Linear Models: The full rank model — inference
Recap

In this section, we develop various forms of hypothesis testing on the full rank model. To recap, the full rank model is

\[ y = X\beta + \varepsilon \]

where the errors \( \varepsilon \) have:

- mean 0;
- variance \( \sigma^2 I \);
- (for some theorems) are normally distributed.
The first thing we want to test is for *model adequacy* — is the model that we have appropriate?

If our model is inappropriate, then none of the $x$ variables have any relevance to predicting $y$. This means that all the parameters will be 0.

Therefore we want to test the null hypothesis

$$H_0 : \beta = 0.$$
What is the alternative hypothesis?
What is the alternative hypothesis?

If the model is useful, then at least some of the $x$ variables are relevant to predicting $y$, and the corresponding parameters will be nonzero. So our alternative hypothesis is

$$H_1 : \beta \neq 0.$$ 

To test these hypotheses, we assume throughout the section that the errors $\varepsilon$ are normally distributed.
The method used to test the hypotheses is ANOVA, which is probably familiar to you. Here we go into it in depth and provide a sound theoretical basis for the tests.

The principle behind ANOVA is that if $\beta = 0$, then $y = \varepsilon$ consists entirely of errors. In this case, $y^T y$, the sum of squares of the errors, measures the variability of the errors.

However, if $\beta \neq 0$, then $y = X\beta + \varepsilon$. In this case, $y^T y$ is not made up solely of the errors but also of the model predictions. Some of $y^T y$ will come from the errors and some from the model predictions.

By separating $y^T y$ into the two parts, we can compare them to see how well the model is doing.
More precisely, the sum of squares of the residuals can be shown (via linear algebra) to equal

$$SS_{Res} = (y - Xb)^T(y - Xb) = y^T y - y^T X(XX^T)^{-1}X^T y$$

which means that

$$y^T y = y^T X(XX^T)^{-1}X^T y + SS_{Res}.$$ 

We call $y^T X(XX^T)^{-1}X^T y$ the regression sum of squares and denote it by $SS_{Reg}$. It reflects the variation in the response variable that is accounted for by the model. If we call the total variation in the response variable $SS_{Total} = y^T y$, then we have divided it into:

$$SS_{Total} = SS_{Reg} + SS_{Res}.$$  

**Example.** Consider the paint cracking example of the earlier section. The data matrices are

\[
y = \begin{bmatrix} 1.9 \\ 2.7 \\ 4.2 \\ 4.8 \\ 4.8 \\ 5.1 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \end{bmatrix}
\]

and we estimated the variance to be

\[ s^2 = 0.27. \]

This means that

\[ SS_{Res} = (n - p)s^2 = (6 - 2)s^2 = 1.1. \]
Since

\[
SS_{Total} = y^T y = \begin{bmatrix}
1.9 & 2.7 & 4.2 & 4.8 & 4.8 & 5.1 \\
\end{bmatrix}
\begin{bmatrix}
1.9 \\
2.7 \\
4.2 \\
4.8 \\
4.8 \\
5.1 \\
\end{bmatrix} = 100.63,
\]

we get

\[
SS_{Reg} = SS_{Total} - SS_{Res} = 99.53.
\]

Since 99.53 > 1.1, informally we would say that the model is very appropriate!
**Example.** Suppose the model predicts the response variable exactly: then $y = X\beta$. We have

$$SS_{Reg} = y^T X (X^T X)^{-1} X^T y$$

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

$$= \beta^T X^T X \beta$$

$$= y^T y = SS_{Total}$$

and $SS_{Res} = 0$. 
On the other hand, suppose the model is useless: then $\beta = 0$ and $y = \varepsilon$. Suppose also that we know that the model is useless, so $b = 0$. Then

\[ SS_{Res} = (y - Xb)^T (y - Xb) = y^T y = SS_{Total} \]

and $SS_{Reg} = 0$.

These are the two extremes of the spectrum.
To create a formal test of $\beta = 0$, we compare $SS_{Reg}$ against $SS_{Res}$. If $SS_{Reg}$ is large compared to $SS_{Res}$, then the model is useful. If $SS_{Res}$ is large compared to $SS_{Reg}$, then the model is not useful.

To know exactly how large, we must first derive the distributions of $SS_{Reg}$ and $SS_{Res}$. Of course, we already know the latter (which we re-state).
Theorem

In the full rank linear model, $SS_{\text{Res}} / \sigma^2$ has a $\chi^2$ distribution with $n - p$ degrees of freedom.

Theorem

In the full rank linear model, $SS_{\text{Reg}} / \sigma^2$ has a noncentral $\chi^2$ distribution with $p = k + 1$ degrees of freedom and noncentrality parameter

$$\lambda = \frac{1}{2\sigma^2} \beta^T X^T X \beta.$$
Proof. By definition

\[ \frac{SS_{Reg}}{\sigma^2} = \frac{1}{\sigma^2} y^T X (X^T X)^{-1} X^T y. \]

By assumption, \( y \) is a normal random vector with mean \( X \beta \) and variance \( \sigma^2 I \). Also \( X (X^T X)^{-1} X^T \) is an idempotent and symmetric matrix. Therefore its rank is equal to its trace:

\[
\begin{align*}
    r(X(X^T X)^{-1}X^T) &= tr(X(X^T X)^{-1}X^T) \\
    &= tr((X^T X)^{-1}X^T X) = tr(I_{k+1}) \\
    &= k + 1.
\end{align*}
\]
By a previous linear algebra result, \( SS_{Reg}/\sigma^2 \) must have a noncentral \( \chi^2 \) distribution with \( k + 1 \) degrees of freedom and noncentrality parameter

\[
\lambda = \frac{1}{2\sigma^2} (X\beta)^T X (X^T X)^{-1} X^T (X\beta) \\
= \frac{1}{2\sigma^2} \beta^T X^T X \beta.
\]

We need one final result to create our test.
Theorem

In the full rank linear model, $\frac{SS_{Reg}}{\sigma^2}$ and $\frac{SS_{Res}}{\sigma^2}$ are independent.

This can be proved by observing that they are both quadratic forms in $y$ and by applying the result for independence of quadratic forms.
The test for $\beta = 0$ comes about when we observe that if the null hypothesis is true, the noncentrality parameter for $SS_{\text{Reg}}/\sigma^2$ must be 0.

This gives use two independent $\chi^2$ variables... and we know what to do with those!
The test for $\beta = 0$ comes about when we observe that if the null hypothesis is true, the noncentrality parameter for $SS_{Reg}/\sigma^2$ must be 0.

