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

\[ \text{RSS} = \sum_i (y_i - \hat{y}_i)^2 \text{ is minimized (aka “least squares”) \quad where } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} \]

We estimate \( \sigma \) by

\[ \hat{\sigma} = \sqrt{\frac{\text{RSS}}{n - (k + 1)}} \]
Trust me . . .

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 $t = \text{quantile of t dist'n with n–(k+1) d.f.}$
- 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.

$y = \text{the OD measurement}$.

The model: $y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \epsilon$

i.e.,

$$y = \begin{cases} 
\beta_0 + \beta_1X_1 + \epsilon & \text{if } X_2 = 0 \\
(\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1 + \epsilon & \text{if } X_2 = 1 
\end{cases}$$

$\beta_2 = 0 \quad \longrightarrow \quad \text{Same intercepts.}$

$\beta_3 = 0 \quad \longrightarrow \quad \text{Same slopes.}$

$\beta_2 = \beta_3 = 0 \quad \longrightarrow \quad \text{Same lines.}$
> lm.out <- lm(y ~ x1 * x2, data=mydat)
> summary(lm.out)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.35305 0.00544 64.9 < 2e-16
x1 -0.00387 0.00019 -20.2 8.86e-15
x2 -0.01992 0.00769 -2.6 0.0175
x1:x2 -0.00055 0.00027 -2.0 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 $\beta$’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., \( y \sim \text{Normal}(\beta_0, \sigma^2) \))

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

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

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

<table>
<thead>
<tr>
<th></th>
<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>Model 1: y ~ x1</td>
<td>22</td>
<td>0.00975</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2: y ~ x1 + x2 + x1:x2</td>
<td>20</td>
<td>0.00312</td>
<td>2</td>
<td>0.00663</td>
<td>21.22</td>
<td>1.1e-05</td>
</tr>
</tbody>
</table>

Example


$y_t =$ incremental growth rate in year $t$

$x_{1t} =$ population density in year $t$

$x_{2t} =$ rise in temperature in year $t$

$x_{3t} =$ a measure of river flow in year $t$

Fit the model

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \epsilon_t$$

where $\epsilon_t \sim \text{iid Normal}(0, \sigma^2)$
Results

<table>
<thead>
<tr>
<th></th>
<th>Est</th>
<th>SE</th>
<th>t-value</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>17.00</td>
<td>3.22</td>
<td>5.3</td>
<td>0.001</td>
</tr>
<tr>
<td>density</td>
<td>−0.37</td>
<td>0.11</td>
<td>−3.2</td>
<td>0.015</td>
</tr>
<tr>
<td>temp.rise</td>
<td>25.36</td>
<td>9.01</td>
<td>2.8</td>
<td>0.026</td>
</tr>
<tr>
<td>flow</td>
<td>0.45</td>
<td>0.47</td>
<td>1.0</td>
<td>0.362</td>
</tr>
</tbody>
</table>
## Assumptions

- $\epsilon$’s independent
- $\epsilon$’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

- Plot residuals vs time
- QQ plot of residuals
- Plot residuals vs fitted values
- Plot residuals vs each x
- Plot residuals vs other x’s

### residuals vs time

![Residuals vs Time Plot](image-url)
residual(t) vs residual(t+1)

\[ \text{corr} = -0.09 \]

Permutation test

\[ P\text{-value} = 78\% \]
QQ plot of residuals

Residuals vs fitted values
A fake example

<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>-4.29</td>
<td>4.97</td>
<td>-0.86</td>
<td>0.40</td>
</tr>
<tr>
<td>x1</td>
<td>0.58</td>
<td>0.25</td>
<td>2.36</td>
<td>0.03</td>
</tr>
</tbody>
</table>
Residuals vs another $x$

Regress $y$ on both $x$’s

<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>18.02</td>
<td>8.50</td>
<td>2.12</td>
<td>0.0491</td>
</tr>
<tr>
<td>$x_1$</td>
<td>1.21</td>
<td>0.29</td>
<td>4.13</td>
<td>0.0007</td>
</tr>
<tr>
<td>$x_2$</td>
<td>-0.77</td>
<td>0.26</td>
<td>-3.00</td>
<td>0.0080</td>
</tr>
</tbody>
</table>
Sediment ingestion by the mud snail, *Hyrobia minutaa*.

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

A model

Let's consider the model

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