laboratory determinations of properties of the crude oil. The following table shows actual percentage yields of petroleum spirit, \( y \), and four properties, \( x_1, x_2, x_3, x_4 \), of the crude oil, for samples from 32 different crudes.

<table>
<thead>
<tr>
<th>Data on yields of petroleum spirit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y ) ( x_1 ) ( x_2 ) ( x_3 ) ( x_4 )</td>
</tr>
<tr>
<td>6.9  38.4  6.1  220  235</td>
</tr>
<tr>
<td>14.4 40.3  4.8  231  307</td>
</tr>
<tr>
<td>7.4  40.0  6.1  217  212</td>
</tr>
<tr>
<td>8.5  31.8  0.2  316  365</td>
</tr>
<tr>
<td>8.0  40.8  3.5  210  218</td>
</tr>
<tr>
<td>2.8  41.3  1.8  267  235</td>
</tr>
<tr>
<td>5.0  38.1  1.2  274  285</td>
</tr>
<tr>
<td>12.2  50.8  8.6  190  205</td>
</tr>
<tr>
<td>9.0  32.2  5.2  236  267</td>
</tr>
<tr>
<td>15.2  38.4  6.1  220  300</td>
</tr>
<tr>
<td>26.8  40.3  4.8  231  367</td>
</tr>
<tr>
<td>14.0  32.2  2.4  284  351</td>
</tr>
<tr>
<td>14.7  31.8  0.2  316  379</td>
</tr>
<tr>
<td>6.4  41.3  1.8  267  275</td>
</tr>
<tr>
<td>17.6  38.1  1.2  274  365</td>
</tr>
<tr>
<td>22.3  50.8  8.6  190  275</td>
</tr>
<tr>
<td>24.8  32.2  5.2  236  360</td>
</tr>
<tr>
<td>26.0  38.4  6.1  220  365</td>
</tr>
<tr>
<td>34.9  40.3  4.8  231  395</td>
</tr>
<tr>
<td>18.2  40.0  6.1  217  272</td>
</tr>
<tr>
<td>23.2  32.2  2.4  284  424</td>
</tr>
<tr>
<td>18.0  31.8  0.2  316  428</td>
</tr>
<tr>
<td>13.1  40.8  3.5  210  273</td>
</tr>
<tr>
<td>16.1  41.3  1.8  267  358</td>
</tr>
<tr>
<td>32.1  38.1  1.2  274  444</td>
</tr>
<tr>
<td>34.7  50.8  8.6  190  345</td>
</tr>
<tr>
<td>31.7  32.2  5.2  236  402</td>
</tr>
<tr>
<td>33.6  38.4  6.1  220  410</td>
</tr>
<tr>
<td>30.4  40.0  6.1  217  340</td>
</tr>
<tr>
<td>26.6  40.8  3.5  210  347</td>
</tr>
<tr>
<td>27.8  41.3  1.8  267  416</td>
</tr>
<tr>
<td>45.7  50.8  8.6  190  407</td>
</tr>
</tbody>
</table>
The variables recorded are as follows.

\[ y: \] percentage yield of petroleum spirit
\[ x_1: \] specific gravity of the crude
\[ x_2: \] crude oil vapour pressure, measured in pounds per square inch
\[ x_3: \] the ASTM 10% distillation point, in °F
\[ x_4: \] the petroleum fraction end point, in °F

It is required to use these data to provide an equation for predicting \( y \) from measurements of the four explanatory variables, \( x_1, x_2, x_3, x_4 \), (or some subset of them).

The data has been read into an S-PLUS data frame `oil`. The function `names` is used to assign names to the five variables. The `linear model` function `lm` is then used to carry out a multiple linear regression of the response variable `spirit` upon the four regressor variables, `gravity`, `pressure`, `distil` and `endpoint`, the results of which are stored in the object `oil.lm`.

```r
> names(oil) <- c("spirit","gravity","pressure","distil","endpoint")
> oil.lm <- lm(spirit ~ gravity + pressure + distil + endpoint, data = oil)
> summary(oil.lm)

Call: lm(formula = spirit ~ gravity + pressure + distil + endpoint, data = oil)
Residuals:
       Min        1Q    Median        3Q       Max
-3.58000 -1.52200 -0.10880  1.42400  4.62100

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.820837   10.123192  -0.6738   0.5062
  gravity     0.227179    0.099909   2.2739   0.0311
  pressure    0.553690    0.369785   1.4976   0.1458
  distil     -0.149526    0.029198  -5.1160  0.0000 **
  endpoint    0.154706    0.006424  23.9922  0.0000 **

Residual standard error: 2.234 on 27 degrees of freedom
Multiple R-Squared: 0.9622, Adjusted R-Squared: 0.9596
F-statistic: 171.7 on 4 and 27 degrees of freedom, p-value: 0

Correlation of Coefficients:

                     (Intercept) gravity pressure distil
  gravity          -0.6682
  pressure         0.0369  0.3816
  distil         -0.9080  0.8461  0.426
  endpoint       -0.0173  0.0426 -0.1944 -0.3091
```

