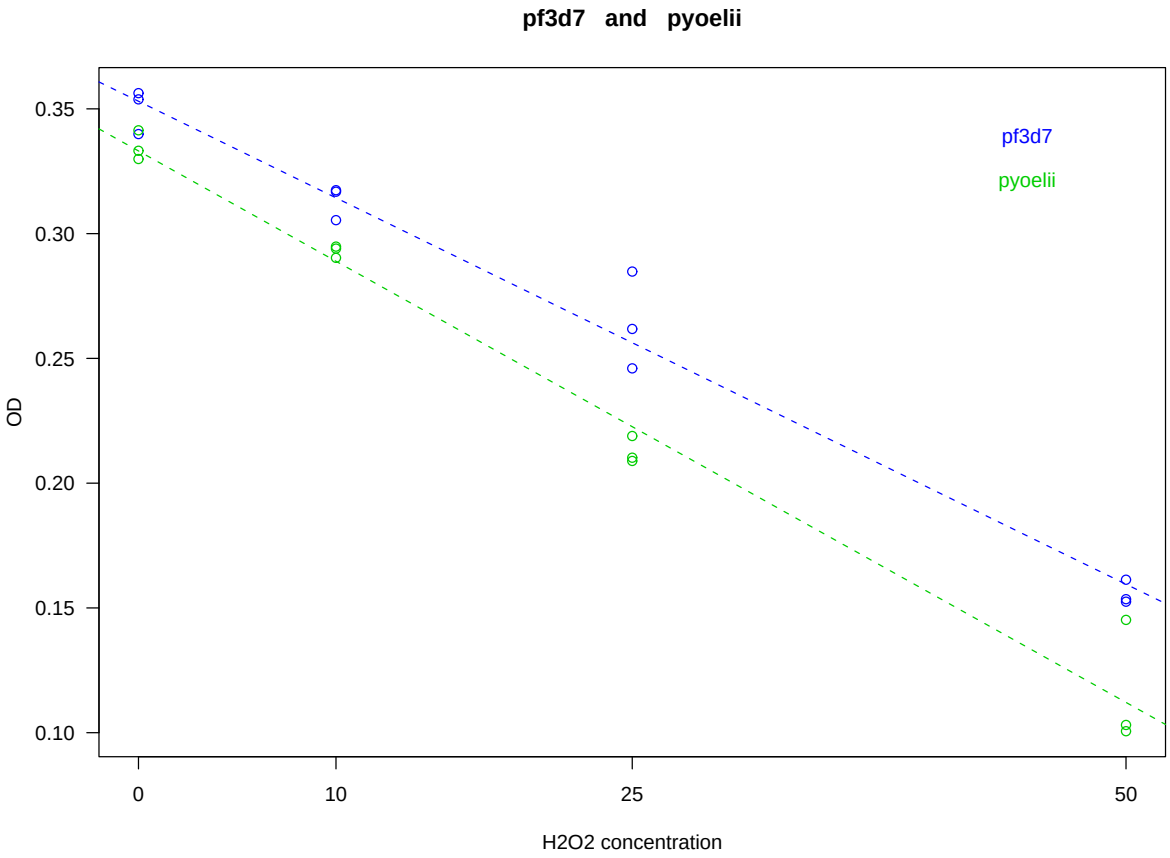
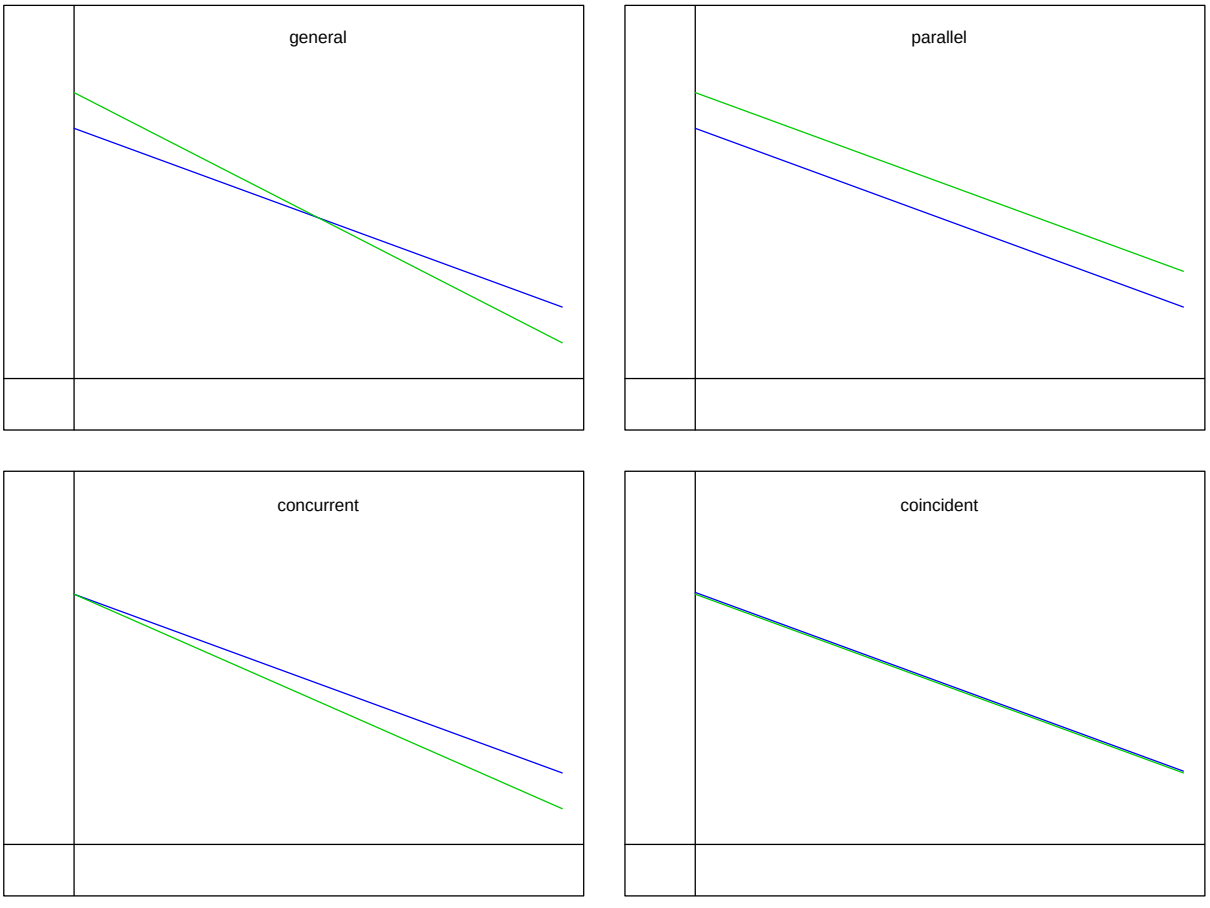