This gives us two independent $\chi^2$ variables... and we know what to do with those!

We divide them by each other of course!

$$\frac{SS_{Reg}/p\sigma^2}{SS_{Res}/(n-p)\sigma^2} = \frac{SS_{Reg}/p}{SS_{Res}/(n-p)}$$

has an $F$ distribution with $p$ and $n - p$ degrees of freedom (under $H_0$).
What happens if $H_0$ is not true? It can be shown that the expected value of the numerator (denoted $MS_{Reg}$) of this statistic is

$$E\left[\frac{SS_{Reg}}{p}\right] = \sigma^2 + \frac{1}{p} \beta^T X^T X \beta$$

and that because $X$ is of full rank, $X^T X$ is positive definite.

We already know the expected value of the denominator (which we denote $MS_{Res}$):

$$E\left[\frac{SS_{Res}}{n-p}\right] = E[s^2] = \sigma^2.$$
So if $\beta = 0$, $E\left[\frac{SS_{Reg}}{p}\right] = \sigma^2$ and the statistic should be close to 1.

But if $\beta \neq 0$, we get $E\left[\frac{SS_{Reg}}{p}\right] > \sigma^2$ and the statistic should be bigger than one.

Therefore, we should use a one-tailed test and reject $H_0$ if the statistic is large.
To lay out the workings, we use a familiar ANOVA table.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>degrees of freedom</th>
<th>Mean square</th>
<th>F ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$y^T X (X^T X)^{-1} X^T y$</td>
<td>$p$</td>
<td>$\frac{SS_{Reg}}{MS_{Res}}$</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>$y^T y - y^T X (X^T X)^{-1} X^T y$</td>
<td>$n - p$</td>
<td>$\frac{SS_{Res}}{n-p}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$y^T y$</td>
<td>$n$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example. A data processing system uses three types of structural elements: files, flows and processes. Files are permanent records, flows are data interfaces, and processes are logical manipulations of the data. The costs of developing software for the system are based on the number of these three elements. A study is conducted with the following results:

<table>
<thead>
<tr>
<th>Cost (y)</th>
<th>Files (x₁)</th>
<th>Flows (x₂)</th>
<th>Processes (x₃)</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.6</td>
<td>4</td>
<td>44</td>
<td>18</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>33</td>
<td>15</td>
</tr>
<tr>
<td>78.1</td>
<td>20</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>28</td>
<td>6</td>
<td>24</td>
<td>21</td>
</tr>
<tr>
<td>80.5</td>
<td>6</td>
<td>227</td>
<td>50</td>
</tr>
<tr>
<td>24.5</td>
<td>3</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>20.5</td>
<td>4</td>
<td>41</td>
<td>13</td>
</tr>
<tr>
<td>147.6</td>
<td>16</td>
<td>187</td>
<td>137</td>
</tr>
<tr>
<td>4.2</td>
<td>4</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>48.2</td>
<td>6</td>
<td>50</td>
<td>21</td>
</tr>
<tr>
<td>20.5</td>
<td>5</td>
<td>48</td>
<td>17</td>
</tr>
</tbody>
</table>
The model we use is

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i. \]

We want to test the hypothesis

\[ H_0 : \beta = 0 \text{ vs. } H_1 : \beta \neq 0. \]

Simple matrix calculations give us

\[
\begin{align*}
SS_{Reg} &= y^T X (X^T X)^{-1} X^T y = 38978 \\
y^T y &= 39667 \\
SS_{Res} &= y^T y - SS_{Reg} = 689 \\
MS_{Reg} &= SS_{Reg} / 4 = 9745 \\
MS_{Res} &= SS_{Res} / (11 - 4) = 98 \\
F_{4,7} &= MS_{Reg} / MS_{Res} = 99
\end{align*}
\]
The $F$ ratio is very large and we would expect $H_0$ to be rejected based on it. Indeed, the critical value for $\alpha = 0.01$ is 7.85, so we can say that $\beta \neq 0$ with (near) certainty.

<table>
<thead>
<tr>
<th>Variation</th>
<th>SS</th>
<th>d.f.</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>38978</td>
<td>4</td>
<td>9745</td>
<td>99</td>
</tr>
<tr>
<td>Residual</td>
<td>689</td>
<td>7</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>39667</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Testing a subvector of $\beta$

The previous test is good for testing if the entire model is adequate, but it does not tell us the full story.

If we find that $\beta \neq 0$, we cannot say which $\beta_i$ is nonzero — we do not need all of them to be nonzero! Even one will suffice.

So if we find that the model is adequate, we cannot say it is the best model. There may still be irrelevant variables.
A common example of this occurring is when the only relevant part of the model

\[ y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \]

is the $\beta_0$ term — in other words, the average $y$ is not 0, but it does not depend on the $x$’s.

In this case, we will find that $\beta \neq 0$, but the model is still terrible!

We need to find a way of testing whether parts of the parameter vector are 0 or not.
We need to be careful (and formal) about this: we split the parameter vector

\[
\beta = \begin{bmatrix}
\beta_0 \\
\vdots \\
\beta_{r-1} \\
\beta_r \\
\vdots \\
\beta_k 
\end{bmatrix} = \begin{bmatrix}
\gamma_1 \\
\gamma_2
\end{bmatrix}
\]

and test the hypotheses

\[H_0 : \gamma_1 = 0 \text{ vs. } H_1 : \gamma_1 \neq 0.\]

By relabelling the indices, we can test the zero-ness of any subset of the parameters.
The important thing to note is that we are testing $\gamma_1 = 0$ in the presence of the other parameters, not by itself.

In other words, we are comparing two models: in $H_1$, the full model

$$y = X\beta + \varepsilon,$$

and in $H_0$, the reduced model

$$y = X_2\gamma_2 + \varepsilon_2$$

where $X_2$ contains the last $k + 1 - r$ columns of $X$ (suitably re-ordered of course).
How do we do this? Recall that the amount of variation in $y$ explained by the full model is

$$SS_{Reg} = y^T X^T (X^T X)^{-1} X^T y.$$ 

We call this $R(\beta)$. If we consider the reduced model, the amount of variation explained is now

$$R(\gamma_2) = y^T X_2 (X_2^T X_2)^{-1} X_2^T y.$$ 

The difference is the amount of variation that is not random but cannot be accounted for by the reduced model.
We call this difference the \textit{sum of squares for regression on }\gamma_1 \textit{ in the presence of }\gamma_2:\n
\[ R(\gamma_1|\gamma_2) = R(\beta) - R(\gamma_2). \]

If \(H_0\) is true, the reduced model is sufficient and \(R(\gamma_1|\gamma_2)\) will be small. However, if \(H_1\) is true, \(R(\gamma_1|\gamma_2)\) will be large.

