Multiple Linear Regression

Multiple linear regression

pf3d7 and pyoelii

H2O2 concentration vs. OD
Multiple linear regression

- General
- Parallel
- Concurrent
- Coincident

pf3d7 and pyoelii

H2O2 concentration
**More than one predictor**

<table>
<thead>
<tr>
<th>#</th>
<th>Y</th>
<th>X₁</th>
<th>X₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3399</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.3563</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.3538</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.3168</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.3054</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.3174</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0.2460</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0.2618</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0.2848</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0.1535</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0.1613</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0.1525</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0.3332</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0.3414</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0.3299</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>0.2940</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>0.2948</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>0.2903</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>0.2089</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>0.2189</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>0.2102</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>0.1006</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>0.1031</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>0.1452</td>
<td>50</td>
<td>1</td>
</tr>
</tbody>
</table>

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 (or, 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.
**Interpretation**

Let $X_1$ be the age of a subject (in years).

$$E[Y] = \beta_0 + \beta_1 X_1$$

$\rightarrow$ Comparing two subjects who differ by one year in age, we expect the responses to differ by $\beta_1$.

$\rightarrow$ Comparing two subjects who differ by five years in age, we expect the responses to differ by $5\beta_1$.

---

**Interpretation**

Let $X_1$ be the age of a subject (in years), and let $X_2$ be an indicator for the treatment arm (0/1).

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$\rightarrow$ Comparing two subjects from the same treatment arm who differ by one year in age, we expect the responses to differ by $\beta_1$.

$\rightarrow$ Comparing two subjects of the same age from the two different treatment arms ($X_2=1$ versus $X_2=0$), we expect the responses to differ by $\beta_2$. 
Interpretation

Let $X_1$ be the age of a subject (in years), and let $X_2$ be an indicator for the treatment arm (0/1).

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

$\longrightarrow$  $E[Y] = \beta_0 + \beta_1 X_1$ (if $X_2=0$)

$\longrightarrow$  $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = \beta_0 + \beta_2 + (\beta_1 + \beta_3) X_1$ (if $X_2=1$)

$\longrightarrow$  Comparing two subjects who differ by one year in age, we expect the responses to differ by $\beta_1$ if they are in the control arm ($X_2=0$), and expect the responses to differ by $\beta_1 + \beta_3$ if they are in the treatment arm ($X_2=1$).

Estimation

We have the model

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

$\longrightarrow$  We estimate the $\beta$’s by the values for which

$$\text{RSS} = \sum_i (y_i - \hat{y}_i)^2$$

is minimized where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$ (aka “least squares”).

$\longrightarrow$  We estimate $\sigma$ by

$$\hat{\sigma} = \sqrt{\frac{\text{RSS}}{n - (k + 1)}}$$
FYI

Calculation of the $\hat{\beta}$’s (and their SEs and correlations) is not that complicated, but without matrix algebra, the formulas are nasty.

Here is what you need to know:

- 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 $\hat{\beta} \pm t \times \hat{\text{SE}}(\hat{\beta})$
  where $t$ is a quantile of t dist’n with $n-(k+1)$ d.f.
- Test $H_0: \beta = 0$ using $|\hat{\beta}|/\hat{\text{SE}}(\hat{\beta})$
  Compare this to a t distribution with $n-(k+1)$ d.f.

→ Use the R function lm()!

The example: a full model

$x_1 = [\text{H}_2\text{O}_2]$.

$x_2 = 0$ 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.,

$$y = \begin{cases} 
\beta_0 + \beta_1 x_1 + \epsilon & \text{if } X_2 = 0 \\
(\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon & \text{if } X_2 = 1 
\end{cases}$$

$\beta_2 = 0 \implies$ Same intercepts.

$\beta_3 = 0 \implies$ Same slopes.

