3 Inference for Linear Models

3.1 Hypotheses about the parameters

We now outline the theory that will enable us to interpret the fitted regression model

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \qquad i = 1, \ldots, n \tag{3.1}$$

which, as described in Section 2.1, may be written in matrix algebra terms as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.2}$$

and to carry out further analyses.

To test the effect of different covariates, and to help us choose an appropriate model, we will assume normality of the errors, i.e. $\boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I})$, so that

$$\mathbf{y} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

It follows that the (ordinary least squares) estimates, $\mathbf{b}$, of $\boldsymbol{\beta}$ are given by

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \sim \text{MVN}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \tag{3.3}$$

Recall also that the error variance σ^2 is estimated by

$$s^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-k-1} = \frac{(\mathbf{y}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\mathbf{b})}{n-k-1} = \frac{SS_R}{n-k-1} \tag{3.4}$$

where $\dfrac{\mathbf{e}^T\mathbf{e}}{\sigma^2} = \dfrac{(n-k-1)s^2}{\sigma^2} \sim \chi^2_{n-k-1}$.

3.1.1 Assessing the regression

Write

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}. \tag{3.5}$$

Computing the inner product of $\mathbf{y}$ with itself, on the basis of the decomposition of Equation (3.5), and using Equation (2.12) to eliminate the cross-product term on the right hand side,

$$\mathbf{y}^T\mathbf{y} = \hat{\mathbf{y}}^T\hat{\mathbf{y}} + \mathbf{e}^T\mathbf{e}, \tag{3.6}$$

where on the right hand side of Equation (3.6) 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.

Now, consider the following expressions for $\hat{\mathbf{y}}^T\hat{\mathbf{y}}$ and $\mathbf{e}^T\mathbf{e}$ in (3.6).

Recall,

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{y}, \tag{3.7}$$

where,

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{3.8}$$

Thus

$$\begin{aligned}\hat{\mathbf{y}}^T\hat{\mathbf{y}} &= \mathbf{b}^T\mathbf{X}^T\mathbf{H}\mathbf{y} \\ &= \mathbf{b}^T\mathbf{X}^T\mathbf{y}.\end{aligned} \tag{3.9}$$

Also, from

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{3.10}$$

we have

$$\begin{aligned}\mathbf{e}^T\mathbf{e} &= \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y} \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y}.\end{aligned} \tag{3.11}$$

In the absence of the explanatory variables $x_1, \ldots, x_k$, it would follow that the regression model could be written as

$$y_i = \beta_0 + \epsilon_i$$

so that $\mathrm{E}[y_i] = \mu_i = \beta_0$, and the best estimate of β_0 is simply $\bar{y}$. The total (*corrected*) sum of squares is hence given by

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \\ &= \mathbf{y}^T\mathbf{y} - n\bar{y}^2\end{aligned} \tag{3.12}$$

Then, writing

$$\mathbf{y}^T\mathbf{y} - n\bar{y}^2 = (\mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y}) + (\mathbf{b}^T\mathbf{X}^T\mathbf{y} - n\bar{y}^2) \tag{3.13}$$

we see that the total (corrected) sum of squares can be decomposed to give

$$SS_T = SS_{Reg} + SS_R \tag{3.14}$$

(swapping the bracketed terms in (3.13)), where SS_T, SS_{Reg} and SS_R are the total (corrected) regression, and residual, sums of squares, respectively. The decomposition of the sums of squares can be summarized in the following ANOVA table.

ANOVA Table

Source of Variation	d.f.	Sum of Squares (SS)	Mean Square (MS)
Regression	k	$SS_{Reg} = \mathbf{b}^T\mathbf{X}^T\mathbf{y} - n\bar{y}^2$	$MS_{Reg} = SS_{Reg}/k$
Residual	$n-k-1$	$SS_R = \mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y}$	$MS_R = s^2 = SS_R/(n-k-1)$
Total	$n-1$	$SS_T(= S_{yy}) = \mathbf{y}^T\mathbf{y} - n\bar{y}^2$	

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

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

The two terms on the right hand side of (3.14) are independently distributed, with $SS_R/\sigma^2 \sim \chi^2_{n-k-1}$ and, under H_0, $SS_{Reg}/\sigma^2 \sim \chi^2_k$. 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.

