H2O2 concentration

OD

pf3d7 and pyoelii

general

parallel

concurrent

coincident
More than one predictor

The model with two parallel lines can be described as

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \]

In other words (ur...equations):

\[ Y = \begin{cases} 
\beta_0 + \beta_1 X_1 + \epsilon & \text{if } X_2 = 0 \\
(\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon & \text{if } X_2 = 1 
\end{cases} \]
Multiple linear regression

A multiple linear regression model has the form

\[ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \]

The predictors (the X's) can be categorical or numerical.

Often, all predictors are numerical or all are categorical.

And actually, categorical variables are converted into a group of numerical ones.

ANOVA as linear regression

ANOVA:
- \( k \) groups; \( n_i \) observations in group \( i \)
- \( y_i \) = response for individual \( i \)
- \( g_i \) = group to which individual \( i \) belongs

Model: \( y \)'s indep't; \( y_i \sim \text{normal}(\mu_{g_i}, \sigma^2) \)

\( H_0: \mu_1 = \mu_2 = \cdots = \mu_k \)

Linear regression:
- Let \( x_{ij} = 1 \) if individual \( i \) is in group \( j \)
  (and = 0 otherwise).

Model: \( y_i = \mu_1 x_{i1} + \mu_2 x_{i2} + \cdots + \mu_k x_{ik} + \epsilon_i \)
where \( \epsilon_i \) iid \( \sim \text{Normal}(0, \sigma^2) \)
You could also write...

\[ y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \]

In which case:

\[ \beta_1 = \mu_1 \quad \beta_j = \mu_j - \mu_1 \text{ for } j > 1 \]

Here \( H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \)

is equivalent to \( H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0 \)

---

**Estimation**

We have the model

\[ y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad \epsilon_i \sim \text{iid Normal}(0, \sigma^2) \]

We estimate the \( \beta \)'s by the values for which

\[ \text{RSS} = \sum_i (y_i - \hat{y}_i)^2 \quad \text{is minimized} \text{ (aka “least squares”)} \]

where \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} \)

We estimate \( \sigma \) by

\[ \hat{\sigma} = \sqrt{\frac{\text{RSS}}{n - (k + 1)}} \]
Calculation of the $\hat{\beta}$’s (and their SEs and correlations) is not that complicated, but without matrix algebra, the formulas are exceedingly nasty.

- The SEs of the $\hat{\beta}$’s involve $\sigma$ and the x’s.
- The $\hat{\beta}$’s are normally distributed.
- Obtain confidence intervals for the $\beta$’s using $\hat{\beta} \pm t \times \hat{SE}(\hat{\beta})$
  where $t = $ quantile of t dist’n with $n-(k+1)$ d.f.
- Test $H_0 : \beta = 0$ using $|\hat{\beta}| / \hat{SE}(\hat{\beta})$
  Compare this to a t dist’n with $n-(k+1)$ d.f.

The example: a full model

$x_1 = [H_2O_2]$.

$x_2 = 0$ or $1$, indicating species of heme.

$y = $ the OD measurement.

The model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$

i.e.,

$$y = \begin{cases} 
\beta_0 + \beta_1 x_1 + \epsilon & \text{if } x_2 = 0 \\
(\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon & \text{if } x_2 = 1 
\end{cases}$$

$\beta_2 = 0 \quad \rightarrow \quad $ Same intercepts.

$\beta_3 = 0 \quad \rightarrow \quad $ Same slopes.

$\beta_2 = \beta_3 = 0 \quad \rightarrow \quad $ Same lines.
Testing many $\beta$'s

We have the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad \epsilon_i \sim \text{iid Normal}(0, \sigma^2)$$

We seek to test

$$H_0: \beta_{r+1} = \cdots = \beta_k = 0.$$

In other words, do we really have just:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir} + \epsilon_i, \quad \epsilon_i \sim \text{iid Normal}(0, \sigma^2)$$
What to do...

1. Fit the “full” model (with all \( k \) x’s).

2. Calculate the residual sum of squares, \( \text{RSS}_{\text{full}} \).

3. Fit the “reduced” model (with only \( r \) x’s).

4. Calculate the residual sum of squares, \( \text{RSS}_{\text{red}} \).

5. Calculate \( F = \frac{(\text{RSS}_{\text{red}} - \text{RSS}_{\text{full}})/(\text{df}_{\text{red}} - \text{df}_{\text{full}})}{\text{RSS}_{\text{full}}/\text{df}_{\text{full}}} \).

   where \( \text{df}_{\text{red}} = n - r - 1 \) and \( \text{df}_{\text{full}} = n - k - 1 \).

6. Under \( H_0 \), \( F \sim F(\text{df}_{\text{red}} - \text{df}_{\text{full}}, \text{df}_{\text{full}}) \).

In particular...

\[
y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad \epsilon_i \sim \text{iid Normal}(0, \sigma^2)
\]

We seek to test

\[
H_0 : \beta_1 = \cdots = \beta_k = 0.
\]

(i.e., none of the x’s are related to y.)

**Full model:** All the x’s

**Reduced model:** \( y = \beta_0 + \epsilon \) (i.e., \( y \sim \text{Normal}(\beta_0, \sigma^2) \))

\[
\text{RSS}_{\text{red}} = \sum_i (y_i - \bar{y})^2
\]

\[
F = \left[ \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2}{k} \right] / \left[ \sum_i (y_i - \hat{y}_i)^2 / (n - k - 1) \right]
\]

and compare to \( F(k, n - k - 1) \) dist’n.
The example

To test $\beta_2 = \beta_3 = 0$ . . .

```r
> lm.red <- lm(y ~ x1, data=dat)
> lm.full <- lm(y ~ x1*x2, data=dat)
> anova(lm.red, lm.full)
```

Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2 + x1:x2

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>0.00975</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>0.00312</td>
<td>2</td>
<td>0.00663</td>
<td>21.22</td>
</tr>
</tbody>
</table>