2
8.2 Hypotheses about the parameters

We now outline the theory that will enable us to interpret the above analysis and to carry out further analyses. In Section 7.5 we found that the sum of squares of the response variable may be decomposed as:

\[ y'y = \hat{y}'\hat{y} + e'e, \]  

(1)

where on the right hand side of Equation (1) the second term is the residual sum of squares and the first term is that part of the sum of squares of the response variable that is accounted for by the fitted regression. This term may be rewritten, using the result of Section 7.4 that

\[ \hat{y} = Xb = Hy, \]  

(2)

where,

\[ H = X(X'X)^{-1}X'. \]  

(3)

Thus

\[ \hat{y}'\hat{y} = b'X'Hy = b'X'y = b_0 \sum_{i=1}^{n} y_i + \sum_{r=1}^{k} b_r \sum_{i=1}^{n} x_{ir}y_i. \]  

(4)

After some algebraic manipulation, subtracting \( n\bar{y}^2 \) from both sides of Equation (1) and using the first of the \( k+1 \) normal equations of Section 7.1, we find that we may partition the total (corrected) sum of squares \( SS_T \equiv \sum_{i=1}^{n}(y_i - \bar{y})^2 \) as

\[ SS_T = SS_{Reg} + SS_R. \]  

(5)

Here \( SS_R \equiv \sum_{i=1}^{n}e_i^2 \) is the residual sum of squares. \( SS_{Reg} \) is the regression sum of squares, that part of the total sum of squares that is accounted for by the fitted regression, and which may be written as

\[ SS_{Reg} = \sum_{r=1}^{k} b_r S_{ry}, \]

where

\[ S_{ry} = \sum_{i=1}^{n}(x_{ir} - \bar{x}_r)(y_i - \bar{y}) \quad r = 1, \ldots, k. \]

Corresponding to the partition of the sum of squares of Equation (5) we have the following ANOVA, where, as we saw in Section 7.7, \( s^2 \equiv MS_R \) is an unbiased estimator of the error variance \( \sigma^2 \).

**ANOVA TABLE**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>( k )</td>
<td>( \sum b_r S_{ry} )</td>
<td>( SS_{Reg}/k )</td>
</tr>
<tr>
<td>Residual</td>
<td>( n-k-1 )</td>
<td>by subtraction</td>
<td>( s^2 \equiv SS_R/(n-k-1) )</td>
</tr>
<tr>
<td>Total</td>
<td>( n-1 )</td>
<td>( \sum_{i=1}^{n}(y_i - \bar{y})^2 )</td>
<td></td>
</tr>
</tbody>
</table>
We may wish to test the hypothesis that there is no linear relationship between the response variable \( y \) and the regressor variables \( x_1, x_2, \ldots, x_k \). Formally, we test the null hypothesis

\[
H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0
\]

against the alternative

\[
H_1 : \beta_j \neq 0 \quad \text{for some} \quad j = 1, \ldots, k.
\]

The two terms on the right hand side of Equation (5) are independently distributed, with \( SS_R/\sigma^2 \sim \chi^2_{n-k-1} \) and, under \( H_0 \), \( SS_{Reg}/\sigma^2 \sim \chi^2_k \). Note that here we are using the assumption that the errors \( \epsilon_i \) in the regression model are NID(0, \( \sigma^2 \)). The hypothesis \( H_0 \) is tested using a one-tail test with test statistic

\[
F = \frac{MS_{Reg}}{MS_R},
\]

which under \( H_0 \) has the \( F_{k,n-k-1} \) distribution.

