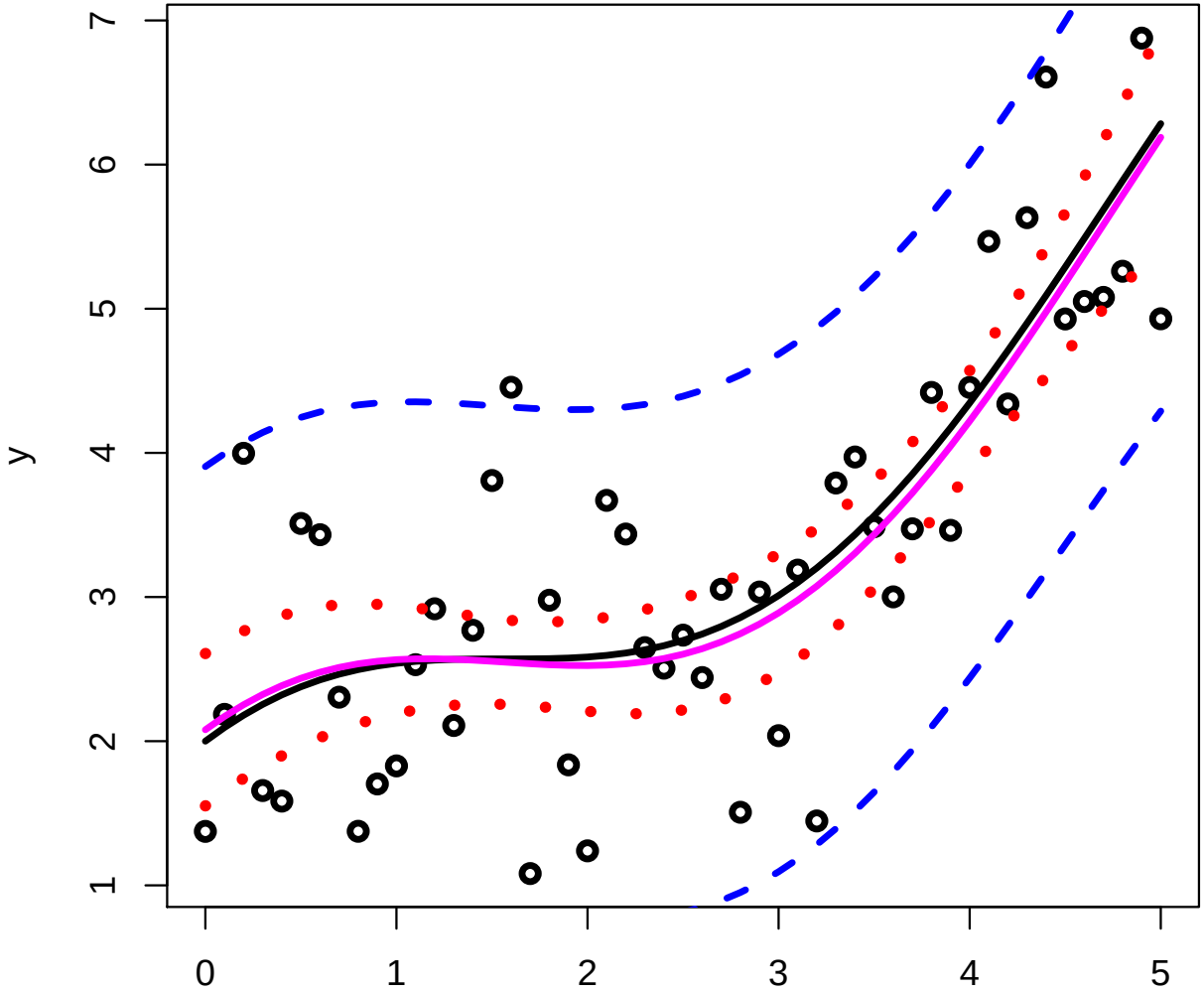
Linear Models with R: A demonstration.