$\beta_2 = \beta_3 = 0 \implies$ Same lines.
Results

```r
> lm.out <- lm(y ~ x1 * x2, data=mydat)
> summary(lm.out)

Coefficients:  
                         Estimate Std. Error t value Pr(>|t|)  
(Intercept)            0.35305   0.00544  64.90  < 2e-16  
x1                    -0.00387   0.00019  -20.20    8.86e-15  
x2                    -0.01992   0.00769   -2.60    0.0175  
x1:x2                 -0.00055   0.00027   -2.02    0.0563  

Residual standard error: 0.0125 on 20 degrees of freedom  
Multiple R-Squared: 0.98, Adjusted R-squared: 0.977  
F-statistic: 326.4 on 3 and 20 DF, p-value: < 2.2e-16
```

Testing many parameters

We have the model

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

We seek to test \( H_0 : \beta_{r+1} = \cdots = \beta_k = 0 \).

In other words, do we really have just:

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

?
What to do...

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

2. Calculate the residual sum of squares, $RSS_{full}$.

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

4. Calculate the residual sum of squares, $RSS_{red}$.

5. Calculate $F = \frac{(RSS_{red} - RSS_{full})/(df_{red} - df_{full})}{RSS_{full}/df_{full}}$.
   
   where $df_{red} = n - r - 1$ and $df_{full} = n - k - 1$.

6. Under $H_0$, $F \sim F(df_{red} - df_{full}, df_{full})$.

In particular...

Assume the model

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

We seek to test $H_0 : \beta_1 = \cdots = \beta_k = 0$ (i.e., none of the x’s are related to y).

$\rightarrow$ Full model: All the x’s

$\rightarrow$ Reduced model: $y = \beta_0 + \epsilon \quad RSS_{red} = \sum_i (y_i - \hat{y})^2$

$\rightarrow$ $F = \frac{(\sum_i (y_i - \hat{y})^2 - \sum_i (y_i - \hat{y}_i)^2)/k}{\sum_i (y_i - \hat{y}_i)^2/(n - k)}$

Compare this to a $F(k, n - k - 1)$ dist’n.
The example

To test $\beta_2 = \beta_3 = 0$

> 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>0.00975</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.00312</td>
<td>2</td>
<td>0.00663</td>
<td>21.22</td>
<td>1.1e-05</td>
</tr>
</tbody>
</table>

Summary

- $R^2$ is called the coefficient of determination: it is equal to the proportion of the variability in $Y$ explained by the regression model.

- The sample (multiple) correlation coefficient in a regression setting can be defined as the correlation between the observed values $Y$ and the fitted values $\hat{Y}$ from the regression model. Mathematically, we have $R = \text{cor}(Y, \hat{Y})$

- $R^2$ tells us nothing about model violations.
Summary

• The notion “the higher $R^2$, the better the model” is simply wrong.

• Assuming we have an intercept in the (linear regression) model, the more predictors we include in the model, the higher $R^2$.

• There is a test for “significant” reductions in $R^2$.

• In a linear model, over-fitting does not cause bias, but (slightly) inflates the standard error.

• Under-fitting on the other hand can cause bias.

• Randomization controls for bias due to unfitted covariates.

Diagnostics

Assumptions

- $\epsilon$’s normally distributed
- $\epsilon$’s have constant SD
- $y$’s linear in each of the $x$’s
- No other $x$’s belong in the model

Diagnostics

- QQ plot of residuals
- Plot residuals vs fitted values
- Plot residuals vs each $x$
- Plot residuals vs other $x$’s
Another example

Sediment ingestion by the mud snail, *Hydrobia minuta*.

\[ y = \text{Amount ingested} \]
\[ x = \text{Time allowed to eat} \]

Let's consider the model

\[ y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{iid Normal}(0, \sigma^2) \]
## Estimated coefficients

<table>
<thead>
<tr>
<th></th>
<th>Est</th>
<th>SE</th>
<th>t-val</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-339</td>
<td>127</td>
<td>-2.66</td>
<td>0.019</td>
</tr>
<tr>
<td>time</td>
<td>75.7</td>
<td>15.4</td>
<td>4.91</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>time^2</td>
<td>-1.55</td>
<td>0.48</td>
<td>-3.22</td>
<td>0.006</td>
</tr>
<tr>
<td>time^3</td>
<td>0.010</td>
<td>0.004</td>
<td>2.52</td>
<td>0.024</td>
</tr>
</tbody>
</table>

## Diagnostic plots

- Residuals vs. Fitted
- QQ plot of residuals
- Residuals vs. time