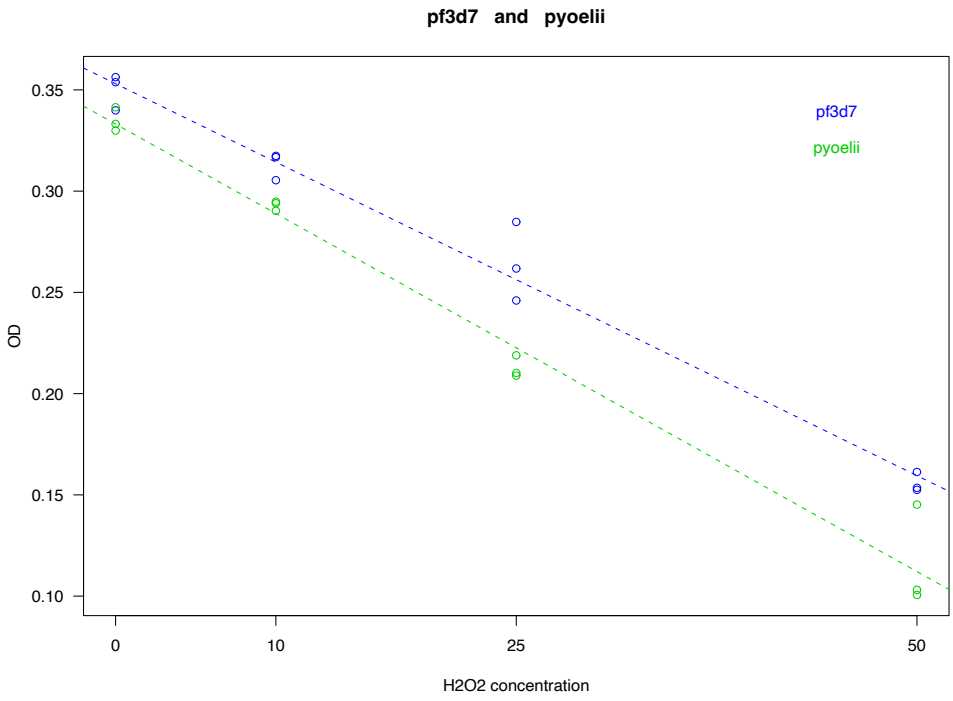
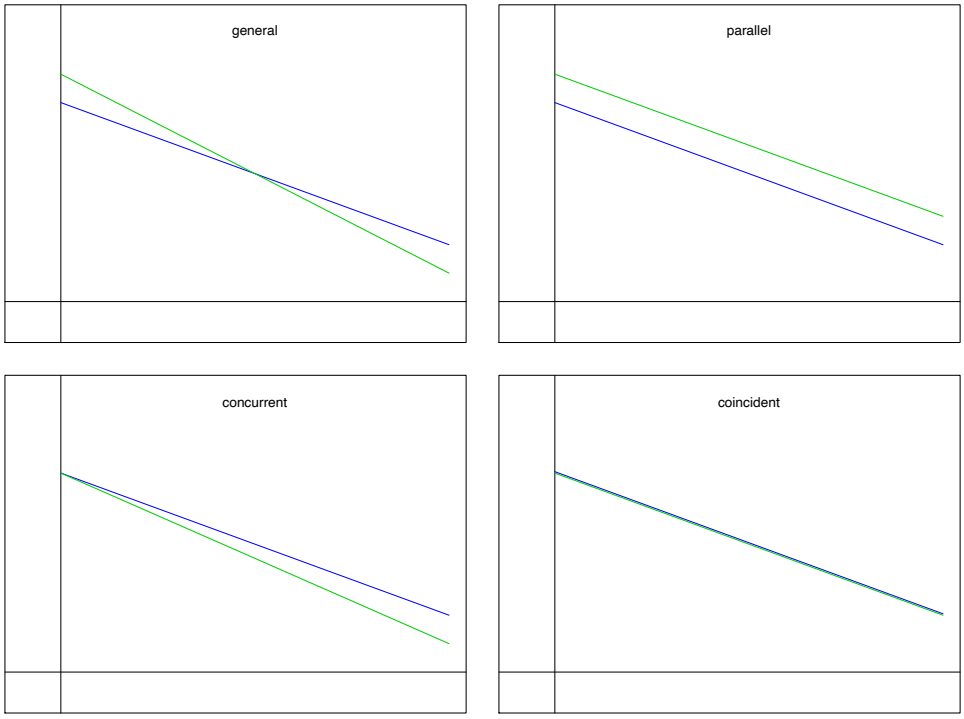
Multiple Linear Regression