I think you can see where this is going!
We want to develop a test statistic that is based on $R(\gamma_1|\gamma_2)$. To do this we need to lay the necessary groundwork.

**Lemma**

The following matrix properties hold:

- The rank of $X_2(X_2^TX_2)^{-1}X_2^T$ is $p - r$;
- The matrix $X(X^TX)^{-1}X^T - X_2(X_2^TX_2)^{-1}X_2^T$ is idempotent and has rank $r$; and
- The rank of $I - X(X^TX)^{-1}X^T$ is $n - p$.

These are all easy to prove using our linear algebra framework.
We continue with a result very similar to one we examined before.

**Theorem (Cochran-Fisher Theorem)**

Let $\mathbf{z}$ be a $n \times 1$ normal random vector with mean $\mu$ and variance $I$. Decompose the sum of squares of $\mathbf{z}$ into the quadratic forms

$$
\mathbf{z}^T \mathbf{z} = \sum_{i=1}^{m} \mathbf{y}^T A_i \mathbf{y}.
$$

Then the quadratic forms are independent and have noncentral $\chi^2$ distributions with parameters $r(A_i)$ and $\frac{1}{2} \mu^T A_i \mu$, respectively, if and only if

$$
\sum_{i=1}^{m} r(A_i) = n.
$$
To apply this theorem, we let $z = y/\sigma$, which gives us

$$E[z] = \frac{X\beta}{\sigma}, \text{var } z = \frac{1}{\sigma^2} \text{var } y = I.$$ 

Then we split $z^Tz$ into the following quadratic forms:

$$z^Tz = \frac{1}{\sigma^2} y^T y$$

$$= \frac{y^T [X_2 (X_2^T X_2)^{-1} X_2^T] y}{\sigma^2}$$

$$+ \frac{y^T [X (X^T X)^{-1} X^T - X_2 (X_2^T X_2)^{-1} X_2^T] y}{\sigma^2}$$

$$+ \frac{y^T [I - X (X^T X)^{-1} X^T] y}{\sigma^2}.$$
The above matrix lemma shows that the ranks of the matrices sum to

\[(p - r) + r + (n - p) = n.\]

Therefore, all the quadratic forms are independent and have noncentral \(\chi^2\) distributions.

We are interested in particular in the middle form, which is 

\[\frac{1}{\sigma^2} R(\gamma_1 | \gamma_2).\]
We now know (and probably should have suspected) that it has a noncentral $\chi^2$ distribution with $r$ degrees of freedom and noncentrality parameter

$$\lambda = \frac{1}{2\sigma^2} \beta^T X^T [X(X^T X)^{-1} X^T - X_2(X_2^T X_2)^{-1} X_2^T] X \beta.$$  

We would like this to be 0 if $H_0$ is true, so we can generate an $F$ statistic.

**Theorem**

If $H_0$ is true, then $\frac{1}{\sigma^2} R(\gamma_1 | \gamma_2)$ has noncentrality parameter 0.
The proof is a simple (but tedious) matter of partitioning matrices and then setting $\gamma_1 = 0$.

Therefore our test statistic is

$$\frac{R(\gamma_1 | \gamma_2)/r}{SS_{Res}/(n - p)}.$$ 

It has an $F$ distribution with $r$ and $n - p$ degrees of freedom.

Again, we perform a one-sided test as we should reject $H_0$ if $R(\gamma_1 | \gamma_2)$ is large.
We can again express the test in an ANOVA table.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>degrees of freedom</th>
<th>Mean square</th>
<th>F ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full model</td>
<td>$R(\beta)$</td>
<td>$p$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced model</td>
<td>$R(\gamma_2)$</td>
<td>$p - r$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_1$ in presence of $\gamma_2$</td>
<td>$R(\gamma_1</td>
<td>\gamma_2)$</td>
<td>$r$</td>
<td>$\frac{R(\gamma_1</td>
</tr>
<tr>
<td>Residual</td>
<td>$y^T y - R(\beta)$</td>
<td>$n - p$</td>
<td>$\frac{SS_{Res}}{n-p}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$y^T y$</td>
<td>$n$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example. Consider the data processing system example given above. We rejected the null hypothesis $\beta = 0$. But that is obvious because the cost of a system is not going to have average 0!

The question we want to test is, does the cost depend on the files, flows or processes? In other words, is one of $\beta_1$, $\beta_2$, or $\beta_3$ nonzero?

To do this, we re-arrange the parameter vector as

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_0 \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}.$$
To keep things ‘in sync’, we must rearrange the columns of $X$:

$$
X = \begin{bmatrix}
4 & 44 & 18 & 1 \\
2 & 33 & 15 & 1 \\
\vdots & \vdots & \vdots & \vdots \\
5 & 48 & 17 & 1
\end{bmatrix}
= \begin{bmatrix}
X_1 \\
X_2
\end{bmatrix}.
$$

We want to test $H_0 : \gamma_1 = 0$ (the intercept alone is adequate) against $H_1 : \gamma_1 \neq 0$. The reduced model is

$$
y = X_2 \beta_0 + \varepsilon_2.
$$
The sum of squares of regression for the reduced model is

\[
R(\gamma_2) = y^T X_2 (X_2^T X_2)^{-1} X_2^T y \\
= (X_2^T y)^T (n)^{-1} X_2^T y \\
= \frac{1}{11} \left( \sum_{i=1}^{1} 1y_i \right)^2 \\
= 21800.
\]

From before,

\[
R(\beta) = SS_{Reg} = 38978, \ MS_{Res} = 98,
\]

so

\[
R(\gamma_1|\gamma_2) = R(\beta) - R(\gamma_2) = 38978 - 21800 = 17178.
\]
Our $F$ statistic is now

$$\frac{\frac{R(\gamma_1|\gamma_2)/r}{SS_{Res}/(n-p)}}{17178/3} = 58.2.$$ 

We check this against the $F$ distribution with 3 and $n-p=7$ degrees of freedom. The critical point for $\alpha = 0.01$ is 8.45, so we can again say that $H_0$ can be rejected.