Arthur Berg
aberg@ucsd.edu
http://www.math.ucsd.edu/∼aberg/



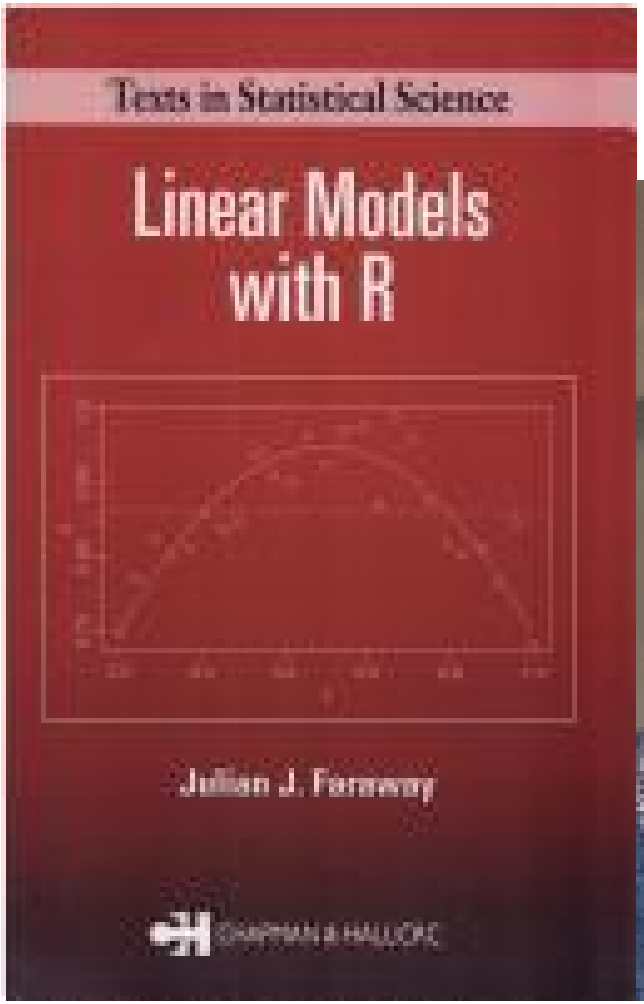
Outline

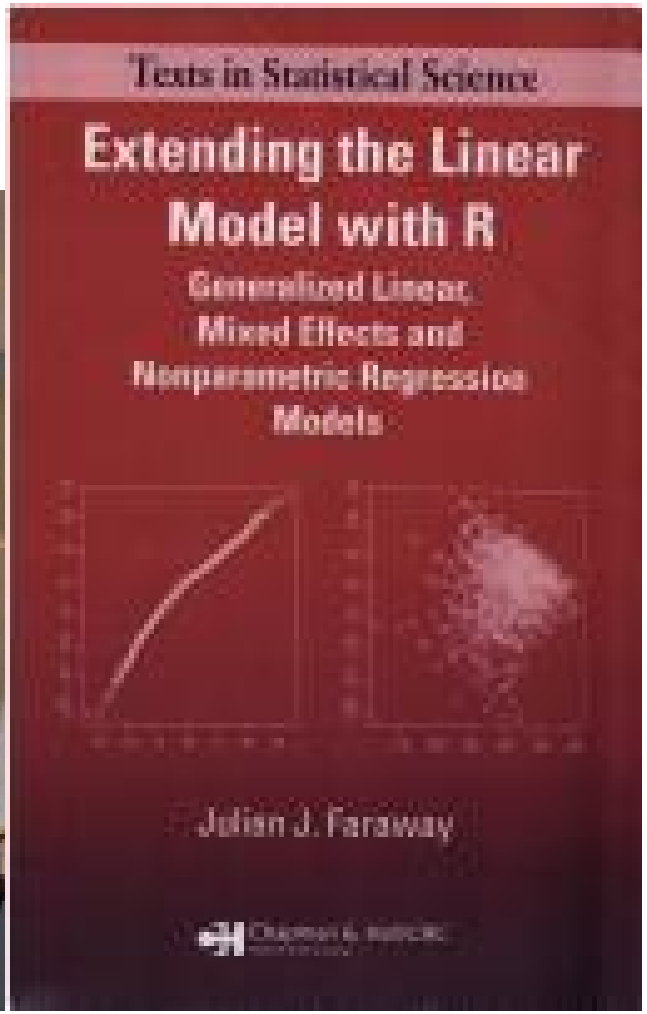
1. Introduction
2. `R` basics
3. `lm`
4. Model Selection
5. ANOVA

R and these slides

- R is free and very powerful.
- Slides created with LaTeX-Beamer, R, and Sweave
- Interested in Computational Statistics? Check out Math 185
- Other courses:
  - ECE 271AB (Statistical Learning)
  - CSE 190 (Stats intro course)
- R books:
  - Venables and Ripley–Modern Applied Statistics with S.
  - Venables and Smith–An Introduction to R.
  - Verzani–Simple R and Using R for introductory statistics.

R books for Linear Models





- We will follow Professor Julian J. Faraway's free text Practical Regression and ANOVA using R (213 pages) in the R basics and ANOVA sections.

Importing data

- R has many buit-in datasets – `data()`
- Many other datasets in `Library(MASS)` (V & R)
- Use functions `read.table()` and `scan()`.

```
> library(MASS)
> library(faraway)
> data(stackloss)
> dim(stackloss)
```

```
[1] 21  4
```

```
> stackloss[1:5, ]
```

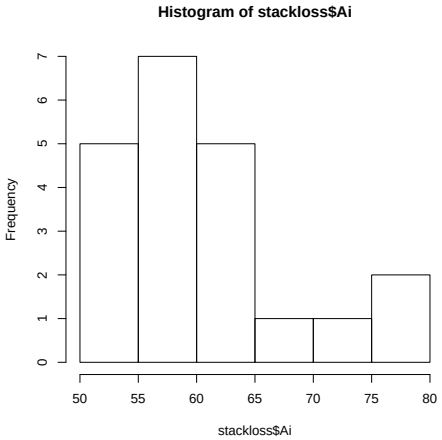
```
  Air.Flow Water.Temp Acid.Conc. stack.loss
1       80         27         89         42
2       80         27         88         37
3       75         25         90         37
4       62         24         87         28
5       62         22         87         18
```

Numerical Summaries

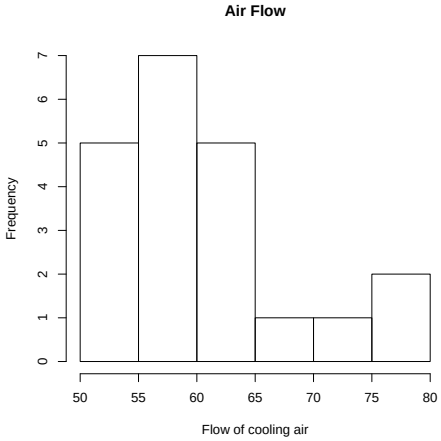
```
> summary(stackloss)
    Air.Flow       Water.Temp     Acid.Conc.      stack.loss
 Min.   :50.0   Min.   :17.0   Min.   :72.0   Min.   : 7.0
 1st Qu.:56.0   1st Qu.:18.0   1st Qu.:82.0   1st Qu.:11.0
 Median :58.0   Median :20.0   Median :87.0   Median :15.0
 Mean   :60.4   Mean   :21.1   Mean   :86.3   Mean   :17.5
 3rd Qu.:62.0   3rd Qu.:24.0   3rd Qu.:89.0   3rd Qu.:19.0
 Max.   :80.0   Max.   :27.0   Max.   :93.0   Max.   :42.0
> stackloss$Air.Flow
 [1] 80 80 75 62 62 62 62 62 58 58 58 58 58 58 50 50 50 50 50 56 70
> mean(stackloss$Ai)
[1] 60.43
> median(stackloss$Ai)
[1] 58
> range(stackloss$Ai)
[1] 50 80
> quantile(stackloss$Ai)
  0%  25%  50%  75% 100%
  50   56   58   62   80
> var(stackloss$Ai)
[1] 84.06
> sd(stackloss$Ai)
[1] 9.168
> cor(stackloss)
           Air.Flow Water.Temp Acid.Conc. stack.loss
Air.Flow     1.0000     0.7819     0.5001     0.9197
Water.Temp   0.7819     1.0000     0.3909     0.8755
Acid.Conc.   0.5001     0.3909     1.0000     0.3998
stack.loss   0.9197     0.8755     0.3998     1.0000
```

