610 Handout R1-Regression, Basic with missing data regression

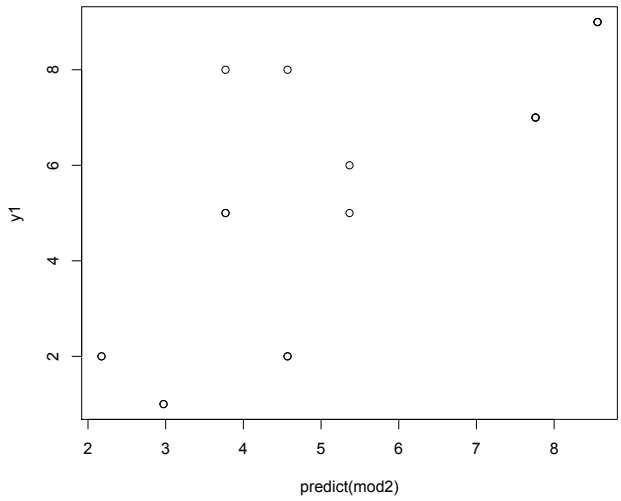
Missing data creates headaches for us R novices. The key to getting predicted and residuals in R that can be plotted is to run the model with the option 'na.action = na.exclude'.

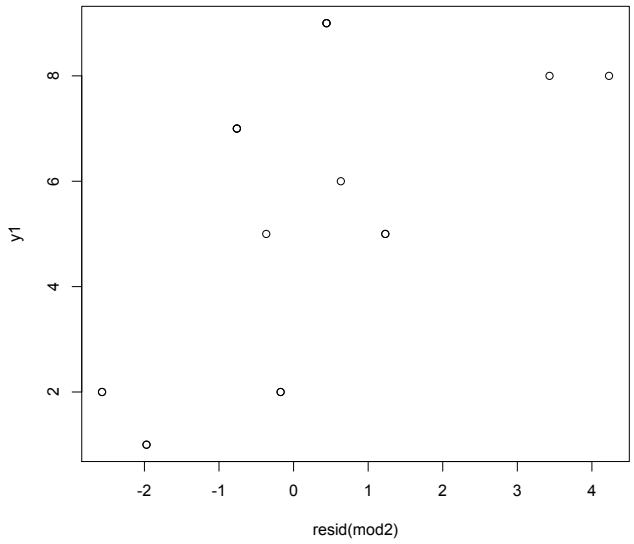
Quick overview:

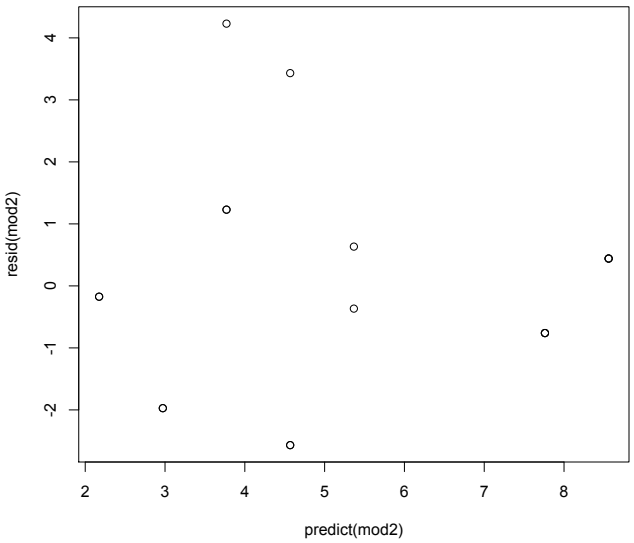
1. Make a data set, then remove a couple of observations, replace with NaN to indicate missing.
2. Input data to R. I pasted in my Mac.
3. Use 'lm' for regression with na.action = na.exclude. This allows you to plot predicted and residuals etc. Otherwise, R doesn't handle the missing data in the way you probably want.
4. Plot predictions vs observed values.
5. Plot residuals vs predicted values

```
> toydata=read.table(pipe("pbpaste"),header=T)
> toydata
   x1  y1
1   1   2
2   4   2
3   5   6
4   8   7
5   9   9
6   3   5
7   2   1
8   5 NaN
9   4   8
10  1   2
11  4   2
12 NaN   6
13  8   7
14  9   9
15  3   5
16  2   1
17  5   5
18  3   8
19  8   7
20  9   9
> attach(toydata)  ## attaching has disadvantages in certain circumstances
> mod2=lm(y1~x1, na.action=na.exclude) # na.exclude pads the vectors so they are equal length
> mod2

Call:
lm(formula = y1 ~ x1, na.action = na.exclude)

Coefficients:
```

```
(Intercept)         x1
   1.3760      0.7981
> summary(mod2)

Call:
lm(formula = y1 ~ x1, na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5684 -0.7607 -0.1741  0.5855  4.2297

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3760     0.8951   1.537 0.143754
x1            0.7981     0.1591   5.018 0.000126 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.88 on 16 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.6114,  Adjusted R-squared:  0.5871
F-statistic: 25.18 on 1 and 16 DF,  p-value: 0.0001263

> predict(mod2)  ## list the predicted values to show the missing values
       1        2        3        4        5        6        7        8
9
2.174086 4.568362 5.366455 7.760731 8.558824 3.770270 2.972178       NA
4.568362
      10       11       12       13       14       15       16       17
18
2.174086 4.568362       NA 7.760731 8.558824 3.770270 2.972178 5.366455
3.770270
      19       20
7.760731 8.558824

> resid(mod2)  ## list residuals
        1         2         3         4         5         6         7         8
-0.1740859 -2.5683625  0.6335453 -0.7607313  0.4411765  1.2297297 -1.9721781        NA
        9        10        11        12        13        14        15        16
 3.4316375 -0.1740859 -2.5683625        NA -0.7607313  0.4411765  1.2297297 -1.9721781
       17        18        19        20
-0.3664547  4.2297297 -0.7607313  0.4411765


> plot(predict(mod2),y1) ## predicted vs observed
> plot(resid(mod2),y1)  ## residual vs observed
> plot(predict(mod2),resid(mod2))## predicted vs residual
```


y1
predict(mod2)
2
3
4
5
6
7
8


y1
resid(mod2)
-2
-1
0
1
2
3
4


resid(mod2)
predict(mod2)
-2
-1
0
1
2
3
4
2
3
4
5
6
7
8