3.1.2 Testing a subset of variables

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 $\mathbf{b}_k$ of $\boldsymbol{\beta}_k \equiv (\beta_0, \beta_1, \ldots, \beta_k)^T$;
2. fit the *reduced model* by carrying out the regression of y on $x_1, x_2, \ldots, x_m$ to obtain the estimates $\mathbf{b}_m$ of $\boldsymbol{\beta}_m \equiv (\beta_0, \beta_1, \ldots, \beta_m)^T$.

- Note that in general, when some of the variables are removed from a regression, the estimates of the remaining parameters are altered. Thus $\mathbf{b}_m$ is not equal to the first $m + 1$ components of $\mathbf{b}_k$.
- 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

$$SS_{Reg}(k) - SS_{Reg}(m). \tag{3.15}$$

The quantity in Formula (3.15) 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

$$F = \frac{(SS_{Reg}(k) - SS_{Reg}(m))/(k-m)}{MS_R}, \tag{3.16}$$

where MS_R is the residual mean square for the *full* model, has the $F_{k-m,n-k-1}$ distribution. The test statistic (3.16) is used to carry out a (one-tail) test of $H_0(m)$.

The F-statistic (3.16) may be described as

$$F = \frac{\text{change in regression ss / change in df}}{\text{residual ms for the larger model}}$$

3.1.3 Testing individual parameters

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

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

$$F = \frac{SS_{Reg}(k) - SS_{Reg}(k-1)}{MS_R} \tag{3.17}$$

with 1 and $n-k-1$ degrees of freedom, where MS_R is, as previously, the residual mean square for the full model.

The value of the F-statistic (3.17) is equal to t^2, where t is the t-ratio corresponding to x_k in the full model as given in the R output (see later). 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.

The t-statistics can be found directly by reference to (3.3), since we have

$$\mathrm{E}(b_j) = \beta_j \quad \text{and} \quad \mathrm{Var}(b_j) = \sigma^2 c_{jj} \quad j = 0, \ldots, k$$

where c_{jj} is the $(j+1)$th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. We have further that

$$b_j \sim \mathrm{N}(\beta_j, \sigma^2 c_{jj}) \tag{3.18}$$

so that, to test $H_0 : \beta_j = 0$, we use

$$\frac{b_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}} \sim \mathrm{N}(0,1) \tag{3.19}$$

and

$$\frac{b_j - \beta_j}{\sqrt{s^2 c_{jj}}} \sim t_{n-k-1} \tag{3.20}$$

where $\sqrt{s^2 c_{jj}} = s\sqrt{c_{jj}}$ is the *standard error* of b_j.

3.2 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 (albeit not the ultimate or definitive one).

An alternative measure, which is often preferred, is the *adjusted coefficient of determination* (adjusted for the number of regressor variables),

$$\bar{R}^2 = 1 - \frac{MS_R}{MS_T},$$

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 $\bar{R}^2$ decreases.

3.3 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 $\mathbf{X}^T\mathbf{X}$ 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 ± 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.

3.4 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 $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)^T$, we use the fitted regression equation to obtain

$$\hat{y} = \mathbf{x}^T\mathbf{b}. \tag{3.21}$$

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

$$y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon,$$

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

$$E(y) = \mathbf{x}^T\boldsymbol{\beta}$$

and

$$y = E(y) + \epsilon. \tag{3.22}$$

The $\hat{y}$ defined in Equation (3.21) 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 $\mathbf{b}$ instead of the vector of actual parameter values $\boldsymbol{\beta}$, and the other due to the presence of the error term ϵ.

Since $\mathbf{b}$ is an unbiased estimator of $\boldsymbol{\beta}$,

$$\mathrm{E}(\hat{y}) = \mathrm{E}(\mathbf{x}^T\mathbf{b}) = \mathbf{x}^T\mathrm{E}(\mathbf{b}) = \mathbf{x}^T\boldsymbol{\beta} = \mathrm{E}(y).$$

Also,

$$\begin{aligned}\mathrm{Var}(\hat{y}) &= \mathrm{Var}(\mathbf{x}^T\mathbf{b}) \\ &= \mathbf{x}^T\mathrm{Cov}(\mathbf{b}, \mathbf{b})\mathbf{x} \\ &= \sigma^2\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x},\end{aligned} \tag{3.23}$$

using (3.3).