Graphical Summaries I

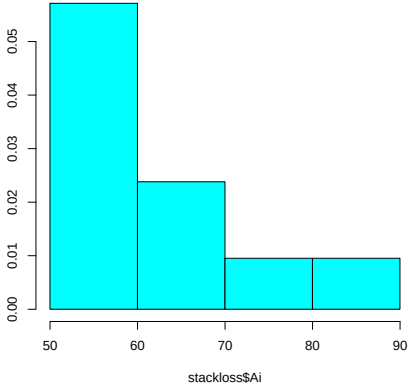
```
> hist(stackloss$Ai)
```



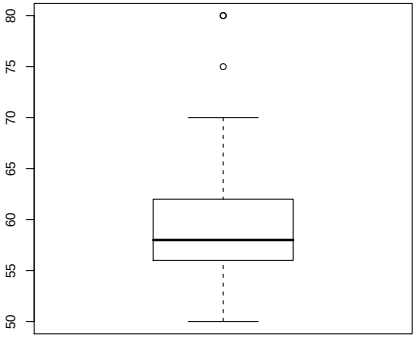

```
> hist(stackloss$Ai, main = "Air Flow",
+      xlab = "Flow of cooling air")
```




```
> truehist(stackloss$Ai)
```

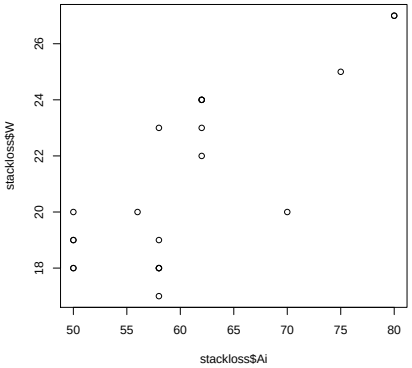



```
> boxplot(stackloss$Ai)
```

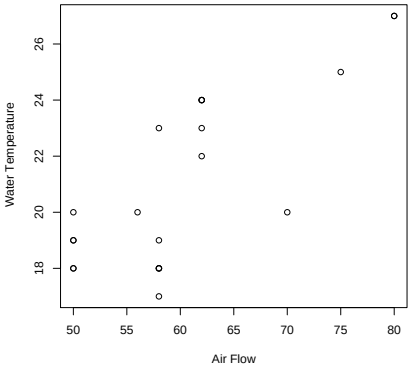


Graphical Summaries II

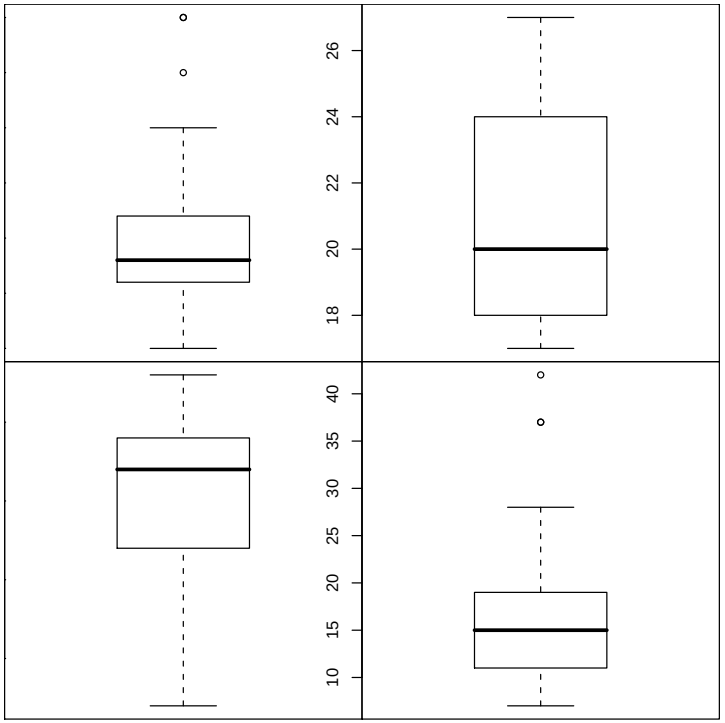
```
> plot(stackloss$Ai, stackloss$W)
```




```
> plot(Water.Temp ~ Air.Flow,
+      stackloss, xlab = "Air Flow",
+      ylab = "Water Temperature")
```