In other words, the intercept alone does not explain the variation in the response variable adequately, and we are (reasonably) certain that we need at least one of the terms in the model.
<table>
<thead>
<tr>
<th>Variation</th>
<th>SS</th>
<th>d.f.</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full</td>
<td>38978</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced</td>
<td>21800</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_1$ in presence of $\gamma_2$</td>
<td>17178</td>
<td>3</td>
<td>5726</td>
<td>58.2</td>
</tr>
<tr>
<td>Residual</td>
<td>689</td>
<td>7</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>39667</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How do we decide which parameter sets to test? There are many ways to partition the parameters.

With the help of a computer, we can test all the possible parameter sets.

It is much more useful to select parameter sets based on knowledge of the subject. This will tell us which parameters are certain to be in the model, and which may or may not be.
Corrected sum of squares

Usually it is pretty obvious that \( y \) does not have an average of 0, so we can say right off the bat that \( \beta_0 \) is nonzero. In this case, there isn’t much point in including \( \beta_0 = 0 \) in any of our tests!

The model adequacy test then becomes, are all the other parameters zero in the presence of the intercept (or not)?

This is the same test that we performed in the previous example. As we noted there, the reduced model regression sum of squares can be written as

\[
R(\gamma_2) = \frac{1}{n} \left( \sum_{i=1}^{n} y_i \right)^2.
\]
The ANOVA table for this test is:

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>degrees of freedom</th>
<th>Mean square</th>
<th>F ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full model</td>
<td>$R(\beta)$</td>
<td>$k + 1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced model</td>
<td>$(\sum_{i=1}^{n} y_i)^2 / n$</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_1$ in presence of $\gamma_2$</td>
<td>$R(\gamma_1</td>
<td>\gamma_2)$</td>
<td>$k$</td>
<td>$\frac{R(\gamma_1</td>
</tr>
<tr>
<td>Residual</td>
<td>$y^T y - R(\beta)$</td>
<td>$n - k - 1$</td>
<td>$\frac{\text{SS}_{\text{Res}}}{n - p}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$y^T y$</td>
<td>$n$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
However, we can also think of it in a different way. If we are assuming that $\beta_0$ is nonzero, then we don’t have to explain all the variation in the response variable. Instead, we can just explain the variation around the mean,

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = y^T y - R(\gamma_2).$$

This is called the *corrected sum of squares*, and $R(\gamma_2)$ is the *correction factor*.

Because we have to estimate the mean, we lose one degree of freedom (it is now $n – 1$).
Then we break down the corrected sum of squares into \( R(\gamma_1|\gamma_2) \) and \( SS_{Res} \), and test using an \( F \) statistic ratio. The end result is the same, but the table looks slightly different.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>degrees of freedom</th>
<th>Mean square</th>
<th>( F ) ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>( SS_{Reg} - (\sum_{i=1}^{n} y_i)^2 / n )</td>
<td>( k )</td>
<td>( R(\gamma_1</td>
<td>\gamma_2) )</td>
</tr>
<tr>
<td>Residual</td>
<td>( SS_{Res} )</td>
<td>( n - k - 1 )</td>
<td>( SS_{Res} )</td>
<td>( n-k-1 )</td>
</tr>
<tr>
<td>Total</td>
<td>( y^T y - (\sum_{i=1}^{n} y_i)^2 / n )</td>
<td>( n - 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some computer outputs will use a corrected sum of squares layout instead of an uncorrected sum, so you should be familiar with both.
Example. In the data processing example, we rejected the hypothesis that $[\beta_1 \beta_2 \beta_3] = 0$. The ANOVA table for a corrected sum of squares test is

<table>
<thead>
<tr>
<th>Variation</th>
<th>SS</th>
<th>d.f.</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>17178</td>
<td>3</td>
<td>5726</td>
<td>58.2</td>
</tr>
<tr>
<td>Residual</td>
<td>689</td>
<td>7</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>17867</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The actual test does not change — the $F$ statistic and degrees of freedom are the same.
Right-tailed test

The next theorem formally shows that we should be using a right-tailed test. Its proof uses our linear algebra results.

Theorem

Let \( X \) be a \( n \times p \) full rank matrix. Let \( X \) and \( \beta \) be partitioned as

\[
X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}, \quad \beta = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix},
\]

where \( X_1 \) is \( n \times r \) and \( \gamma_1 \) is \( r \times 1 \). Then

\[
R(\gamma_1|\gamma_2) = \hat{\gamma}_1^T A_{11}^{-1} \hat{\gamma}_1,
\]

where \( \hat{\gamma}_1 \) is the least squares estimator for \( \gamma_1 \), and \( A_{11} \) is the \( r \times r \) principal minor of \( (X^T X)^{-1} \):

\[
A_{11}^{-1} = X_1^T X_1 - X_1^T X_2 (X_2^T X_2)^{-1} X_2^T X_1.
\]
Now the $F$ statistic we use for testing $\gamma_1$ in the presence of $\gamma_2$ is

$$\frac{R(\gamma_1|\gamma_2)/r}{SS_{Res}/(n-p)} = \frac{\gamma_1^TA_{11}^{-1}\gamma_1/r}{SS_{Res}/(n-p)}.$$

We know that $E[\frac{SS_{Res}}{n-p}] = \sigma^2$. It is also easy to work out that the least squares estimator for $\gamma_1$ has the properties

$$E[\hat{\gamma}_1] = \gamma_1,$$
$$\text{var } \hat{\gamma}_1 = A_{11}\sigma^2.$$
Now we can say that

\[
E[\frac{\hat{\gamma}_1^T A_{11}^{-1} \hat{\gamma}_1}{r}] = \frac{1}{r} [tr(A_{11}^{-1} A_{11} \sigma^2) + \gamma_1^T A_{11}^{-1} \gamma_1]
\]

\[
= \frac{1}{r} [tr(I_r \sigma^2) + \gamma_1^T A_{11}^{-1} \gamma_1]
\]

\[
= \sigma^2 + \frac{1}{r} \gamma_1^T A_{11}^{-1} \gamma_1.
\]

Now \(A_{11}\) is a principal minor of the positive definite matrix \((X^T X)^{-1}\), and so both it and its inverse are positive definite. Thus

\[
\frac{1}{r} \gamma_1^T A_{11}^{-1} \gamma_1 \geq 0
\]

with equality happening at \(\gamma_1 = 0\). Therefore we should indeed be using a right-tailed test.
In the previous section, we showed how to test if any number of parameters are 0, given that the rest of the parameters are in the model.