Another hypothesis that we may wish to test is that some of the \( k \) regressor variables are redundant, i.e., that a subset of them, \( x_1, x_2, \ldots, x_m \), say, is sufficient, where \( m < k \). Formally, we test the null hypothesis

\[
H_0(m) : \beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0.
\]

To test \( H_0(m) \),

1. fit the full model by carrying out the regression of \( y \) on \( x_1, x_2, \ldots, x_k \) to obtain the estimates \( b_k \) of \( \beta_k \equiv (\beta_0, \beta_1, \ldots, \beta_k)' \);

2. fit the reduced model by carrying out the regression of \( y \) on \( x_1, x_2, \ldots, x_m \) to obtain the estimates \( b_m \) of \( \beta_m \equiv (\beta_0, \beta_1, \ldots, \beta_m)' \).

- Note that in general, when some of the variables are removed from a regression, the estimates of the remaining parameters are altered. Thus \( b_m \) is not equal to the first \( m + 1 \) components of \( b_k \).

- It is merely a matter of notational convenience that the last \( k - m \) variables are considered for redundancy here. The method to be described applies equally well for any selection of variables.

Denote by \( SS_{Reg}(m) \) the regression sum of squares, associated with \( m \) degrees of freedom, that is obtained from fitting the reduced model and by \( SS_{Reg}(k) \) the regression sum of squares from the full model, associated with \( k \) degrees of freedom. Thus the change in the regression sum of squares when the full model is replaced by the reduced one is

\[
SS_{Reg}(k) - SS_{Reg}(m).
\]

The quantity in Formula (6) is also known as the sum of squares due to \( x_{m+1}, x_{m+2}, \ldots, x_k \) adjusted for the presence of \( x_1, x_2, \ldots, x_m \) and is associated with \( k - m \) degrees of freedom.
Under $H_0(m)$, the F-statistic
\[ F = \frac{(SS_{Reg}(k) - SS_{Reg}(m))/(k - m)}{MS_R}, \tag{7} \]
where $MS_R$ is the residual mean square for the full model, has the $F_{k-m,n-k-1}$ distribution. The test statistic (7) is used to carry out a (one-tail) test of $H_0(m)$.

The F-statistic (7) may be described as
\[ F = \frac{\text{change in regression ss}}{\text{change in df}} \times \frac{\text{residual ms for the larger model}}{\text{residual ms}} \]
The commonest case arises when we test whether one particular variable, $x_k$ say, is redundant. Then the sum of squares due to $x_k$ adjusted for the presence of $x_1, x_2, \ldots, x_{k-1}$ is given by
\[ SS_{Reg}(k) - SS_{Reg}(k-1), \]
and the corresponding test statistic for testing whether the variable $x_k$ is redundant is
\[ F = \frac{SS_{Reg}(k) - SS_{Reg}(k-1)}{MS_R} \tag{8} \]
with 1 and $n-k-1$ degrees of freedom, where $MS_R$ is, as previously, the residual mean square for the full model.

The above analysis provides the basis for a systematic approach to searching for a “best” subset of the regressor variables to use.

The value of the F-statistic (8) is equal to $t^2$, where $t$ is the t-ratio corresponding to $x_k$ in the full model as given in the S-PLUS output. The p-values of these F and t-statistics are identical. Thus, in the regression output, the p-values associated with each of the regressor variables are the p-values for the significance of each regressor variable in the presence of the others.

### 8.3 The coefficient of determination (multiple $R^2$)

The coefficient of determination is defined by
\[ R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_R}{SS_T}. \]
It is the proportion of the total sum of squares that is accounted for by the fitted regression and may be regarded as a measure of the goodness of fit of the regression model.

An alternative measure, which is often preferred, is the adjusted coefficient of determination (adjusted for the number of regressor variables),
\[ \tilde{R}^2 = 1 - \frac{MS_R}{MS_T}, \]
where $MS_T = SS_T/(n - 1)$. Both coefficients take the value 1 if the fit is perfect, i.e., $SS_R = 0$.

In comparing the effects of different sets of regressor variables, we may start with a minimal set, possibly an empty set, and think of adding in variables one or more at a time. Whenever we introduce an extra regressor variable, $SS_R$ decreases and the proportion of the total sum of squares accounted for by the regression, $R^2$, necessarily increases. However, it is also the case that the error degrees of freedom are necessarily reduced by 1 whenever we introduce an extra regressor variable, and in certain cases the introduction of an extra variable is totally counterproductive, so that $MS_R$ increases and $R^2$ decreases.

