Chapter 4:

Vineyard Soil Data: Relationship between pH and Organic Matter and EC

A matrix plot is a good plot to explore relationships among several variables.

> pairs(subset(soil, select = c('pH', 'OM', 'EC')), pch = 16)

The upper right plot is the plot we’re familiar with: pH vs. OM. Notice that EC (electrical conductivity) also appears to be related to pH and EC and OM are related to each other. EC and OM (both predictor variables in our model) are said to be collinear. This collinearity has a profound effect when both EC and OM are used simultaneously to predict pH.
Order Matters
The two models below are identical, but the summaries of them differ based on sequential sums of squares. Here we see that while EC and OM are significant predictors alone, when together, each becomes less significant. (masking)

```r
> m1 = lm(pH ~ EC + OM, data = soil)
> coef(m1)
(Intercept)          EC          OM
  8.4892205  -0.4124082  -0.1081872

> anova(m1)
Analysis of Variance Table
Response: pH
  Df Sum Sq Mean Sq F value Pr(>F)
EC     1 0.32159 0.32159  24.7189 0.0001162 ***
OM     1 0.06524 0.06524   5.0144 0.0387968 *
Residuals 17 0.22117 0.01301

> drop1(m1, test = 'F')
Single term deletions

Model: pH ~ EC + OM
  Df Sum of Sq RSS AIC F value  Pr(F)
<none>              0.22117 -84.091
EC     1  0.02918 0.25035 -83.613  2.2428 0.1526
OM     1  0.06524 0.28641 -80.921  5.0144 0.0388 *

> m2 = lm(pH ~ OM + EC, data = soil)
> coef(m2)
(Intercept)          OM          EC
  8.4892205  -0.1081872  -0.4124082

> anova(m2)
Analysis of Variance Table
Response: pH
  Df Sum Sq Mean Sq F value Pr(>F)
OM     1 0.35765 0.35765  27.4905 6.61e-05 ***
EC     1 0.02918 0.02918   2.2428 0.1526
Residuals 17 0.22117 0.01301

> drop1(m2, test = 'F')
Single term deletions

Model: pH ~ OM + EC
  Df Sum of Sq RSS AIC F value Pr(F)
<none>              0.22117 -84.091
OM     1  0.06524 0.28641 -80.921  5.0144 0.0388 *
EC     1  0.02918 0.25035 -83.613  2.2428 0.1526
```
> m1 = lm(pH ~ EC, data = soil)
> anova(m1)
Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC</td>
<td>1</td>
<td>0.32159</td>
<td>0.32159</td>
<td>20.211</td>
<td>0.0002796 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>18</td>
<td>0.28641</td>
<td>0.01591</td>
<td></td>
<td>0.01591</td>
</tr>
</tbody>
</table>

---

> m2 = lm(pH ~ OM, data = soil)
> anova(m2)
Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OM</td>
<td>1</td>
<td>0.35765</td>
<td>0.35765</td>
<td>25.715</td>
<td>7.964e-05 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>18</td>
<td>0.25035</td>
<td>0.01391</td>
<td></td>
<td>0.01391</td>
</tr>
</tbody>
</table>

---

SSY = 0.387 + 0.222 = 0.608

SSE = 0.222

SSR = 0.358 + 0.029 = 0.387

SSR = 0.322 + 0.065 = 0.387

SSR = 0.358 + 0.029 = 0.387
Another example (improve performance of predictor)

This example looks similar, but is actually quite different. While WHC (water holding capacity) is not significant by itself, when EC (electrical conductivity) is included in the model, it becomes a significant predictor.

```r
> m1 = lm(pH ~ WHC, data = soil)
> summary(m1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   7.681   0.121   63.39   <2e-16 ***
WHC         0.014   0.009    1.56    0.14    

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.173 on 18 degrees of freedom
Multiple R-squared: 0.118,   Adjusted R-squared: 0.069
F-statistic: 2.42 on 1 and 18 DF,  p-value: 0.137

> m2 = lm(pH ~ WHC + EC, data = soil)
> summary(m2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   8.039   0.099    81.5   <2e-16 ***
WHC          0.016   0.006    2.91   0.01 **  
EC         -0.917   0.166   -5.54 3.62e-05 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.106 on 17 degrees of freedom
Multiple R-squared: 0.686,   Adjusted R-squared: 0.649
F-statistic: 18.53 on 2 and 17 DF,  p-value: 5.36e-05

> anova(m2)

Analysis of Variance Table

Response: pH  
           Df Sum Sq Mean Sq  F value Pr(>F) 
WHC         1 0.0720 0.07203   6.4033  0.0215 *  
EC         1 0.3448 0.34482  30.6618 3.62e-05 ***
Residuals 17 0.1912 0.01125

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```r
> drop1(m2, test = 'F')
Single term deletions

Model:
pH ~ WHC + EC

                     Df Sum of Sq  RSS    AIC F value  Pr( F)
<none>                        0.19118 -87.005
WHC     1   0.09523 0.28641 -80.921  8.4677  0.0097 55 **
EC      1   0.34482 0.53600 -68.387 30.6618 3.619e- 05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```