Often we are interested in whether a particular parameter is 0 or not, in the presence of all the other parameters. This is called a \textit{partial} test.

We can do this using the methodology developed in the previous section.
Ideally, we would like to do this for every parameter, to determine the best subset of parameters to use in the final model. However, there is a difficulty: the regression sums of squares for these tests do not add up to the full regression sum of squares

$$R(\beta) \neq R(\beta_0|\beta_1, \ldots, \beta_k) + R(\beta_1|\beta_0, \beta_2, \ldots, \beta_k) + \ldots + R(\beta_k|\beta_0, \ldots, \beta_{k-1})$$

So we are *not* breaking down the regression sum of squares into a part which is explained by $\beta_0$, a part which is explained by $\beta_1$, etc., and testing each of these individually.
This is because we are testing the importance of each parameter in a model which contains all the parameters.

Therefore, acceptance or rejection of $H_0$ does not mean that the parameter is useful or useless in the best model, just useful or useless in the full model.

What we would like to do is to start with a simple model and sequentially add parameters until we reach the full model, rather than just comparing the full model with a nearly-full model.
With possibly some relabelling, we look at the series of models

\[ \begin{align*}
    y &= \beta_0 + \varepsilon^{(0)} \\
    y &= \beta_0 + \beta_1 x_1 + \varepsilon^{(1)} \\
    &\vdots \\
    y &= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon^{(k)}.
\end{align*} \]

We denote the corresponding \( X \) matrices by \( X^{(j)} \), which are the first \( j + 1 \) columns of \( X \).

The regression sum of squares for each of these models is calculated in the usual way:

\[ R(\beta_0, \beta_1, \ldots, \beta_j) = y^T X^{(j)}((X^{(j)})^T X^{(j)})^{-1}(X^{(j)})^T y. \]
Note that these are ‘full’ regression sums of squares, i.e. we are looking at the total variation explained by the model in the presence of no other parameters.

Now by taking the difference between the sums of squares, we can get the extra variation explained as we add variables to the model one at a time:

\[
R(\beta_1|\beta_0) = R(\beta_0, \beta_1) - R(\beta_0)
\]
\[
R(\beta_2|\beta_0, \beta_1) = R(\beta_0, \beta_1, \beta_2) - R(\beta_0, \beta_1)
\]
\[\vdots\]
\[
R(\beta_k|\beta_0, \beta_1, \ldots, \beta_{k-1}) = R(\beta) - R(\beta_0, \beta_1, \ldots, \beta_{k-1}).
\]
These regression sum of squares do add up to the total:

\[ R(\beta) = R(\beta_0) + R(\beta_1|\beta_0) + R(\beta_2|\beta_0, \beta_1) + \ldots + R(\beta_k|\beta_0, \beta_1, \ldots, \beta_{k-1}) \]

so we can decompose the regression sum of squares like this and test each part using an \( F \) test to assess the usefulness of each parameter.
Each parameter set has 1 more parameter than the previous one, so each sequential regression sum of squares has 1 degree of freedom. Therefore we test (for each $j$)

$$R(\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1}) \frac{SS_{Res}}{(n - p)}$$

which have $F$ distributions with 1 and $n - p$ degrees of freedom.

This is still not entirely satisfactory, because the result will depend heavily on the order of the parameters. We can either try all possible orders (a bit hard!), or try to arrange them in a sensible way using knowledge of the data.
Example. An experiment was conducted to study the size of squid. The response is the weight of the squid, and the predictors are

- $x_1$: Beak length
- $x_2$: Wing length
- $x_3$: Beak to notch length
- $x_4$: Notch to wing length
- $x_5$: Width

A total of 22 squid are sampled. We skip over the data/calculations...
The first thing we test is whether $\beta = 0$. The ANOVA table is

<table>
<thead>
<tr>
<th>Variation</th>
<th>SS</th>
<th>d.f.</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>595.16</td>
<td>6</td>
<td>99.19</td>
<td>200.47</td>
</tr>
<tr>
<td>Residual</td>
<td>7.92</td>
<td>16</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>603.08</td>
<td>22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The null hypothesis $\beta = 0$ is rejected strongly (at $\alpha = 0.01$ the critical value is 4.20)!
Next we test to see which parameters should be in the model. The sequential sums of squares are:

\[
\begin{align*}
R(\beta_0) &= 387.16 \\
R(\beta_1|\beta_0) &= 199.15 \\
R(\beta_2|\beta_0, \beta_1) &= 0.127 \\
R(\beta_3|\beta_0, \beta_1, \beta_2) &= 4.12 \\
R(\beta_4|\beta_0, \beta_1, \beta_2, \beta_3) &= 0.263 \\
R(\beta_5|\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) &= 4.35
\end{align*}
\]

Note that these sum to the regression sum of squares for the full model, 595.16.
Each of these sums of squares should be compared against the critical $F$ value with 1 and $n - p$ degrees of freedom, multiplied by $SS_{Res}/(n - p)$. With $\alpha = 0.01$, this is

$$\frac{7.92}{22 - 6} \times 8.53 = 4.22.$$

So in a model with no parameters, we should definitely add $\beta_0$ and then $\beta_1$, but not $\beta_2$. We should probably add $\beta_3$ and $\beta_5$ as well.

The $\beta_4$ test is a little harder to interpret: if $\beta_0, \beta_1, \beta_2$, and $\beta_3$ is in the model, we should not add it. But $\beta_2$ is not in the model!

In this case, we should test the usefulness of $\beta_4$ given the parameters that we do have in the model.
We can also use a $t$ test for a partial test of one parameter. We are testing $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$ in the presence of all the other parameters.