### 8.4 Example (continued)

We are now ready to examine in detail the S-PLUS output for our example. The fitted regression coefficients are given by

$$b = (-6.8208, 0.2272, 0.5537, -0.1495, 0.1547)'$$

The corresponding p-values show that, in the presence of the other regressor variables, only the effect of the variable pressure is not significant at the 5% level. Thus, according to the test statistic that we have described, the removal of this variable from the regression does not lead to a significant reduction in the goodness of the fit, and we might consider omitting it.

The Residual standard error: 2.234 is the square root $s$ of the residual mean square from the ANOVA. The Multiple R-Squared: 0.9622 is the value of the coefficient of determination, $R^2$. The value 171.7 of the overall F-statistic from the regression ANOVA, with its corresponding p-value of 0, shows that there is highly significant evidence of a linear relationship between the response variable and the four regressor variables. Finally, the correlation matrix of the fitted regression coefficients is given, obtained from the expression for the covariance matrix of Equation (17) of Section 7.6.

In view of the fact that pressure does not have a significant p-value, we repeat the regression analysis without it.
> oil3.lm <- lm(spirit ~ gravity + distil + endpoint, data = oil)
> summary(oil3.lm)

Call: lm(formula = spirit ~ gravity + distil + endpoint, data = oil)
Residuals:
    Min     1Q Median     3Q    Max
-3.53 -1.361 -0.2681  1.391  4.766

Coefficients:
             Value Std. Error  t value Pr(>|t|)
(Intercept) 4.0320     7.2233  0.5582 0.5811
  gravity   0.2217     0.1021  2.1725 0.0384
  distil  -0.1866     0.0159 -11.7177 0.0000
   endpoint 0.1565     0.0065 24.2238 0.0000

Residual standard error: 2.283 on 28 degrees of freedom
Multiple R-Squared: 0.959
F-statistic: 218.5 on 3 and 28 degrees of freedom, the p-value is 0

Correlation of Coefficients:
                (Intercept) gravity distil
  gravity -0.9199
  distil -0.8150  0.6579
  endpoint -0.1780  0.0508 -0.2766

All the regressor variables are now significant, so we might very reasonably decide to adopt this model. However, the variable gravity is not very highly significant, so we also consider the model after removing gravity as well.

> oil2.lm <- lm(spirit ~ endpoint + distil, data = oil)
> summary(oil2.lm)

Call: lm(formula = spirit ~ endpoint + distil, data = oil)
Residuals:
    Min     1Q Median     3Q    Max
-3.959 -1.906 -0.3711  1.624  4.38

Coefficients:
             Value Std. Error  t value Pr(>|t|)
(Intercept) 18.4676     3.0090  6.1374 0.0000
  endpoint  0.1558     0.0069 22.7308 0.0000
  distil  -0.2093     0.0127 -16.4349 0.0000

Residual standard error: 2.426 on 29 degrees of freedom
Multiple R-Squared: 0.9521
F-statistic: 288.4 on 2 and 29 degrees of freedom, the p-value is 0

Correlation of Coefficients:
                (Intercept) endpoint
  endpoint -0.3351
  distil -0.7104  -0.4122
Now we have a model for which both the remaining regressor variables are highly significant, so we could not remove either of them without a serious loss of fit of the model.

To summarize our results so far and to extend them somewhat, consider the following table, where \( m \) denotes the number of regressor variables being used.

<table>
<thead>
<tr>
<th>( m )</th>
<th>( s )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2.234</td>
<td>0.962</td>
</tr>
<tr>
<td>3</td>
<td>2.283</td>
<td>0.959</td>
</tr>
<tr>
<td>2</td>
<td>2.426</td>
<td>0.952</td>
</tr>
<tr>
<td>1</td>
<td>7.659</td>
<td>0.506</td>
</tr>
</tbody>
</table>

For each value of \( m \), we have chosen the best set of \( m \) regressor variables, in the sense that it provides the best fit, giving the smallest value of \( s \) and the largest value of \( R^2 \). For \( m = 2 \), as may be checked by carrying out regressions on all possible pairs of regressor variables, \text{endpoint} \ and \text{distil} \ is the best pair of regressor variables to use. For \( m = 1 \), as we shall see later, the best single regressor variable to use is \text{endpoint}.