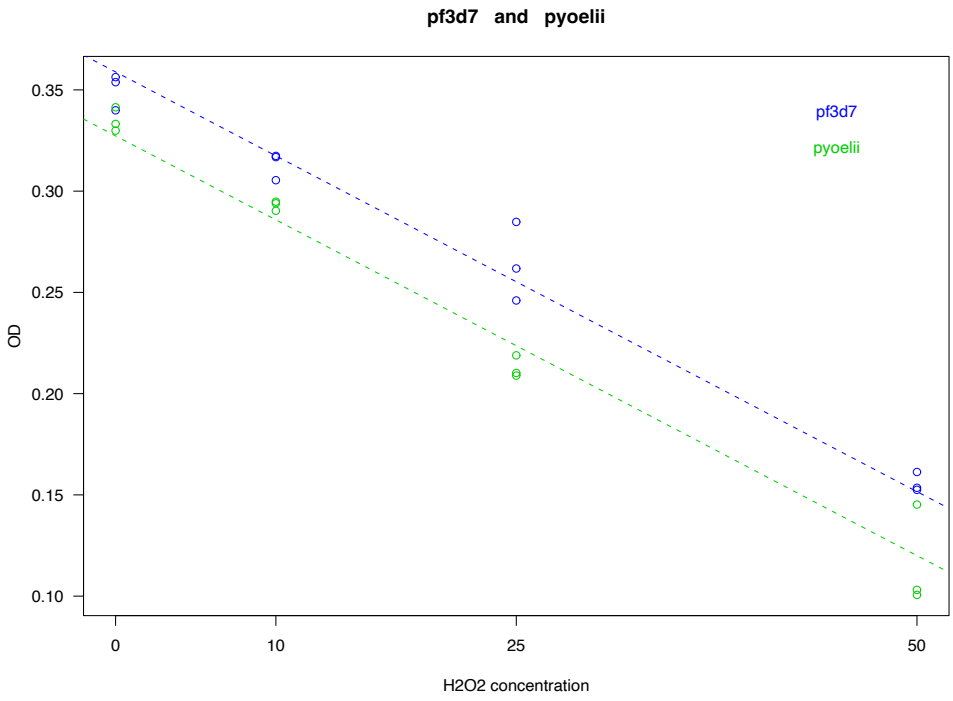
Multiple linear regression



Multiple linear regression



Multiple linear regression



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 (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

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

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.

Interpretation

Let X_1 be the concentration of H2O2.

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

$\longrightarrow$ Comparing two experiments that differ by one unit concentration, we expect the responses to differ by β_1.

$\longrightarrow$ Comparing two experiments that differ by five units concentration, we expect the responses to differ by $5\beta_1$.

Interpretation

Let X_1 be the concentration of H2O2 and let X_2 be the indicator for the species of heme (0/1).

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

$\longrightarrow$ Comparing two experiments on the same species of heme that differ by one unit concentration, we expect the responses to differ by β_1.

$\longrightarrow$ Comparing two experiments at the same concentration on the two different species of heme (X_2=1 versus X_2=0), we expect the responses to differ by β_2.