Remember that we developed a confidence interval on $\beta_i$ to be

$$b_i \pm t_{\alpha/2} s \sqrt{c_{ii}}$$

where $c_{ii}$ is the $(i, i)$th entry of $(X^T X)^{-1}$, and we use a $t$ distribution with $n - p$ degrees of freedom. If this confidence interval includes 0, we do not reject $H_0$; otherwise, we can reject it.
In other words, we use the $t$ statistic (with $n - p$ degrees of freedom)

$$\frac{b_i}{s\sqrt{c_{ii}}}.$$ 

Let us compare this with our existing partial $F$ test. The statistic we use for this is

$$R(\beta_i|\beta_0, \beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_k) \quad \frac{SS_{Res}/(n - p)}{SS_{Res}/(n - p)}.$$ 

The denominator is of course $s^2$. 
We see from above that the numerator is

\[ R(\beta_i|\beta_0, \beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_k) = \hat{\gamma}_1^T A_{11}^{-1} \hat{\gamma}_1 \]

where \( \hat{\gamma}_1 = b_i \), and \( A_{11} \) is the top left element of \((X^TX)^{-1}\) after the columns have been re-arranged so that the \( i \)th column comes first. In other words, \( A_{11} = c_{ii} \) and

\[ R(\beta_i|\beta_0, \beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_k) = b_i (c_{ii})^{-1} b_i = \frac{b_i^2}{c_{ii}}. \]

So the statistic (using an \( F \) distribution with 1 and \( n - p \) degrees of freedom) is

\[ \frac{b_i^2}{c_{ii} s^2}. \]

This is exactly the square of the \( t \) statistic!
This is actually not too surprising. The $t$ distribution can be expressed as a normal variable divided by the square root of a $\chi^2$ variable, and therefore when we square it we get the square of a normal variable divided by a $\chi^2$ variable. But the square of a normal variable is a $\chi^2$ variable with 1 d.f.

Therefore the square of a $t$ variable with $n$ d.f. is an $F$ variable with 1 and $n$ d.f.

This means that the $t$ test and the $F$ test are (nearly) identical; the $t$ test is actually slightly more useful, because it also gives an indication of the sign of the parameter.
Example. In the inference section, we modelled the amount of a chemical which dissolves in water, held at a certain temperature. We found that the 95% confidence interval for $\beta_1$ was

$$0.31 \pm 2.78 \times 0.86 \sqrt{0.00057} = [0.25, 0.36].$$

A $t$ test would use the statistic

$$\frac{b_1}{s \sqrt{c_{11}}} = \frac{0.31}{0.86 \sqrt{0.00057}} = 14.89$$

which, using a $t$ distribution with $n - p = 6 - 2 = 4$ degrees of freedom, would reject the hypothesis $\beta_1 = 0$ at the 0.05 level (critical value 2.78). We can also say that $\beta_1$ is almost certainly positive.
On the other hand, if we use an $F$ test then we find that

\[ R(\beta) = y^T X (X^T X)^{-1} X^T y = 663.77 \]

\[ R(\beta_0) = y^T X_1 (X_1^T X_1)^{-1} X_1^T y = 498.68 \]

\[ R(\beta_1 | \beta_0) = 663.77 - 498.68 = 165.09 \]

and the $F$ statistic is

\[ \frac{R(\beta_1 | \beta_0)}{s^2} = \frac{165.09}{0.74} = 221.7 = 14.89^2. \]
Based on the $F$ distribution with 1 and 4 degrees of freedom, the critical value is $7.71 = 2.77^2$. So we can again reject the null hypothesis of $\beta_1 = 0$.

<table>
<thead>
<tr>
<th>Variation</th>
<th>SS</th>
<th>d.f.</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full</td>
<td>663.77</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced</td>
<td>498.68</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$ in presence of $\beta_0$</td>
<td>165.09</td>
<td>1</td>
<td>165.09</td>
<td>221.7</td>
</tr>
<tr>
<td>Residual</td>
<td>2.98</td>
<td>4</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>666.75</td>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A more general hypothesis

The framework we have constructed allows us to test that the entire parameter set, or some of it, is 0 or not.

What if we want to test whether the parameter set is some specified nonzero value?

In other words, we want to test $H_0 : \beta = \beta^*$ against $H_1 : \beta \neq \beta^*$. 
To generate an appropriate test for this, we consider the least squares estimator \( \mathbf{b} \). Under the traditional assumptions, \( \mathbf{b} \) is a normal random vector with mean \( \mathbf{\beta} \) and variance \( (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \).

Now define

\[
\mathbf{z} = \mathbf{b} - \mathbf{\beta}^*.
\]

Obviously \( \mathbf{z} \) is a normal random vector with mean \( \mathbf{\beta} - \mathbf{\beta}^* \) and variance \( (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \).

Let \( \mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} \) and consider the quadratic form

\[
\mathbf{z}^T\mathbf{A}\mathbf{z} = \frac{(\mathbf{b} - \mathbf{\beta}^*)^T\mathbf{X}^T\mathbf{X}(\mathbf{b} - \mathbf{\beta}^*)}{\sigma^2}.
\]
We know that $A$ is a symmetric $p \times p$ matrix, and

$$A(X^T X)^{-1} \sigma^2 = \frac{1}{\sigma^2} X^T X (X^T X)^{-1} \sigma^2 = I_p$$

is idempotent of rank $p$. Therefore the quadratic form has a noncentral $\chi^2$ distribution with $p$ degrees of freedom and noncentrality parameter

$$\lambda = \frac{1}{2\sigma^2} (\beta - \beta^*)^T X^T X (\beta - \beta^*).$$

If the null hypothesis is true, $\beta = \beta^*$ and $\lambda = 0$, which means that the quadratic form has a $\chi^2$ distribution with $p$ d.f.
We would expect from previous derivations that we would now end up with an $F$ statistic

$$\frac{(b - \beta^*)^T X^T X (b - \beta^*)/p}{SS_{Res}/(n - p)}.$$ 

To show that this does have an $F$ distribution, we need to show that the numerator and denominator are independent. 

But the numerator depends only on $b$ (randomness-wise), and the denominator is $s^2$. We know that they are independent, so the numerator and denominator are independent.
Therefore the statistic
\[
\frac{(\mathbf{b} - \mathbf{\beta}^*)^T \mathbf{X}^T \mathbf{X} (\mathbf{b} - \mathbf{\beta}^*)}{p} / \frac{SS_{\text{Res}}}{n - p}
\]

has an \( F \) distribution with \( p \) and \( n - p \) degrees of freedom (under the null hypothesis).