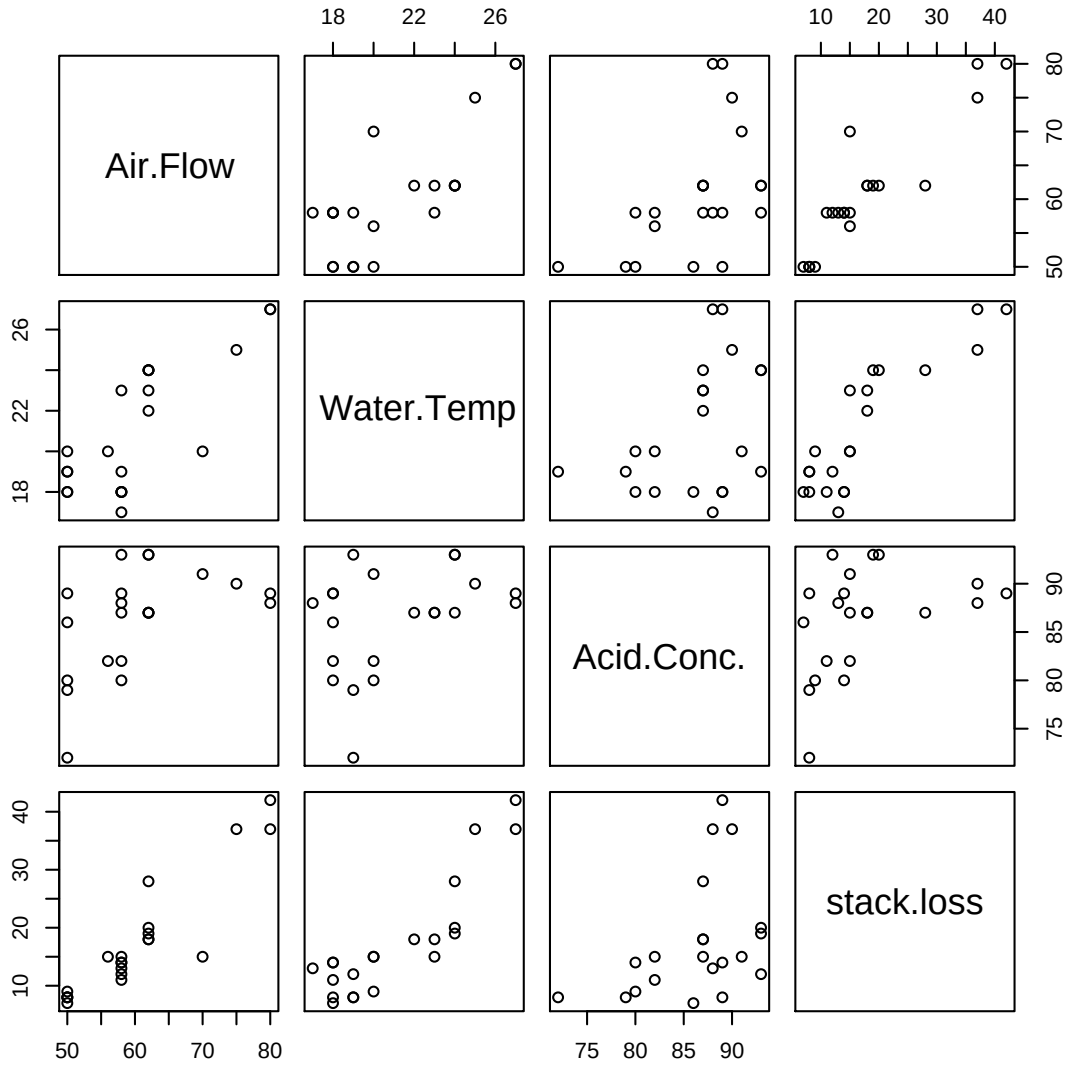



```
> par(mfrow = c(2, 2))
> par(mar = c(0, 0, 0, 0))
> boxplot(stackloss$Ai)
> boxplot(stackloss$Wa)
> boxplot(stackloss$Ac)
> boxplot(stackloss$s)
> par(mfrow = c(1, 1))
```



Graphical Summaries III

```
> plot(stackloss)
```



Selecting Subsets

```
> stackloss[2, ]

  Air.Flow Water.Temp Acid.Conc. stack.loss
2       80         27         88         37

> stackloss[, 3]

 [1] 89 88 90 87 87 87 93 93 87 80 89 88 82 93 89 86 72 79 80 82 91

> stackloss[2, 3]

[1] 88

> c(1, 2, 4)

[1] 1 2 4

> stackloss[c(1, 2, 4), ]

  Air.Flow Water.Temp Acid.Conc. stack.loss
1       80         27         89         42
2       80         27         88         37
4       62         24         87         28

> 3:7

[1] 3 4 5 6 7

> stackloss[3:6, ]

  Air.Flow Water.Temp Acid.Conc. stack.loss
3       75         25         90         37
4       62         24         87         28
5       62         22         87         18
6       62         23         87         18

> stackloss[-(2:21), ]

  Air.Flow Water.Temp Acid.Conc. stack.loss
1       80         27         89         42

> stackloss[stackloss$Ai > 72, ]

  Air.Flow Water.Temp Acid.Conc. stack.loss
1       80         27         89         42
2       80         27         88         37
3       75         25         90         37
```

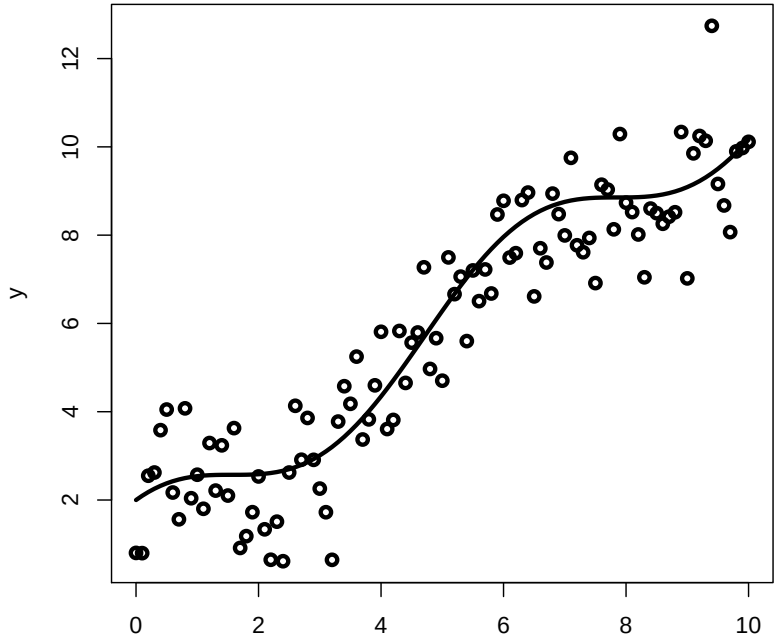
A Simple Simulation

Simulate data from the model

$$y_i = 1 + x_i + \cos(x_i) + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