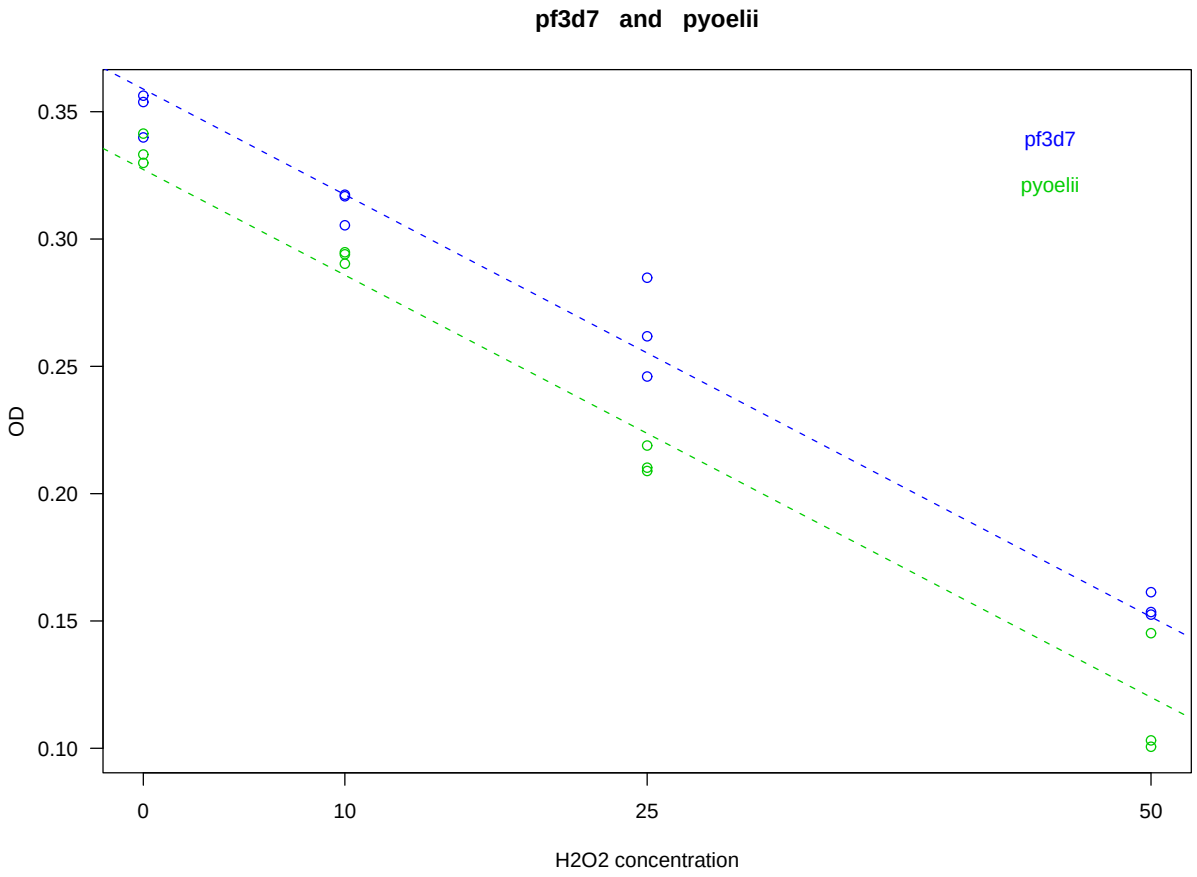