From inspection of the table, we see that there is relatively little difference in the fit of the model, whether 2, 3 or 4 regressor variables are used. However, the use of only one regressor variable gives a much poorer fit. In choosing between the models with 2, 3 or 4 regressor variables, it is a matter of judgement whether we prefer a more complex model that give us a slightly better fit or a simpler model that gives a slightly poorer fit.

On this basis we may, at least for the present, opt for the model that uses the two regressor variables \text{endpoint} \ and \text{distil} — we shall look at this problem again later. From the output for oil2.lm, we see that the fitted regression equation is

\[
\text{spirit} = 18.4676 + 0.1558 \times \text{endpoint} - 0.2093 \times \text{distil}.
\]

To investigate the adequacy of the model we may examine plots of the residuals and other quantities discussed in Section 7.8. In S-PLUS, a simple way of doing this is by the following commands.

```r
> par(mfrow = c(3,2))
> plot(oil2.lm)
```

The command `par(mfrow = c(3,2))` formats the graphics window so that it will produce six plots, laid out in three rows and two columns.

For example, one of the plots is a normal quantile-quantile plot, in which the ordered residuals \( e_{(i)} \) are plotted against quantiles of the standard normal distribution,

\[
\Phi^{-1}\left(\frac{i - \frac{1}{2}}{n}\right), \quad i = 1, \ldots, n \quad [\text{if } n \geq 11],
\]

where \( \Phi \) is the cumulative distribution function of the standard normal distribution. If the residuals are approximately normally distributed then the plot should be approximately linear.
8.5 Multicollinearity and ill-conditioning

The problem of multicollinearity arises when some of the regressor variables in a multiple linear regression are highly correlated with each other, i.e., there are strong linear relationships among them. The matrix $X'X$ is then ill-conditioned. This means that the matrix is close to being singular, which results in at least some of the estimates $b_r$ having excessively large standard deviations. Problems of numerical accuracy in the calculation of the estimates also arise, these being the more serious the closer are the correlations to $\pm 1$. Fitted models are unstable in the sense that a small perturbation to an observation can result in a very different fitted model.

Just on the basis of intuition, it should be possible to use regressor variables which are highly correlated with each other as alternatives to each other in the regression. As one way of dealing with the problem of multicollinearity, decisions will often have to be made on which of the highly correlated variables to remove from the regression.

Generally, in multiple linear regression, a choice may have to be made between several models with varying numbers of regressor variables and hence with varying numbers of parameters. The choice may not be a clear-cut one, although the use of the F-tests described earlier provides one possible basis for a systematic approach.

A model with more terms in it will account for a greater part of the variability in the response variable and give better predictions; but a model with fewer terms, although it does not explain as much of the variability in the response variable, may have other advantages:

1. it may be easier to interpret;
2. it may be less prone to problems of multicollinearity, so that its parameters are more accurately estimated;
3. it may be more accurate for prediction when the model is extrapolated to values of the regressor variables not covered in the experiment, i.e., the simpler model has greater scope.

However, great care has to be taken with extrapolation – the fitted model may not be at all appropriate for use outside the range of the regressor variables covered by the experiment.

8.6 The AIC and a stepwise approach

In attempting to compare models with different numbers of parameters, a commonly adopted approach is to use the Akaike information criterion (AIC), which is defined by

$$
\text{AIC} = -2 \text{ maximized log-likelihood} + 2 \text{ number of parameters} \\
= n \ln(SS_R/n) + 2p + \text{ const.}
$$

The model with the smallest value of the AIC is judged to be the most appropriate.

- A detailed discussion of the concept of likelihood will be given in other courses. Here we merely present the result of the mathematical analysis.
• As we are only interested in comparing the AICs of different models, it is sufficient to evaluate the AIC up to an arbitrary additive constant.

• Note that the AIC weighs the desirability of having a small value of the residual sum of squares $SS_R$ against that of having a small number of parameters $p$.

The S-PLUS function \texttt{step} is used to carry out a stepwise regression procedure which, by sequentially deleting or adding regressor variables, attempts to find a “best” set of regressor variables. In fact the S-PLUS function \texttt{step} does not use the exact AIC as its criterion function. Somewhat confusingly, S-PLUS nevertheless in some contexts calls its criterion function the AIC, but in others calls it the \texttt{Cp}, a reference to another criterion statistic, \textit{Mallows’} $C_p$. Venables and Ripley prefer to use their own function, \texttt{stepAIC}, which uses the exact AIC. This function is made accessible from the MASS library in S-PLUS by the command \texttt{library(MASS)}. The function \texttt{step} in R is similar to \texttt{stepAIC} in S-PLUS.