We know the expected value of the denominator is \( \sigma^2 \). The expected value of the numerator is:
\[ E \left[ \frac{(\mathbf{b} - \beta^*)^T X^T X (\mathbf{b} - \beta^*)}{p} \right] \]
\[ = \frac{1}{p} \left[ tr(X^T X (X^T X)^{-1} \sigma^2) + (\beta - \beta^*)^T X^T X (\beta - \beta^*) \right] \]
\[ = \frac{1}{p} \left[ p \sigma^2 + (\beta - \beta^*)^T X^T X (\beta - \beta^*) \right] \]
\[ = \sigma^2 + \frac{1}{p} (\beta - \beta^*)^T X^T X (\beta - \beta^*). \]

Since \( X^T X \) is positive definite (\( X \) is of full rank), the second term is non-negative, and 0 if and only if \( \beta = \beta^* \), i.e. \( H_0 \) is true. Therefore we can use a one-tailed test and reject \( H_0 \) when the \( F \) statistic is large.
In a similar vein, it can be shown that if we want to test a subvector of $\beta$, $H_0 : \gamma_1 = \gamma_1^*$, where $\gamma_1$ is $r \times 1$, then the appropriate statistic is an $F$ statistic with $r$ and $n - p$ degrees of freedom,

\[
\frac{(\hat{\gamma}_1 - \gamma_1^*)^T A_{11}^{-1} (\hat{\gamma}_1 - \gamma_1^*) / r}{SS_{Res} / (n - p)},
\]

where

\[
A_{11}^{-1} = X_1^T X_1 - X_1^T X_2 (X_2^T X_2)^{-1} X_2^T X_1.
\]
**Example.** Consider the data processing system example we looked at earlier. Suppose we want to test the hypothesis $H_0 : \beta = \begin{bmatrix} 2 & 0 & 0 & 1 \end{bmatrix}^T$. The least squares estimate is

$$
\mathbf{b} = (X^T X)^{-1} X^T y = \begin{bmatrix} 1.96 \\ 0.12 \\ 0.18 \\ 0.8 \end{bmatrix}
$$

so

$$
\mathbf{b} - \beta^* = \begin{bmatrix} -0.04 \\ 0.12 \\ 0.18 \\ -0.2 \end{bmatrix}.
$$
Our calculations proceed:

\[(b - \beta^*)^T X^T X (b - \beta^*) = 1110.18\]

\[SS_{Res} = y^T [I - X(X^T X)^{-1} X^T] y = 668.63\]

\[p = 4\]

\[F_{4,7} = \frac{1110.18/4}{668.63/7} = 2.8.\]

The critical value at \(\alpha = 0.05\) is 4.12, so we cannot reject the null hypothesis. This doesn’t mean that it is true — just that it is close!
We can now progress to testing the general linear hypothesis, which tests

\[ H_0 : C\beta = \delta^* \text{ vs. } H_1 : C\beta \neq \delta^* \]

where \( C \) is an \( r \times p \) matrix of rank \( r \leq p \) and \( \delta^* \) is an \( r \times 1 \) vector of constants.

This hypothesis makes it possible to test for relationships among the parameters, as well as testing the individual parameters against a constant.
**Example.** Consider the regression model with 4 parameters (3 predictors)

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i. \]

Let

\[ C = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \delta^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \]

Suppose we want to test \( H_0 : C\beta = \delta^*. \) Then what we are really testing for is

\[ \beta_1 - \beta_2 = 0 \\
\beta_2 - \beta_3 = 0. \]

In other words, we are testing the hypothesis \( \beta_1 = \beta_2 = \beta_3. \)
To develop a test statistic, we start with $Cb - \delta^*$, the least squares estimator for $C\beta - \delta^*$. Because it is a vector containing linear combinations of variables which have a joint normal distribution, it is a normal random vector. The mean and variance rules can be applied to show that

$$E[Cb - \delta^*] = C\beta - \delta^*, \quad \text{var} (Cb - \delta^*) = C(X^TX)^{-1}C^T\sigma^2.$$
Therefore, the quadratic form

$$\frac{(C\beta - \delta^*)^T [C(X^T X)^{-1} C^T]^{-1} (C\beta - \delta^*)}{\sigma^2}$$

has a noncentral $\chi^2$ distribution with $r$ degrees of freedom and noncentrality parameter

$$\lambda = \frac{(C\beta - \delta^*)^T [C(X^T X)^{-1} C^T]^{-1} (C\beta - \delta^*)}{2\sigma^2}.$$
If the null hypothesis is true, then $C\beta = \delta^*$ and the quadratic form has a $\chi^2$ distribution.

Since the numerator depends (stochastically) only on $b$, and therefore is independent from $s^2$, we now know that the statistic

$$
\frac{(Cb - \delta^*)^T [C(X^T X)^{-1} C^T]^{-1} (Cb - \delta^*)}{r} \frac{SS_{Res}}{(n - p)}
$$

has an $F$ distribution with $r$ and $n - p$ degrees of freedom, under the null hypothesis. We use this statistic to test the general linear hypothesis.
To justify a one-tailed test, it can be shown that the expected value of the numerator is

\[
E\left[\frac{(Cb - \delta^*)^T [C(X^T X)^{-1} C]^T]^{-1}(Cb - \delta^*)}{r}\right]
\]

\[
= \sigma^2 + \frac{1}{r} (C\beta - \delta^*)^T [C(X^T X)^{-1} C]^T]^{-1}(C\beta - \delta^*).
\]

Furthermore, \(C(X^T X)^{-1} C^T\) is positive definite.

If the null hypothesis is true, then the expectation is \(\sigma^2\). However, if it is false, the expectation will be greater than \(\sigma^2\). Therefore we reject \(H_0\) when the statistic is large.
**Example.** Again consider the data processing system example. Suppose we wish to test the hypothesis given in the previous example, $H_0 : C\beta = \delta^*$ where

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \delta^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$ 

*Out least squares estimates are (still)*

$$b = (X^T X)^{-1} X^T y = \begin{bmatrix} 1.96 \\ 0.12 \\ 0.18 \\ 0.8 \end{bmatrix}.$$
Therefore

\[
Cb - \delta^* = \begin{bmatrix}
-0.06 \\
-0.62
\end{bmatrix}
\]

\[
C(X^TX)^{-1}C^T = \begin{bmatrix}
0.013 & 0.0024 \\
0.0024 & 0.00077
\end{bmatrix}
\]

\[
(Cb - \delta^*)^T [C(X^TX)^{-1}C^T]^{-1}(Cb - \delta^*) = 1138.35
\]
Since $SS_{Res}$ was calculated earlier to be 668.63, our $F$ statistic (with 2 and 7 degrees of freedom) is

$$\frac{1138.35/2}{668.63/7} = 5.79.$$ 

The critical value for $\alpha = 0.01$ is 9.55, so we cannot reject the null hypothesis.