pf3d7 and pyoelii
pf3d7
pyoelii
OD
0.35
0.30
0.25
0.20
0.15
0.10
0
10
25
50
H2O2 concentration


general
parallel
concurrent
coincident



More than one predictor

#	Y	X_1	X_2
1	0.3399	0	0
2	0.3563	0	0
3	0.3538	0	0
4	0.3168	10	0
5	0.3054	10	0
6	0.3174	10	0
7	0.2460	25	0
8	0.2618	25	0
9	0.2848	25	0
10	0.1535	50	0
11	0.1613	50	0
12	0.1525	50	0
13	0.3332	0	1
14	0.3414	0	1
15	0.3299	0	1
16	0.2940	10	1
17	0.2948	10	1
18	0.2903	10	1
19	0.2089	25	1
20	0.2189	25	1
21	0.2102	25	1
22	0.1006	50	1
23	0.1031	50	1
24	0.1452	50	1

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

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

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 ϵ_i iid $\sim$ Normal(0, σ^2)

You could also write...

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

In which case:

$\beta_1 = \mu_1$ $\beta_j = \mu_j - \mu_1$ 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 β's by the values for which

$RSS = \sum_i (y_i - \hat{y}_i)^2$ is minimized (aka "least squares")

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

We estimate σ by $\hat{\sigma} = \sqrt{\frac{RSS}{n - (k+1)}}$

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 σ and the x's.
- The $\hat{\beta}$'s are normally distributed.
- Obtain confidence intervals for the β's using $\hat{\beta} \pm \mathrm{t} \times \widehat{\mathrm{SE}}(\hat{\beta})$

  where t = quantile of t dist'n with n–(k+1) d.f.
- Test $H_0 : \beta = 0$ using $|\hat{\beta}|/\widehat{\mathrm{SE}}(\hat{\beta})$

  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.,

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

$\beta_2 = 0 \longrightarrow$ Same intercepts.

$\beta_3 = 0 \longrightarrow$ Same slopes.

$\beta_2 = \beta_3 = 0 \longrightarrow$ Same lines.

Results

```
> 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 β'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, 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. . .

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

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., $y \sim \text{Normal}(\beta_0, \sigma^2)$

$RSS_{red} = \sum_i (y_i - \bar{y})^2$

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

and compare to F(k, n – k – 1) dist'n.

The example

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

```
> 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

        Res.Df       RSS   Df Sum of Sq       F    Pr(>F)
    1       22   0.00975
    2       20   0.00312    2   0.00663   21.22   1.1e-05
```