where $x_i = i/10$ for $i = 1, \ldots, 100$.

```
> x <- seq(0, 10, 0.1)
> y <- 1 + x + cos(x) + rnorm(x)
> yy <- 1 + x + cos(x)
> plot(x, y, lwd = 3)
> lines(x, yy, lwd = 3)
```



lm

Fit the linear model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \cos(x) + \varepsilon_i$$

```
> fit <- lm(y ~ x + I(cos(x)))
> fit

Call:
lm(formula = y ~ x + I(cos(x)))

Coefficients:
(Intercept)            x    I(cos(x))
      0.854        0.989        0.923
```

Notes:

- Intercept is automatically included; use `lm(y~0+x+I(x^2)+I(cos(x)))` or `lm(y~-1+x+I(x^2)+I(cos(x)))` to exclude intercept.
- The function `I()` is specific to `R`; `S` users can simply type `lm(y~x+x^2+cos(x))`.

Summary of lm

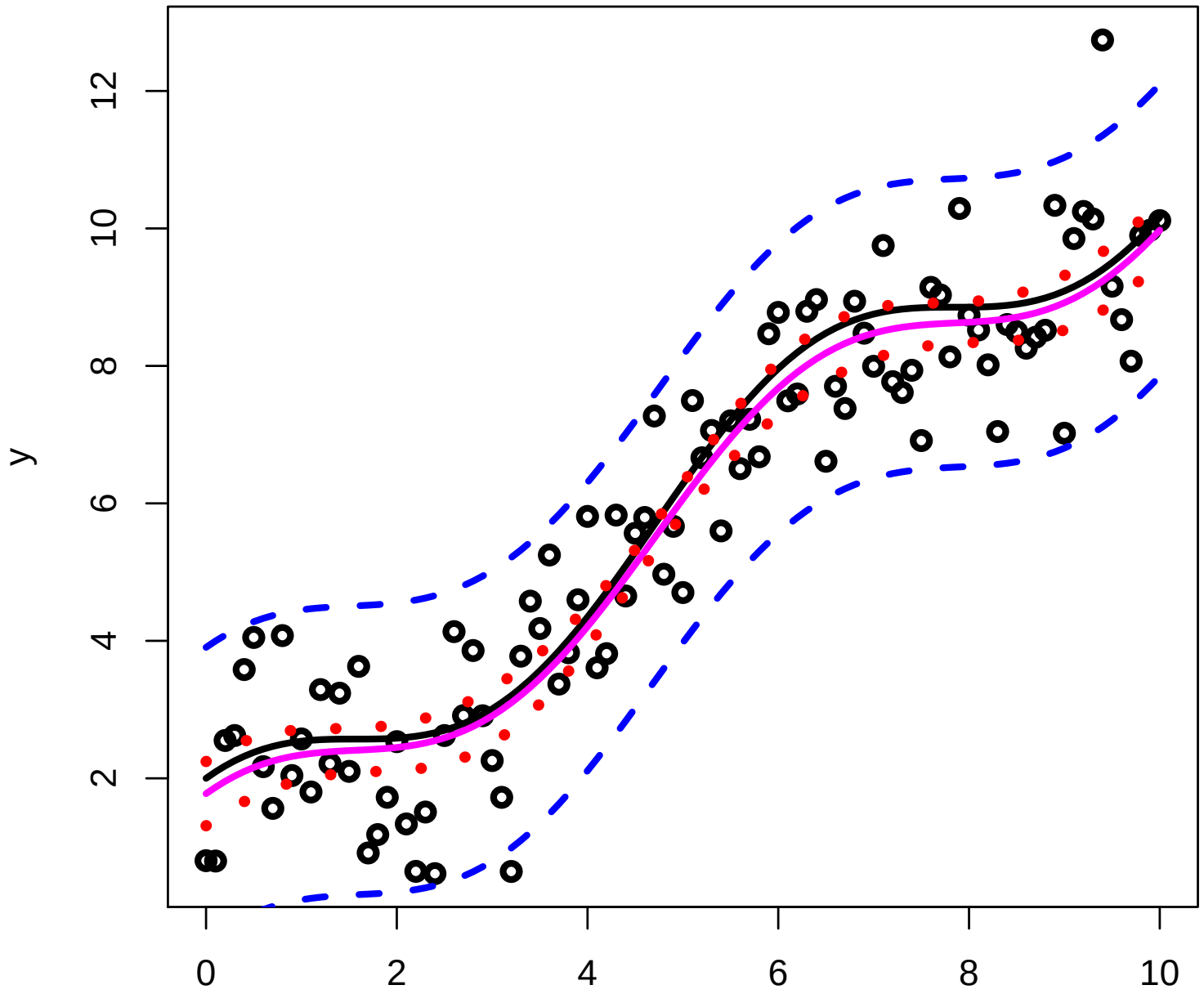
```
> summary(fit)

Call:
lm(formula = y ~ x + I(cos(x)))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4537 -0.7110 -0.0566  0.7269  3.5102 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.8538     0.2097    4.07  9.5e-05 ***
x             0.9895     0.0368   26.92  < 2e-16 ***
I(cos(x))     0.9229     0.1481    6.23  1.2e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.05 on 98 degrees of freedom
Multiple R-Squared: 0.881, Adjusted R-squared: 0.878 
F-statistic:  362 on 2 and 98 DF,  p-value: <2e-16
```

