Chapter 3: General Linear Model

Vineyard Soil Data: Relationship between pH and Organic Matter

Regression as a GLM

> m.soil = lm(pH ~ OM, data = soil)
> summary(m.soil)

Call:
  lm(formula = pH ~ OM, data = soil)

Residuals:
   Min     1Q Median     3Q    Max
-0.19707 -0.08409  0.02841  0.05750  0.28851

Coefficients:  
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.54899    0.13840  61.769  < 2e-16 ***
OM         -0.16346    0.03223  -5.071  7.96e-05 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1179 on 18 degrees of freedom
Multiple R-squared: 0.5882,     Adjusted R-squared:  0.5654
F-statistic: 25.72 on 1 and 18 DF,  p-value: 7.964e-05

So we have the following estimate regression model:

\[ \hat{y} = 8.5490 - 0.1635x \]
Squirrel Length Data: Length differences by location

ANOVA as a GLM
We will now fit an ANOVA using the same model command. (ANOVA and Regression are actually the same type of model!) But first, we need to be very clear with the computer how our factor levels should be treated. Do we want to compare the mean length at each location to the overall mean length (a sum to zero contrast or `contr.sum` in R), or to a reference location (a treatment contrast or `contr.treatment` in R)?

Sum to zero contrast
Using the sum to zero contrast (that our book uses) we have the following model.

```r
> options(contrasts = c('contr.sum', 'contr.sum'))
> m.squirrel = lm(length ~ location, data = squirrel)
> summary(m.squirrel)
```

Call:
`lm(formula = length ~ location, data = squirrel)`

Residuals:
```
 Min 1Q Median 3Q Max
-17.200 -4.563 -0.375 7.600 13.250
```

Coefficients:
```
Estimate Std. Error t value Pr(>|t|)
(Intercept)  263.925      2.240 117.848  < 2e-16 ***
location1     -3.175      4.020  -0.790 0.442825
location2    -11.925      3.733  -3.195 0.006486 **
location3     16.825      4.020   4.185 0.000916 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 9.443 on 14 degrees of freedom
Multiple R-squared: 0.602, Adjusted R-squared: 0.5167
F-statistic: 7.059 on 3 and 14 DF,  p-value: 0.00401

```r
> levels(squirrel$location)
[1] "big bear" "hemet" "loop hill" "susanville"
```

The above (sum to zero contrast) gives the following model(s):

\[
\hat{y} = 263.925 + \begin{cases} 
-3.175 & \text{big bear} \\
-11.925 & \text{hemet} \\
16.825 & \text{loop hill} \\
3.175 + 11.925 - 16.825 & \text{susanville}
\end{cases}
\]

Predict the mean length of a squirrel using the above model.

Predict the mean length of Big Bear squirrels.
**Treatment Contrast**

Using a treatment contrast we have the following model.

```r
> options(contrasts = c('contr.treatment', 'contr.treatment'))
> m.squirrel = lm(length ~ location, data = squirrel)
> summary(m.squirrel)
```

**Call:**
`lm(formula = length ~ location, data = squirrel)`

**Residuals:**
Min      1Q  Median      3Q     Max
-17.200  -4.563  -0.375   7.600  13.250

**Coefficients:**

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 260.750     | 4.721   | 55.228   | < 2e-16 *** |
| locationhemet | -8.750     | 6.334   | -1.381   | 0.18882   |
| locationloop hill | 20.000     | 6.677   | 2.995    | 0.00964 ** |
| locationsusanville | 1.450     | 6.334   | 0.229    | 0.82225   |

---

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

Residual standard error: 9.443 on 14 degrees of freedom
Multiple R-squared: 0.602, Adjusted R-squared: 0.5167
F-statistic: 7.059 on 3 and 14 DF,  p-value: 0.00401

The above gives the following model(s):

\[ \hat{y} = 260.750 + \begin{cases} +0.000 & \text{big bear} \\ -8.750 & \text{hemet} \\ +20.000 & \text{loop hill} \\ +1.450 & \text{susanville} \end{cases} \]

Predict the mean length of a squirrel using the above model.

Predict the mean length of Big Bear squirrels.
Ordered Factors
Recall that levels of a factor, are by default, sorted alphabetically; so, Big Bear is the reference location. It would be more sensible for this problem if the farthest north (or south) were the reference category. The following code is useful for specifying an order to the levels of a factor. In this case, the location column of the squirrel dataset is being characterized as an ordered factor.

```
> squirrel$location = ordered(squirrel$location, 
+     levels = c('hemet', 'big bear', 'susanville', 'loop hill'))
> levels(squirrel$location)
[1] "hemet"      "big bear"   "susanville" "loop hill"
```

Now we fit the model using the exact same commands as above (still with the treatment constrast).

```
> m.squirrel = lm(length ~ location, data = squirrel)
> summary(m.squirrel)

Call:
  lm(formula = length ~ location, data = squirrel)

Residuals:
  Min     1Q Median     3Q    Max
-17.200 -4.563  -0.375   7.600  13.250

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)     252.000      4.223  59.675  < 2e-16 ***
locationbig bear    8.750      6.334   1.381 0.18 8819
locationsusanville 10.200      5.972   1.708 0.10 9714
locationloop hill  28.750      6.334   4.539 0.00 0464 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.443 on 14 degrees of freedom
Multiple R-squared: 0.602, Adjusted R-squared: 0.5167
F-statistic: 7.059 on 3 and 14 DF,  p-value: 0.00401
```

The above gives the following model(s):

\[
\hat{y} = 252.000 + \begin{cases} 
+0.000 & \text{hemet} \\
+8.750 & \text{big bear} \\
+10.200 & \text{susanville} \\
+28.750 & \text{loop hill} 
\end{cases}
\]

Predict the mean length of a squirrel using the above model.

Predict the mean length of Big Bear squirrels.