We compare models by comparing the sums of squares of residuals.

**Extrusion and Wear**

*ex8_3_15* has data on the relationship between extrusion pressure (x, in KPa) and wear (y, in mg).

A plot suggests a quadratic model. Here we fit a quadratic model to the data.

```r
> l=lm(y~x+I(x^2),data=ex8_3_15)
> xscale=data.frame(x=seq(min(ex8_3_15$x),max(ex8_3_15$x),length.out=101))
> plot(y~x,data=ex8_3_15)
> lines(predict(l,xscale)~xscale$x,data=ex8_3_15)
```

Just by looking at the graph, it is clear that the quadratic is a good fit – better than any linear fit and probably about as good as a cubic polynomial will do. We’ll look at the models given by polynomials of 0,1,2,3 degree.

```r
> l0=lm(y~1,data=ex8_3_15)
> l1=lm(y~x,data=ex8_3_15)
> l2=lm(y~x+I(x^2),data=ex8_3_15)
> l3=lm(y~x+I(x^2)+I(x^3),data=ex8_3_15)
> SSE0=sum(residuals(l0)^2)
> SSE1=sum(residuals(l1)^2)
> SSE2=sum(residuals(l2)^2)
> SSE3=sum(residuals(l3)^2)
> SSE0; SSE1; SSE2; SSE3
[1] 16.708
[1] 14.617
[1] 0.52914
[1] 0.46492
```

From the differences in sums of squares of residuals, one can see that the second degree model represents a considerable improvement over the first degree model but the third degree model represents minimal improvement.
We will compare two models: a full model and a reduced model. For example, the full model might have more variables than the reduced model. For example, in the multiple regression case we have:

**Reduced model** \[ y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \]

**Full model** \[ y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + \cdots + \beta_p x_p + \varepsilon \]

We want to test the hypothesis that the full model adds explanatory value over the reduced model. That hypothesis is

\[ H_0 : \beta_{k+1} = \cdots = \beta_p = 0 \]

The statistic we use is the following:

\[
F = \frac{(\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}})/(p - k)}{\text{SSE}_{\text{full}}/(n - p - 1)}
\]

If the null hypothesis is true, F is expected to be about 1 whereas if the null hypothesis is false then F will tend to be larger. Under the null hypothesis (and all the other assumptions of multiple regression), the distribution of the F statistic is known and so we can compute a P-value associated with the null hypothesis.

The relevant values are produced by the `anova` function in R. This function compares a reduced model to a full model.

\[
> \text{anova(11,12)}
\]

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Model 1: y ~ x</th>
<th>Model 2: y ~ x + I(x^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res.Df</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>RSS</td>
<td>14.62</td>
<td>0.53</td>
</tr>
<tr>
<td>Df Sum of Sq</td>
<td>14.1</td>
<td>79.9</td>
</tr>
<tr>
<td>F Pr(&gt;F)</td>
<td>0.003**</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

\[
> \text{anova(12,13)}
\]

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Model 1: y ~ x + I(x^2)</th>
<th>Model 2: y ~ x + I(x^2) + I(x^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res.Df</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>RSS</td>
<td>0.529</td>
<td>0.465</td>
</tr>
<tr>
<td>Df Sum of Sq</td>
<td>0.0642</td>
<td>0.28</td>
</tr>
<tr>
<td>F Pr(&gt;F)</td>
<td>0.65</td>
<td>0.28</td>
</tr>
</tbody>
</table>

These ANOVA tables tell us that there is good reason to use a quadratic model over the linear model but not a cubic model over the quadratic model.
Runs

The Mathematics 241 dataframe a10106 has data on all American League baseball teams for the 2001-2006 seasons. We would like to predict the number of runs that a team will score from certain other statistics. Here are the variables that we might use:

- R  Runs
- H  Hits
- AB At bats
- HR Home runs
- X2B Doubles
- X3B Triples
- BB Walks (base-on-balls)
- SO Strikeouts
- SB Stolen bases
- CS Caught stealing

Investigate different possible models using these variables to predict runs R. Compare the models using ANOVA and write down what you think to be the best model and why.

For each model, try to interpret the coefficients using what you know about baseball. Remember that hits, home runs, doubles, triples, walks and stolen bases are good things and caught stealing and strikeouts are bad things. Also remember that some of the variables overlap: for example every HR is also a H and every H is also an AB.