At each step of the procedure generated by the S-PLUS \texttt{step} function, the following is done:

1. Given the current set of regressor variables, the procedure calculates the criterion statistic for the current set (no variables deleted) and for all models with one of the current regressor variables deleted. The model with the smallest value of the statistic is selected.

2. If no variable has been deleted then the procedure calculates the criterion statistic for the current set (no variables added) and for all models with one regressor variable added from among those not in the current set. The model with the smallest value of the statistic is selected.

3. The procedure stops when no more variables can be deleted or added according to the above criteria.

• Note that this procedure, or any variant of it, is an exploratory tool which produces a single suggested set of regressor variables.

• The procedure is not guaranteed to find in any sense the “best” set of regressor variables, although it may be a useful guide. It is not necessarily the case that there even is a clearly “best” set of regressor variables. If the procedure terminates with, say, $m$ regressor variables, there may be another set of $m$ regressor variables which give a better fit.

• A more systematic way of investigating which variables could be omitted from the regression is to carry out and compare the regression analyses for all possible subsets of the regressor variables, but if $k$ is large then this is prohibitively time-consuming since there are $2^k - 1$ subsets of the regressor variables.
We also have to specify the starting point for the iteration. In the following slightly compressed output for our example, the object \texttt{oil0.lm} is the result of fitting no regressor variables to the response variable \texttt{spirit}. It is used to specify the starting point for the iteration of no regressor variables present. Thus the \texttt{step} function has \texttt{oil0.lm} as its first argument. The second argument is the \texttt{scope} argument, which specifies the set of regressor variables to be considered for inclusion.

At each stage, for each candidate model, the output lists the change in the regression sum of squares due to the deleted or added variable, the resulting residual sum of squares (RSS) and the criterion statistic (Cp). (The quantity \texttt{scale} is an arbitrary constant used in the calculation of the criterion statistic.)

Finally, a brief summary is provided for the chosen model, in the present case the model with the two regressor variables \texttt{endpoint} and \texttt{distil}. This is the same model that we tentatively adopted earlier.

```r
> oil0.lm <- lm(spirit ~ 1, data = oil)
> step(oil0.lm, ~ gravity + pressure + distil + endpoint, data = oil)
Start: AIC=3794.018
  spirit ~ 1

Single term additions
Model:
spirit ~ 1
scale: 114.9702

Df Sum of Sq RSS Cp
<none> 3564.077 3794.018
gravity 1 216.256 3347.821 3807.702
pressure 1 525.738 3038.339 3498.220
distil 1 353.700 3210.377 3670.258
endpoint 1 1804.384 1759.694 2219.575

Step: AIC=2219.574
  spirit ~ endpoint

Single term deletions
Model:
spirit ~ endpoint
scale: 114.9702

Df Sum of Sq RSS Cp
<none> 1759.694 2219.575
endpoint 1 1804.384 3564.077 3794.018

Single term additions
Model:
spirit ~ endpoint
scale: 114.9702

Df Sum of Sq RSS Cp
<none> 1759.694 2219.575
gravity 1 897.747 861.947 1551.768
pressure 1 1389.826 369.867 1059.688
distil 1 1589.082 170.612 860.433
```

12
Step: AIC= 860.4329
spirit ~ endpoint + distil

Single term deletions
Model:
spirit ~ endpoint + distil
scale: 114.9702

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>Cp</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>170.612</td>
<td>860.433</td>
<td></td>
</tr>
<tr>
<td>endpoint</td>
<td>3039.766</td>
<td>3210.377</td>
<td>3670.258</td>
</tr>
<tr>
<td>distil</td>
<td>1589.082</td>
<td>1759.694</td>
<td>2219.575</td>
</tr>
</tbody>
</table>

Single term additions
Model:
spirit ~ endpoint + distil
scale: 114.9702

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>Cp</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>170.6115</td>
<td>860.433</td>
<td></td>
</tr>
<tr>
<td>gravity</td>
<td>24.61041</td>
<td>146.0011</td>
<td>1065.763</td>
</tr>
<tr>
<td>pressure</td>
<td>9.99201</td>
<td>160.6195</td>
<td>1080.381</td>
</tr>
</tbody>
</table>