By estimating σ^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

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

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

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

[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.

3.5 Example: Fitting a Linear Model in R

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.

Data on yields of petroleum spirit

y	x_1	x_2	x_3	x_4
6.9	38.4	6.1	220	235
14.4	40.3	4.8	231	307
7.4	40.0	6.1	217	212
8.5	31.8	0.2	316	365
8.0	40.8	3.5	210	218
2.8	41.3	1.8	267	235
5.0	38.1	1.2	274	285
12.2	50.8	8.6	190	205
10.0	32.2	5.2	236	267
15.2	38.4	6.1	220	300
26.8	40.3	4.8	231	367
14.0	32.2	2.4	284	351
14.7	31.8	0.2	316	379
6.4	41.3	1.8	267	275
17.6	38.1	1.2	274	365
22.3	50.8	8.6	190	275
24.8	32.2	5.2	236	360
26.0	38.4	6.1	220	365
34.9	40.3	4.8	231	395
18.2	40.0	6.1	217	272
23.2	32.2	2.4	284	424
18.0	31.8	0.2	316	428
13.1	40.8	3.5	210	273
16.1	41.3	1.8	267	358
32.1	38.1	1.2	274	444
34.7	50.8	8.6	190	345
31.7	32.2	5.2	236	402
33.6	38.4	6.1	220	410
30.4	40.0	6.1	217	340
26.6	40.8	3.5	210	347
27.8	41.3	1.8	267	416
45.7	50.8	8.6	190	407

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 have been read into R and stored as the *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`.

```
> 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
-3.5804 -1.5223 -0.1098  1.4237  4.6214

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.820774  10.123152  -0.674   0.5062
gravity      0.227246   0.099937   2.274   0.0311 *
pressure     0.553726   0.369752   1.498   0.1458
distil      -0.149536   0.029229  -5.116 2.23e-05 ***
endpoint     0.154650   0.006446  23.992  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.234 on 27 degrees of freedom
Multiple R-squared:  0.9622,    Adjusted R-squared:  0.9566
F-statistic: 171.7 on 4 and 27 DF,  p-value: < 2.2e-16
```

The fitted regression coefficients are given by

$$\mathbf{b} = (-6.8208, 0.2272, 0.5537, -0.1495, 0.1547)^T.$$

and the fitted regression equation is

$$\widehat{\texttt{spirit}} = -6.8208 + 0.2272\ \texttt{gravity} + 0.5537\ \texttt{pressure} - 0.1495\ \texttt{distil} + 0.1547\ \texttt{endpoint}$$

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 at least one of the four regressor variables.

In view of the fact that `pressure` is not significant in the presence of the other variables, on account of its large 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
-3.5303 -1.3606 -0.2681  1.3911  4.7658

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.032034   7.223341   0.558   0.5811
gravity      0.221727   0.102061   2.173   0.0384 *
distil      -0.186571   0.015922 -11.718 2.61e-12 ***
endpoint     0.156527   0.006462  24.224  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.283 on 28 degrees of freedom
Multiple R-squared:  0.959,     Adjusted R-squared:  0.9546
F-statistic: 218.5 on 3 and 28 DF,  p-value: < 2.2e-16
```

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 might also consider the model after removing `gravity` as well.

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

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

Residuals:
    Min      1Q  Median      3Q     Max
-3.9593 -1.9063 -0.3711  1.6242  4.3802

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.467633   3.009009   6.137 1.09e-06 ***
```

```
distil      -0.209329   0.012737 -16.435 3.11e-16 ***
endpoint     0.155813   0.006855  22.731  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.426 on 29 degrees of freedom
Multiple R-squared:  0.9521,    Adjusted R-squared:  0.9488
F-statistic: 288.4 on 2 and 29 DF,  p-value: < 2.2e-16
```

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.

m	s	R^2	$\bar{R}^2$
4	2.234	0.962	0.957
3	2.283	0.959	0.955
2	2.426	0.952	0.949
1	7.659	0.506	0.490