Interpretation

Let X_1 be the concentration of H2O2 and let X_2 be the indicator for the species of heme (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 experiments that differ by one unit concentration, we expect the responses to differ by β_1 if they are in the first heme (X_2=0), and expect the responses to differ by $\beta_1 + \beta_3$ if they are in the second heme (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 β's by the values for which

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

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

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

$\longrightarrow$ Use the R function lm()!

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.,

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

β_2=0 $\longrightarrow$ Same intercepts.
β_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 parameters

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...

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

$\longrightarrow$ Full model: All the x's

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

$\longrightarrow$ $F = [(\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2)/k] \; / \; [\sum_i (y_i - \hat{y}_i)^2/(n - k - 1)]$
Compare this to a F(k, n – k – 1) dist'n.

The example

To test β_2=β_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
```

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 = 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

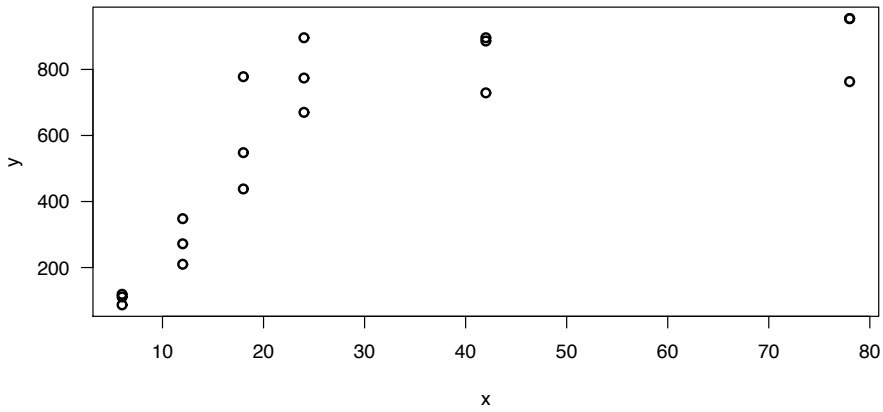
Assumptions	Diagnostics
ϵ's normally distributed	QQ plot of residuals
ϵ's have constant SD	Plot residuals vs fitted values
y's linear in each of the x's	Plot residuals vs each x
No other x's belong in the model	Plot residuals vs other x's

Another example

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

y = Amount ingested

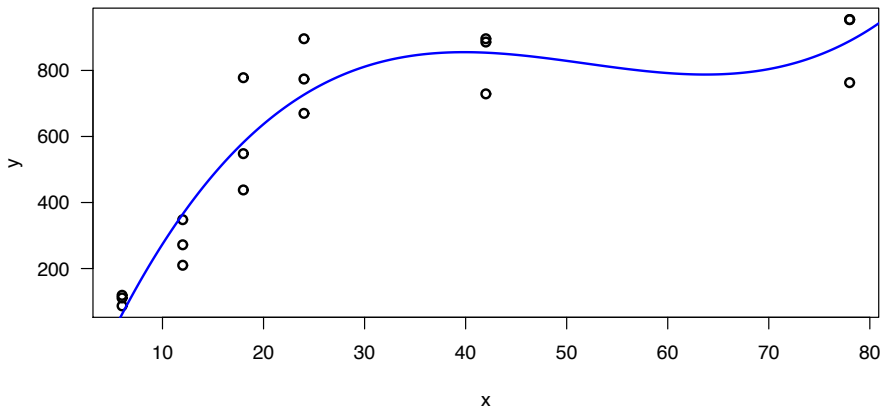
x = Time allowed to eat



A model

Let's consider the model

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



Estimated coefficients

	Est	SE	t-val	P-val
Intercept	-339	127	-2.66	0.019
time	75.7	15.4	4.91	<0.001
time^2	-1.55	0.48	-3.22	0.006
time^3	0.010	0.004	2.52	0.024

Diagnostic plots