In other words we are not certain that the parameters $\beta_1, \beta_2,$ and $\beta_3$ are not identical — the contribution to the response variable from each of the explanatory factors may be the same.
Example. Consider the null hypothesis of model adequacy, $H_0 : \beta = 0$. We can express this in the form of the general linear hypothesis with $C = I_p$ (which has rank $p$) and $\delta^* = 0$. The test statistic then becomes

$$
\frac{(Cb - \delta^*)^T [C(X^T X)^{-1} C^T]^{-1} (Cb - \delta^*)}{SS_{Res}/(n - p)} = \frac{b^T X^T X b}{SS_{Res}/(n - p)}
= \frac{[(X^T X)^{-1} X^T y]^T X^T X [(X^T X)^{-1} X^T y]}{SS_{Res}/(n - p)}$$
\begin{align*}
  & \frac{y^T X (X^T X)^{-1} X^T X (X^T X)^{-1} X^T y}{p} \\
  &= \frac{SS_{Res}/(n - p)}{SS_{Res}/(n - p)} \\
  &= \frac{SS_{Reg}/p}{SS_{Res}/(n - p)}.
\end{align*}

This is the same statistic that we derived earlier.

Unsurprisingly, all the other hypotheses in this section can also be expressed in terms of the general linear hypothesis. However, I thought it best to work up to it rather than throwing it at you all in one go!
Recall the $F$ statistic for testing a subvector of $\beta$:

$$\frac{R(\gamma_1|\gamma_2)/r}{SS_{Res}/(n-p)}$$

where $\gamma_1$ is $r \times 1$ and

$$R(\gamma_1|\gamma_2) = R(\beta) - R(\gamma_2) = y^T X (X^T X)^{-1} X^T y - y^T X_2 (X_2^T X_2)^{-1} X_2^T y.$$  

This is the variation in $y$ that is explained by $\gamma_1$ in the presence of $\gamma_2$. Now consider what happens if $X_1$ and $X_2$ are orthogonal, i.e.

$$X_1^T X_2 = 0.$$
Then

\[ X^T X = \begin{bmatrix} \frac{X_1^T}{X_2^T} \\ \frac{X_1^T X_1}{X_2^T X_1} \end{bmatrix} \begin{bmatrix} X_1 \mid X_2 \end{bmatrix} = \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} = \begin{bmatrix} X_1^T X_1 & 0 \\ 0 & X_2^T X_2 \end{bmatrix} \]

and

\[ (X^T X)^{-1} = \begin{bmatrix} \frac{(X_1^T X_1)^{-1}}{0} & 0 \\ 0 & (X_2^T X_2)^{-1} \end{bmatrix}. \]
\begin{align*}
R(\gamma_1 | \gamma_2) &= y^T X (X^T X)^{-1} X^T y - y^T X_2 (X_2^T X_2)^{-1} X_2^T y \\
&= y^T \left[ \begin{array}{c|c}
X_1 & X_2 \\
\end{array} \right] \left[ \begin{array}{c|c}
(X_1^T X_1)^{-1} & 0 \\
0 & (X_2^T X_2)^{-1} \\
\end{array} \right] \left[ \begin{array}{c}
\frac{X_1^T}{X_2^T} \\
\end{array} \right] y \\
&\quad - y^T X_2 (X_2^T X_2)^{-1} X_2^T y \\
&= y^T \left[ \begin{array}{c|c}
X_1 & X_2 \\
\end{array} \right] \left[ \frac{(X_1^T X_1)^{-1} X_1^T}{(X_2^T X_2)^{-1} X_2^T} \right] y - y^T X_2 (X_2^T X_2)^{-1} X_2^T y \\
&= y^T [X_1 (X_1^T X_1)^{-1} X_1^T + X_2 (X_2^T X_2)^{-1} X_2^T] y - y^T X_2 (X_2^T X_2)^{-1} X_2^T y \\
&= y^T X_1 (X_1^T X_1)^{-1} X_1^T y \\
&= R(\gamma_1).
\end{align*}
Therefore, if $X_1$ and $X_2$ are orthogonal, the variation explained by $\gamma_1$ in the presence of $\gamma_2$ is the same as the variation explained by $\gamma_1$ by itself!

This comes in handy in many different ways.

One way we can use this is to simplify our computation, so that we can test $\gamma_1 = 0$ using the statistic

$$\frac{R(\gamma_1)/r}{SS_{Res}/(n - p)}.$$
The real strength of orthogonality lies in the ability to break down the regression sum of squares into separate components:

$$ R(\beta) = R(\gamma_1 | \gamma_2) + R(\gamma_2) = R(\gamma_1) + R(\gamma_2). $$

So the total variation explained by the model is the variation explained by the model with $\gamma_1$ only, added to the variation explained by the model with $\gamma_2$ only.

This becomes even stronger if all of the columns of $X$ are orthogonal (to each other). In this case, we say that the model is *mutually orthogonal*.
In this case, the regression sum of squares can be broken down into contributions from individual parameters:

\[ R(\beta) = R(\beta_0) + R(\beta_1) + R(\beta_2) + \ldots + R(\beta_k). \]

This makes sequential and partial tests equivalent — and they will both be based on \( R(\beta_i) \).

In other words, the impact of each parameter can be measured without dependence on the other parameters — it does not matter if they are present or absent.

It is advisable (if possible) to ensure that the model is mutually orthogonal. This is called *orthogonal design*. 
Example. Suppose that for a 4-parameter linear model we have

\[ y = \begin{bmatrix} 57 \\ 40 \\ 19 \\ 40 \\ 54 \\ 41 \\ 21 \\ 43 \\ 63 \\ 28 \\ 11 \\ 2 \\ 18 \\ 56 \\ 46 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & -2 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & -2 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & -2 \\ 1 & 0 & 0 & 2 \end{bmatrix} \]
It is easy to see that the model is mutually orthogonal. If we want to test \( H_0 : \beta_1 = 0 \), then the numerator of the \( F \) statistic is

\[
R(\beta_1|\beta_0, \beta_2, \beta_3) = R(\beta_1) = y^T X_1 (X_1^T X_1)^{-1} X_1^T y.
\]

Now \( X_1 \) is the second column of \( X \), and

\[
X_1^T X_1 = 16, \quad (X_1^T X_1)^{-1} = \frac{1}{16}
\]

\[
X_1^T y = -57 + 40 - 19 + 40 - 54 + 41 - 21 + 43 - 2 \times 28 + 2 \times 11 = -21,
\]

so

\[
R(\beta_1) = (-21) \times \frac{1}{16} \times (-21) = 27.56.
\]