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 (and $\bar{R}^2$). For $m = 2$, as may be checked by carrying out regressions on all possible pairs of regressor variables, `endpoint` and `distil` is the best pair of regressor variables to use. For $m = 1$, the best single regressor variable to use is `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 `endpoint` and `distil`. From the output for `oil2.lm`, we see that the fitted regression equation is

$$\widehat{\texttt{spirit}} = 18.4676 - 0.2093\ \texttt{distil} + 0.1558\ \texttt{endpoint}.$$

ANOVA

Note that the *result* of the overall ANOVA F-test for the null hypothesis of no regression effect is presented as part of the summary output, but that this is not in the familiar form of the ANOVA table. However, this can be constructed using the `anova()` function as follows:

```
> anova(oil.lm)
Analysis of Variance Table

Response: spirit
          Df  Sum Sq Mean Sq  F value    Pr(>F)
gravity    1  216.26  216.26  43.3141 4.644e-07 ***
pressure   1  309.85  309.85  62.0603 1.807e-08 ***
distil     1   29.21   29.21   5.8514   0.02258 *
endpoint   1 2873.95 2873.95 575.6263 < 2.2e-16 ***
Residuals 27  134.80    4.99
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Note that this presents the additional sum of squares due to each variable, by adding them sequentially to a model beginning with a single *intercept*. Hence, the additional sum of squares due to adding all four variables is

$$216.26 + 309.85 + 29.21 + 2873.95 = 3429.27$$

and the table may be constructed as follows:

ANOVA Table

Source of Variation	d.f.	Sum of Squares (SS)	Mean Square (MS)
Regression	4	3429.27	857.32
Residual	27	134.80	4.99
Total	31	3564.077	

The F-statistic is hence $\dfrac{857.32}{4.99} = 171.70$, matching the value given in the summary, with p-value calculated as follows:

```
> pf(171.7,4,27, lower.tail = FALSE)
[1] 8.829276e-19
```

Note that the single degree of freedom F-test for the inclusion of `endpoint` in the model is given directly by the `anova()` function. The change in SS due to `endpoint` is `2873.95` which when divided by the residual mean square results in an observed value of the F-statistic of `575.62` on 1 and 27 degrees of freedom (`p=0.0000`). Note also that the F-value is 23.992^2, so that this test is exactly equivalent to the t-test of `endpoint` reported in the summary.

Other ANOVA tests can be constructed directly. To compare two nested models constructed using the `lm()` function, `mod1` and `mod2`, say, we use

> `anova(mod1,mod2)`.

For example:

```
> oil0.lm <- lm(spirit ~ 1, data = oil)
> anova(oil0.lm, oil.lm)
Analysis of Variance Table

Model 1: spirit ~ 1
Model 2: spirit ~ gravity + pressure + distil + endpoint
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     31 3564.1
2     27  134.8  4    3429.3 171.71 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Since `oil0.lm` is the model with intercept only, this output again results in the same overall ANOVA from the *full* model summary.

Prediction

We use the function `predict` in R 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, `distil` and `endpoint`, 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, (200, 400). 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 arguments `se.fit = TRUE` and `interval = c("confidence", "prediction")`, are required so that we obtain confidence and prediction intervals, respectively.

Note that R refers to the quantity $s\sqrt{\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}$ as the *standard error of the fit*, `se.fit`. 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 confidence and prediction intervals as defined in Section 3.4 above.

```
> x <- data.frame(distil = 200, endpoint = 400)
> confidence <-predict(oil2.lm, x, se.fit = T, interval = c("confidence"))
> confidence
$fit
       fit      lwr      upr
1 38.92724 37.00562 40.84885

$se.fit
[1] 0.9395607

$df
[1] 29

$residual.scale
[1] 2.425522

> prediction <-predict(oil2.lm, x, se.fit = T, interval = c("prediction"))
> prediction
$fit
       fit      lwr      upr
1 38.92724 33.60731 44.24717
```

```
$se.fit
[1] 0.9395607

$df
[1] 29

$residual.scale
[1] 2.425522
```

so that, for a crude oil with an ASTM 10% distillation point of 200°F and a petroleum fraction endpoint of 400°F, the model predicts a percentage yield of petroleum spirit of 38.93%, with 95% confidence interval (37.01, 40.85) and 95% prediction interval (33.61, 44.25).