Call:
`lm(formula = spirit ~ endpoint + distil, data = oil)`

Coefficients:

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>endpoint</th>
<th>distil</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.46763</td>
<td>0.1558135</td>
<td>-0.209329</td>
</tr>
</tbody>
</table>

Degrees of freedom: 32 total; 29 residual
Residual standard error (on weighted scale): 2.425522

As an indication of the caution with which the results of stepwise procedures should be treated, note that the function `step` in R and the function `stepAIC` in S-PLUS, using the exact AIC, from the same starting point arrive at the full model with all four regressor variables. This is indeed the model with the smallest value of the AIC. The corresponding stepwise procedure in the package Minitab, based upon the use of the F-tests described in Section 8.2, with its default settings arrives at the model with the three regressor variables `endpoint`, `distil` and `gravity`.

The function `stepAIC` (and the function `step` in R) adopts a slightly simpler sequential procedure. At each stage it considers adding or deleting regressor variables one at a time from the current set, choosing the model that gives the smallest value of the AIC. For our example, starting with no regressor variables present, it arrives at the full set of variables as follows.
> stepAIC(oil0.lm, ~gravity + pressure + distil + endpoint, data = oil)
Start:  AIC= 152.81
spirit ~ 1

Df  Sum of Sq     RSS      AIC
+ endpoint  1 1804.384 1759.694 132.2291
+ pressure  1  525.738 3038.339 149.7066
 + distil    1  353.700 3210.377 151.4690
 + gravity   1  216.256 3347.821 152.8105
 <none>       3564.077 3564.077 152.8136

Step:  AIC= 132.23
spirit ~ endpoint

Df  Sum of Sq     RSS      AIC
+ distil    1 1589.082 170.612  59.5569
+ pressure  1 1389.826 369.867  84.3171
 + gravity   1  897.747 861.947 111.3906
 <none>       1759.694 132.2291
- endpoint  1 1804.384 3564.077 152.8136

Step:  AIC= 59.56
spirit ~ endpoint + distil

Df  Sum of Sq     RSS      AIC
+ gravity    1  24.610 146.001  56.5721
 <none>       170.612  59.5569
+ pressure  1   9.992 160.620  59.6257
 - distil     1 1589.082 1759.694 132.2291
- endpoint  1 3039.766 3210.377 151.4690

Step:  AIC= 56.57
spirit ~ endpoint + distil + gravity

Df  Sum of Sq     RSS      AIC
+ pressure    1 11.197 134.804  56.0187
 <none>       146.001  56.5721
- gravity     1 24.610 170.612  59.5569
 - distil     1  715.946 861.947 111.3906
- endpoint   1 3059.737 3205.738 153.4228

Step:  AIC= 56.02
spirit ~ endpoint + distil + gravity + pressure

Df  Sum of Sq     RSS      AIC
<none>       134.804  56.0187
- pressure    1 11.197 146.001  56.5721
- gravity     1 25.816 160.620  59.6257
 - distil     1 130.676 265.480  75.7057
- endpoint   1 2873.952 3008.756 153.3935

Call:
  lm(formula = spirit ~ endpoint + distil + gravity + pressure, data = oil)

Coefficients:
(Intercept)  endpoint  distil  gravity  pressure
  -6.820774 0.1546501 -0.1495356 0.227246 0.5537262

Degrees of freedom: 32 total; 27 residual
Residual standard error: 2.234444
Returning to the function `step` in S-PLUS, we carry out the stepwise procedure from the starting point with all regressor variables present, which corresponds to the object `oil.lm`. In this case we find that the procedure deletes no variables, so that the suggested model is now the one with all four regressor variables!!!

```r
> step(oil.lm, ~ gravity + pressure + distil + endpoint, data = oil)
Start: AIC= 184.7314
  spirit ~ gravity + pressure + distil + endpoint

Single term deletions

Model:
  spirit ~ gravity + pressure + distil + endpoint

scale: 4.992739

Df  Sum of Sq     RSS       Cp
  <none>         134.804 184.731
  gravity 1     25.816  160.620  200.561
  pressure 1    11.197  146.001  185.943
  distil 1     130.676  265.480  305.421
  endpoint 1   2873.952 3008.756 3048.698
Call:
  lm(formula = spirit ~ gravity + pressure + distil + endpoint, data = oil)

Coefficients:
(Intercept) gravity pressure distil endpoint
  -6.8208  0.2272  0.5537 -0.1495  0.1547

Degrees of freedom: 32 total; 27 residual
Residual standard error: 2.234444
```