Confidence Intervals

```
> confint(fit)

              2.5 % 97.5 %
(Intercept) 0.4376  1.270
x           0.9165  1.062
I(cos(x))   0.6289  1.217

> pr1 <- predict(fit, i = "c")
> pr2 <- predict(fit, i = "p")
> pr1[1:3, ]

    fit   lwr   upr
1 1.777 1.309 2.244
2 1.871 1.409 2.333
3 1.956 1.501 2.412

> pr2[1:3, ]

    fit     lwr   upr
1 1.777 -0.3519 3.905
2 1.871 -0.2564 3.998
3 1.956 -0.1698 4.082
```

```
> yhat <- pr1[, 1]
> yhL1 <- pr1[, 2]
> yhU1 <- pr1[, 3]
> yhL2 <- pr2[, 2]
> yhU2 <- pr2[, 3]
> lines(x, yhat, lwd = 3,
+     lty = 1, col = "magenta")
> lines(x, yhL1, lw = 5,
+     lt = 3, co = "red")
> lines(x, yhU1, lw = 5,
+     lt = 3, co = "red")
> lines(x, yhL2, lw = 3,
+     lt = 2, co = "blue")
> lines(x, yhU2, lw = 3,
+     lt = 2, co = "blue")
```

Graphic



Stepwise AIC

```
> f <- lm(y ~ x + I(x^2) + I(x^3) + sin(x) + cos(x) + exp(x))
> fAIC <- step(f, . ~ .)

         Df Sum of Sq   RSS   AIC
- I(x^2)  1       1.2 102.7  13.7
- I(x^3)  1       1.5 103.0  14.0
<none>                101.5  14.5
- exp(x)  1       2.2 103.7  14.6
- x       1       2.9 104.4  15.4
- sin(x)  1       5.1 106.6  17.5
- cos(x)  1      10.4 112.0  22.4
         Df Sum of Sq   RSS   AIC
- exp(x)  1       1.3 104.0  13.0
<none>                102.7  13.7
- I(x^3)  1       2.2 104.9  13.8
+ I(x^2)  1       1.2 101.5  14.5
- sin(x)  1       4.1 106.8  15.6
- x       1      18.0 120.7  28.0
- cos(x)  1      31.3 134.0  38.5
         Df Sum of Sq   RSS   AIC
- I(x^3)  1       1.5 105.5  12.4
<none>                104.0  13.0
+ exp(x)  1       1.3 102.7  13.7
- sin(x)  1       3.0 107.0  13.8
+ I(x^2)  1       0.3 103.7  14.6
- cos(x)  1      44.1 148.1  46.7
- x       1      91.5 195.5  74.7
         Df Sum of Sq RSS AIC
- sin(x)  1         2 107  12
<none>                105  12
+ I(x^2)  1         2 104  13
+ I(x^3)  1         1 104  13
+ exp(x)  1         1 105  14
- cos(x)  1        43 149  45
- x       1       786 891 226
         Df Sum of Sq RSS AIC
<none>                107  12
+ sin(x)  1         2 105  12
+ exp(x)  1         1 107  14
+ I(x^3)  1      0.36 107  14
+ I(x^2)  1      0.33 107  14
- cos(x)  1        43 150  44
- x       1       794 901 225
```

Stepwise BIC

```
> fBIC <- step(f, . ~ ., k = log(length(x)))

         Df Sum of Sq   RSS   AIC
- I(x^2)  1       1.2 102.7  29.4
- I(x^3)  1       1.5 103.0  29.7
- exp(x)  1       2.2 103.7  30.3
- x       1       2.9 104.4  31.1
<none>                101.5  32.8
- sin(x)  1       5.1 106.6  33.2
- cos(x)  1      10.4 112.0  38.1
         Df Sum of Sq   RSS   AIC
- exp(x)  1       1.3 104.0  26.0
- I(x^3)  1       2.2 104.9  26.9
- sin(x)  1       4.1 106.8  28.7
<none>                102.7  29.4
+ I(x^2)  1       1.2 101.5  32.8
- x       1      18.0 120.7  41.1
- cos(x)  1      31.3 134.0  51.6
         Df Sum of Sq   RSS   AIC
- I(x^3)  1       1.5 105.5  22.8
- sin(x)  1       3.0 107.0  24.3
<none>                104.0  26.0
+ exp(x)  1       1.3 102.7  29.4
+ I(x^2)  1       0.3 103.7  30.3
- cos(x)  1      44.1 148.1  57.1
- x       1      91.5 195.5  85.2
         Df Sum of Sq RSS AIC
- sin(x)  1         2 107  20
<none>                105  23
+ I(x^2)  1         2 104  26
+ I(x^3)  1         1 104  26
+ exp(x)  1         1 105  27
- cos(x)  1        43 149  53
- x       1       786 891 234
         Df Sum of Sq RSS AIC
<none>                107  20
+ sin(x)  1         2 105  23
+ exp(x)  1         1 107  24
+ I(x^3)  1      0.36 107  24
+ I(x^2)  1      0.33 107  24
- cos(x)  1        43 150  49
- x       1       794 901 230
```

AIC, BIC Summaries

```
> summary(fAIC)

Call:
lm(formula = y ~ x + cos(x))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4537 -0.7110 -0.0566  0.7269  3.5102 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.8538     0.2097    4.07  9.5e-05 ***
x             0.9895     0.0368   26.92  < 2e-16 ***
cos(x)        0.9229     0.1481    6.23  1.2e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.05 on 98 degrees of freedom
Multiple R-Squared: 0.881, Adjusted R-squared: 0.878 
F-statistic:  362 on 2 and 98 DF,  p-value: <2e-16

> summary(fBIC)

Call:
lm(formula = y ~ x + cos(x))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4537 -0.7110 -0.0566  0.7269  3.5102 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.8538     0.2097    4.07  9.5e-05 ***
x             0.9895     0.0368   26.92  < 2e-16 ***
cos(x)        0.9229     0.1481    6.23  1.2e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.05 on 98 degrees of freedom
Multiple R-Squared: 0.881, Adjusted R-squared: 0.878 
F-statistic:  362 on 2 and 98 DF,  p-value: <2e-16
```

