Chapter 2: Simple Linear Regression

Vineyard Soil Data: Relationship between pH and Organic Matter

Background
Soil pH and percent organic matter was measured at 10 randomly sampled locations in a vineyard in 2004 and again in 2007. Here we will consider the data as simply 20 soil measurements and ignore the fact that the data was sampled on two occasions.

- Is there a significant linear relationship between soil pH and percent organic matter?
- What fraction of the variation in pH can be explained by percent organic matter?
- What is the expected or average change in pH associated with a 1% increase in organic matter?

Look at the Data
The soil object contains the columns pH and OM. Below is a scatterplot of the data.

> with(soil, plot(pH ~ OM, pch = 16, xlab = 'Organic Matter (%)',
+     ylab = 'pH'))
Models

Anova

\[ Y = \mu + \alpha_i + \varepsilon \]
\[ \hat{Y} = \bar{Y} + a_i \]

with

\[ s = \sqrt{\text{MSE}} \]

To test \( H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_g = 0 \) we compute \( F = \frac{\text{MSR}}{\text{MSE}} \) with \( g - 1 \) numerator df and \( n - g \) denominator df.

Regression

\[ Y = \alpha + \beta X + \varepsilon \]
\[ \hat{Y} = \alpha + bX \]

with

\[ s = \sqrt{\text{MSE}} \]

To test \( H_0: \beta = 0 \) we compute an F-statistic \( F = \frac{\text{MSR}}{\text{MSE}} \) with 1 numerator df and \( n - 2 \) denominator df or equivalently, compute a t-statistic \( t = \frac{b - 0}{s_b} \) with \( n - 2 \) df.

What proportion of the variation in \( Y \) can be explained by the regression line (i.e., the linear relationship between \( X \) and \( Y \)?)

\[ R^2 = \frac{SSY - SSE}{SSY} = \frac{SSR}{SSY} \]

Also, the sample correlation between \( X \) and \( Y \) is given by

\[ r = \sqrt{R^2} = \frac{1}{n - 1} \sum \left( \frac{x - \bar{x}}{s_x} \right) \left( \frac{y - \bar{y}}{s_y} \right) \]

---

1Recall from your earlier statistics course that

\[ s_b = \frac{s}{s_x \sqrt{n - 1}} \]

and

\[ b = r \frac{s_y}{s_x} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \]

and

\[ \alpha = \bar{y} - b \bar{x} \]
**Compute Linear Regression Fit**

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

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)       8.548990   0.138400  61.7691   <2e-16 ***
OM                    -0.163467   0.032232  -5.0710   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