### 8.7 Prediction

One of the reasons for carrying out a linear regression analysis may be that, in future, given a set of values $x_1, x_2, \ldots, x_k$ of the regressor variables, we may wish to be able to predict the corresponding value $y$ of the response variable. Defining the vector $\mathbf{x}$ by

$$
\mathbf{x} = (1, x_1, x_2, \ldots, x_k)'
$$

we use the fitted regression equation to obtain

$$
\hat{y} = \mathbf{x}'\mathbf{b}. \quad (9)
$$

Assuming the validity of the linear regression model, for the given values $x_1, x_2, \ldots, x_k$, the observed value of $y$ will be given by

$$
y = \mathbf{x}'\beta + \epsilon,
$$

where, as before, the error term $\epsilon$ is assumed to have the $N(0, \sigma^2)$ distribution. Hence

$$
E(y) = \mathbf{x}'\beta
$$

and

$$
y = E(y) + \epsilon. \quad (10)
$$
The $\hat{y}$ defined in Equation (9) may be regarded in two ways, either as an estimator of $E(y)$ (the long-term average of all values of $y$ for the given values $x_1, x_2, \ldots, x_k$) or as a predictor of $y$ (one particular value of $y$ for the given values $x_1, x_2, \ldots, x_k$). In the latter case, there are two sources of error in accounting for the difference between an observed value of $y$ and the predicted value $\hat{y}$: one due to using the vector of estimates $b$ instead of the vector of actual parameter values $\beta$, and the other due to the presence of the error term $\epsilon$.

Since $b$ is an unbiased estimator of $\beta$,

$$E(\hat{y}) = E(x'b) = x'E(\beta) = x'\beta = E(y).$$

Also,

$$\text{var}(\hat{y}) = \text{var}(x'b) = x'\text{cov}(b)x = \sigma^2 x'(X'X)^{-1}x,$$

using Equation (17) of Section 7.6.

By estimating $\sigma^2$ by $s^2 \equiv MS_R$ from the ANOVA table, it can be shown that a 100(1$-\alpha$)% confidence interval for $E(y)$ is given by

$$x'b \pm t_{n-k-1,\alpha/2}s\sqrt{x'(X'X)^{-1}x}.$$

Similarly, it can also be shown that a 100(1$-\alpha$)% prediction interval\(^1\) for the value of $y$ is given by

$$x'b \pm t_{n-k-1,\alpha/2}s\sqrt{1 + x'(X'X)^{-1}x}.$$

- S-PLUS refers to the quantity

$$s\sqrt{x'(X'X)^{-1}x}$$

as the standard error of the fit, se.fit.

We use the function predict in S-PLUS to obtain predicted values and their standard errors. We shall continue to use the model with two regressor variables. We construct a data frame $x$ whose variable names are those of our regressor variables, endpoint and distil, and which contains the values of these regressor variables for which we wish to make predictions. In the present case, we shall use a single pair of values, $(400,200)$. The first argument of the predict function is the object oil2.lm that corresponds to our chosen model and the second argument is the data frame $x$ that contains the values of the regressor variables for which we wish to make predictions. The argument se.fit = TRUE is required so that we obtain standard errors for our predictions and so that, subsequently, we can use the function pointwise to produce confidence intervals.

In the output, the term residual.scale refers to the value of $s$. Given this and the value of standard error of the fit, we may, if desired, calculate the prediction interval as defined above, in addition to the confidence interval produced by the function pointwise.

---

\(^1\)So called because there is probability $1 - \alpha$ that the interval limits and $y$ will take values such that the value of $y$ lies within the interval limits.
> x <- data.frame(endpoint = 400, distil = 200)
> predict.result <- predict(oil2.lm, x, se.fit = TRUE)
> predict.result
$fit:
  1
38.92724

$se.fit:
  1
0.9395607

$residual.scale:
[1] 2.425522

$df:
[1] 29

> pointwise(predict.result, 0.95)
$upper:
  1
40.84885

$fit:
  1
38.92724

$lower:
  1
37.00562