leaps

```
> ?leaps
leaps(x=, y=, wt=rep(1, NROW(x)), int=TRUE,
method=c("Cp", "adjr2", "r2"), nbest=10, names=NULL,
df=NROW(x), strictly.compatible=TRUE)
```

```
> library(leaps)
> X <- cbind(x, x^2, x^3, cos(x), sin(x), exp(x))
> fCp <- leaps(X, y)
> str(fCp)
```

```
List of 4
 $ which: logi [1:43, 1:6]  TRUE FALSE FALSE FALSE FALSE FALSE ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:43] "1" "1" "1" "1" ...
  .. ..$ : chr [1:6] "1" "2" "3" "4" ...
 $ label: chr [1:7] "(Intercept)" "1" "2" "3" ...
 $ size : num [1:43] 2 2 2 2 2 2 3 3 3 3 ...
 $ Cp   : num [1:43]  41.7  88.1 173.8 453.1 728.1 ...
```

```
> fCp$which[1:3, ]
```

```
      1     2     3     4     5     6
1  TRUE FALSE FALSE FALSE FALSE FALSE
1 FALSE  TRUE FALSE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE FALSE FALSE
```

```
> fCp$Cp
```

```
 [1]  41.745  88.097 173.832 453.126 728.094 737.501   4.381  21.483
 [9]  40.259  42.596  42.720  43.719  44.727  46.225  75.546  91.551
[17]   4.659   5.895   6.050   6.071   9.313  19.934  20.884  22.423
[25]  24.629  29.218   5.194   5.309   5.716   6.132   7.889   7.893
[33]   8.024   8.065   8.605  13.128   6.127   6.425   7.008   7.715
[41]   9.739  14.670   7.000
```

```
> mCp <- which.min(fCp$Cp)
> fCp$which[mCp, ]
```

```
    1     2     3     4     5     6
 TRUE FALSE FALSE  TRUE FALSE FALSE
```

Coagulation Data

- Blood coagulation times under 4 different diets.

```
> data(coagulation)
> coagulation
```

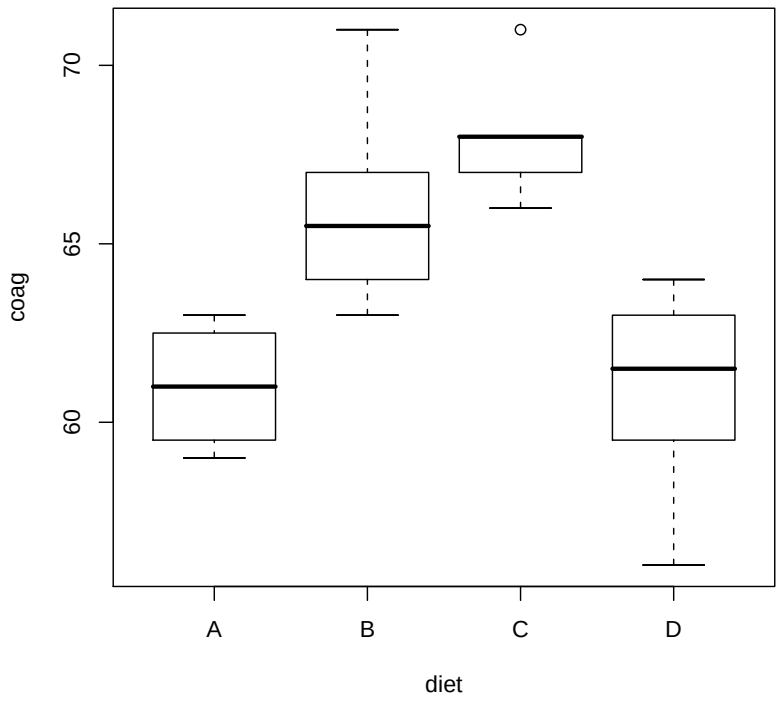
```
   coag diet
1    62    A
2    60    A
3    63    A
4    59    A
5    63    B
6    67    B
7    71    B
8    64    B
9    65    B
10   66    B
11   68    C
12   66    C
13   71    C
14   67    C
15   68    C
16   68    C
17   56    D
18   62    D
19   60    D
20   61    D
21   63    D
22   64    D
23   63    D
24   59    D
```

```
> summary(coagulation)
```

```
      coag      diet
 Min.   :56.0   A:4
 1st Qu.:61.8   B:6
 Median :63.5   C:6
 Mean   :64.0   D:8
 3rd Qu.:67.0
 Max.   :71.0
```

Boxplots

```
> plot(coag ~ diet, data = coagulation)
```



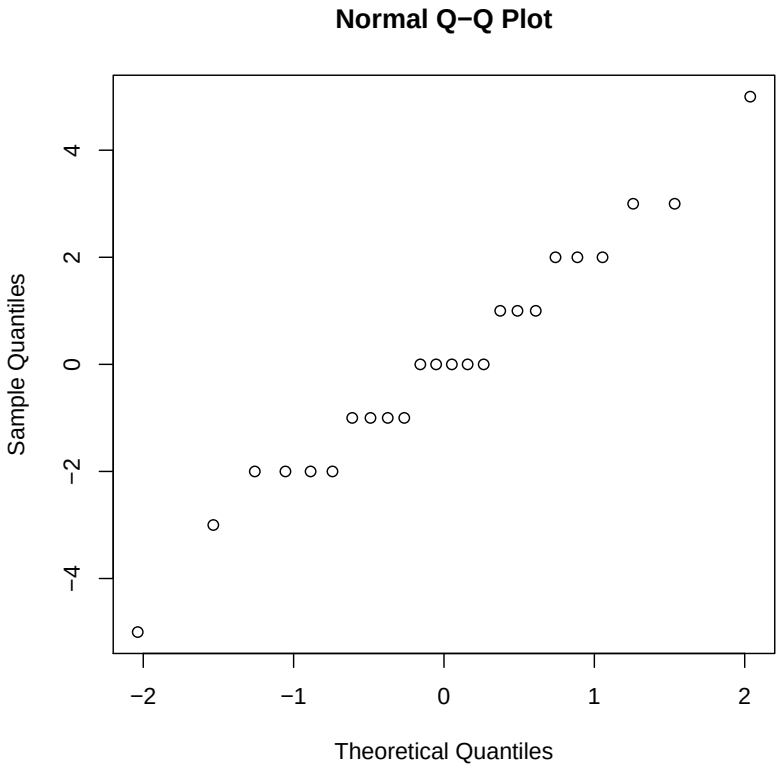
- Outliers?
- Skewness?
- Unequal variance?