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

<table>
<thead>
<tr>
<th>pH</th>
<th>y - \bar{y}</th>
<th>OM (x)</th>
<th>x - \bar{x}</th>
<th>(x - \bar{x})(y - \bar{y})</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2</td>
<td>0.34</td>
<td>3.9</td>
<td>-0.315</td>
<td>-0.1071</td>
</tr>
<tr>
<td>8.1</td>
<td>0.24</td>
<td>3.4</td>
<td>-0.815</td>
<td>-0.1956</td>
</tr>
<tr>
<td>7.8</td>
<td>-0.06</td>
<td>3.4</td>
<td>-0.815</td>
<td>0.0489</td>
</tr>
<tr>
<td>8.0</td>
<td>0.14</td>
<td>3.8</td>
<td>-0.415</td>
<td>-0.0581</td>
</tr>
<tr>
<td>7.9</td>
<td>0.04</td>
<td>3.2</td>
<td>-1.015</td>
<td>-0.0406</td>
</tr>
<tr>
<td>7.9</td>
<td>0.04</td>
<td>4.4</td>
<td>0.185</td>
<td>0.0074</td>
</tr>
<tr>
<td>8.0</td>
<td>0.14</td>
<td>3.6</td>
<td>-0.615</td>
<td>-0.0861</td>
</tr>
<tr>
<td>7.8</td>
<td>-0.06</td>
<td>3.9</td>
<td>-0.315</td>
<td>0.0189</td>
</tr>
<tr>
<td>8.1</td>
<td>0.24</td>
<td>3.0</td>
<td>-1.215</td>
<td>-0.2916</td>
</tr>
<tr>
<td>8.0</td>
<td>0.14</td>
<td>2.9</td>
<td>-1.315</td>
<td>-0.1841</td>
</tr>
<tr>
<td>7.9</td>
<td>0.04</td>
<td>4.3</td>
<td>0.085</td>
<td>0.0034</td>
</tr>
<tr>
<td>7.9</td>
<td>0.04</td>
<td>4.3</td>
<td>0.085</td>
<td>0.0034</td>
</tr>
<tr>
<td>7.8</td>
<td>-0.06</td>
<td>4.4</td>
<td>0.185</td>
<td>-0.0111</td>
</tr>
<tr>
<td>7.8</td>
<td>-0.06</td>
<td>4.6</td>
<td>0.385</td>
<td>-0.0231</td>
</tr>
<tr>
<td>7.8</td>
<td>-0.06</td>
<td>4.9</td>
<td>0.685</td>
<td>-0.0411</td>
</tr>
<tr>
<td>7.5</td>
<td>-0.36</td>
<td>5.6</td>
<td>1.385</td>
<td>-0.4986</td>
</tr>
<tr>
<td>7.6</td>
<td>-0.26</td>
<td>5.8</td>
<td>1.585</td>
<td>-0.4121</td>
</tr>
<tr>
<td>7.8</td>
<td>-0.06</td>
<td>5.0</td>
<td>0.785</td>
<td>-0.0471</td>
</tr>
<tr>
<td>7.7</td>
<td>-0.16</td>
<td>5.3</td>
<td>1.085</td>
<td>-0.1736</td>
</tr>
<tr>
<td>7.6</td>
<td>-0.26</td>
<td>4.6</td>
<td>0.385</td>
<td>-0.1001</td>
</tr>
</tbody>
</table>
Confidence Interval and test for slope

Here’s a table of just the coefficients, their standard errors, etc.

```r
> summary(m1)$coef

                       Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)            8.5489858  0.13840329  61.768660  2.06263 9e-22
OM                     -0.1634605  0.03223435  -5.071002  7.96367 7e-05
```

Computing a $100 \times (1 - \alpha)$CI for $\beta$.

\[
b \pm t_{\alpha/2,n-2}SE_b
\]

```r
> t.mult = qt(.975, 18)
> # 95% confidence interval for beta (the slope)
> summary(m1)$coef[2,1] + c(-1,1) * t.mult * summary(m1)$coef[2,2]
> [1] -0.2311823 -0.0957386
```

Testing $H_0: \beta = -0.3$.

```r
> t.stat = (summary(m1)$coef[2,1] - (-0.3)) / summary(m1)$coef[2,2]
> 2* (1 - pt(t.stat, df = 18))
> [1] 0.0004970428
```

Confidence and Prediction Intervals

- Prediction Interval
  - A certain specimen of soil has an organic matter content of 5.2%. What is its estimated pH?

- Confidence Interval
  - A large area of the vineyard has an organic matter content of 5.2%. What is the average pH for this area?
> new.dat = data.frame(OM = 5.2)
> predict(m1, newdata = new.dat, se.fit = TRUE)
$fit
 1
7.698991

$se.fit
[1] 0.04127382

$df
[1] 18

$residual.scale
[1] 0.1179332

> predict(m1, newdata = new.dat, interval = 'confidence')
 fit  lwr  upr
1 7.698991 7.612278 7.785705
> predict(m1, newdata = new.dat, interval = 'prediction')
 fit  lwr  upr
1 7.698991 7.436487 7.961496

Could we repeat this above questions for soil with an organic matter content of 1.5%? Should we?