Fit a model

```
> g <- lm(coag ~ diet, coagulation)
> summary(g)

Call:
lm(formula = coag ~ diet, data = coagulation)

Residuals:
      Min        1Q    Median        3Q       Max
-5.00e+00 -1.25e+00  1.49e-16  1.25e+00  5.00e+00

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  6.10e+01   1.18e+00    51.55  < 2e-16 ***
dietB        5.00e+00   1.53e+00     3.27  0.00380 **
dietC        7.00e+00   1.53e+00     4.58  0.00018 ***
dietD       -1.07e-14   1.45e+00 -7.4e-15  1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.37 on 20 degrees of freedom
Multiple R-Squared: 0.671, Adjusted R-squared: 0.621
F-statistic: 13.6 on 3 and 20 DF,  p-value: 4.66e-05

> gi <- lm(coag ~ diet - 1, coagulation)
> summary(gi)

Call:
lm(formula = coag ~ diet - 1, data = coagulation)

Residuals:
      Min        1Q    Median        3Q       Max
-5.00e+00 -1.25e+00  1.74e-16  1.25e+00  5.00e+00

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
dietA   61.000      1.183    51.5   <2e-16 ***
dietB   66.000      0.966    68.3   <2e-16 ***
dietC   68.000      0.966    70.4   <2e-16 ***
dietD   61.000      0.837    72.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.37 on 20 degrees of freedom
Multiple R-Squared: 0.999, Adjusted R-squared: 0.999
F-statistic: 4.4e+03 on 4 and 20 DF,  p-value: <2e-16
```

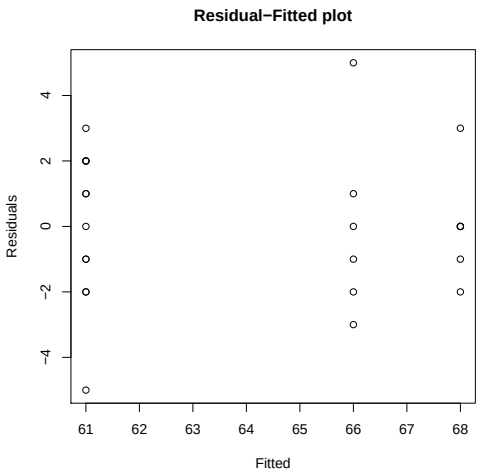
Diagnostics I

```
> qqnorm(g$res)
```

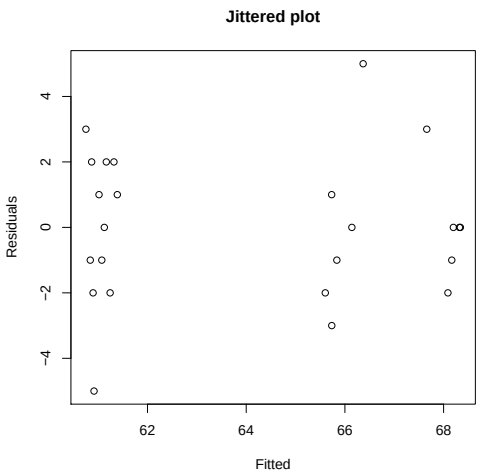


Diagnostics

```
> plot(g$fit, g$res, xlab = "Fitted", ylab = "Residuals",
+     main = "Residual-Fitted plot")
```




```
> plot(jitter(g$fit), g$res, xlab = "Fitted", ylab = "Residuals",
+     main = "Jittered plot")
```



Levene's test of homogeneity

```
> summary(lm(abs(g$res) ~ coagulation$diet))

Call:
lm(formula = abs(g$res) ~ coagulation$diet)

Residuals:
      Min        1Q    Median        3Q       Max
-2.00e+00 -1.00e+00 -1.36e-16  6.25e-01  3.00e+00

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.500      0.716    2.10    0.049 *
coagulation$dietB    0.500      0.924    0.54    0.594
coagulation$dietC   -0.500      0.924   -0.54    0.594
coagulation$dietD    0.500      0.877    0.57    0.575
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.43 on 20 degrees of freedom
Multiple R-Squared: 0.0956, Adjusted R-squared: -0.0401
F-statistic: 0.705 on 3 and 20 DF,  p-value: 0.56
```

aov

```
> summary.aov(g)

            Df Sum Sq Mean Sq F value  Pr(>F)
diet         3  228.0    76.0    13.6 4.7e-05 ***
Residuals   20  112.0     5.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey Honest Significant Difference (HSD)

```
> TukeyHSD(aov(coag ~ diet, coagulation))

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = coag ~ diet, data = coagulation)

$diet
          diff       lwr    upr  p adj
B-A  5.000e+00    0.7246  9.275 0.0183
C-A  7.000e+00    2.7246 11.275 0.0010
D-A -1.421e-14   -4.0560  4.056 1.0000
C-B  2.000e+00   -1.8241  5.824 0.4766
D-B -5.000e+00   -8.5771 -1.423 0.0044
D-C -7.000e+00  -10.5771 -3.423 0.0001
```