Regression Analysis and Lack of Fit

We will look at an example of regression and AOV in R. For more resources on using R, please refer to links under the Computing links on the course website. To start R at the command line under UNIX,

```
okeeffe> R

R : Copyright 2002, The R Development Core Team
Version 1.6.1  (2002-11-01)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.
```

We will examine data from 27 coral reef heads, *Porites lobata*, in the Great Barrier Reef. Risk and Sammarco (1991) found that the density of the coral skeletons increases with distance from the Australian shore, due to differences in inshore and offshore environments. Read in the data from the web site and summarize:

```
> coral <- read.table("http://www.stat.duke.edu/courses/Spring04/sta244/Data/rwg.179.1",header=T)
> summary(coral)
     Sample                   Reef      Distance        Density
 Min.   : 1.0   AlmaBay         :3   Min.   : 3.50   Min.   :1.053
 1st Qu.: 7.5   BowdenReef      :3   1st Qu.:15.40   1st Qu.:1.272
 Median :14.0   GreatPalmIs.    :3   Median :27.80   Median :1.375
 Mean   :14.0   GrubReef        :3   Mean   :33.16   Mean   :1.337
 3rd Qu.:20.5   LittleBroadhurst:3   3rd Qu.:49.50   3rd Qu.:1.435
 Max.   :27.0   MiddleReef      :3   Max.   :74.50   Max.   :1.589
                (Other)         :9
```

The variable `Reef` gives the location of the reef, and is a categorical variable; summary provides counts for each area. The last two variables are the distance and density. It is usually a good idea to look at a plot of all the variables:

```
> postscript("coral-pair-plot.ps")
> pairs(coral)
> dev.off()
```

the postscript and dev.off commands save the graph to a postscript file; omit these if you want to view the plot on the computer screen

RS summarized the relation with a second order polynomial,

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \epsilon_i$$

To fit this model in R we can use the `lm()` function.

```
> coral.lm <- lm(Density ~ Distance + Distance^2, data=coral)
```

The first argument is the *model formula*; the second specifies the dataframe. In the model formula we specify that the response, Density, has a mean with predictors Distance and Distance squared. The intercept is included by default; to omit it add a `-1` to the right-hand side of the model formula. The `I()` function around `Distance^2` protects the function, as usual arithmetic operators have different meanings inside model formula.

```
> summary(coral.lm)

Call:
lm(formula = Density ~ Distance + I(Distance^2), data = coral)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20988 -0.03427  0.01100  0.04247  0.14731

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.167e+00  5.556e-02  20.995   <2e-16 ***
Distance       7.380e-03  3.678e-03   2.006   0.0562 .
I(Distance^2) -4.482e-05  4.447e-05  -1.008   0.3237
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0981 on 24 degrees of freedom
Multiple R-Squared: 0.4935,     Adjusted R-squared: 0.4513
F-statistic: 11.69 on 2 and 24 DF,  p-value: 0.0002851

> anova(coral.lm)
Analysis of Variance Table

Response: Density
              Df   Sum Sq  Mean Sq F value    Pr(>F)
Distance       1 0.215260 0.215260 22.3681 8.261e-05 ***
I(Distance^2)  1 0.009772 0.009772  1.0155    0.3237
Residuals     24 0.230964 0.009623
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> postscript("coral-resid.ps")
> par(mfrow=c(2,2))
> plot(coral.lm, ask=F)
> dev.off()
```

Which is more appropriate? We can compare this model to the model that assumes that each location has its own mean by fitting an one-way AOV model.

```
> coral.aov <- aov(Density ~ Reef, data=coral)
> summary(coral.aov)
            Df  Sum Sq Mean Sq F value    Pr(>F)
Reef         8 0.41909 0.05239  25.549 2.615e-08 ***
Residuals   18 0.03691 0.00205
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> postscript("coral-aov-resid.ps"); par(mfrow=c(2,2)) ; plot(coral.aov, ask=F); dev.off()
```

Now construct an ANOVA table to compare the two models.

```
> anova(coral.lm, coral.aov)
Analysis of Variance Table

Model 1: Density ~ Distance + I(Distance^2)
Model 2: Density ~ Reef
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     24 0.230964
2     18 0.036908  6  0.194056 15.774 2.740e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpret?

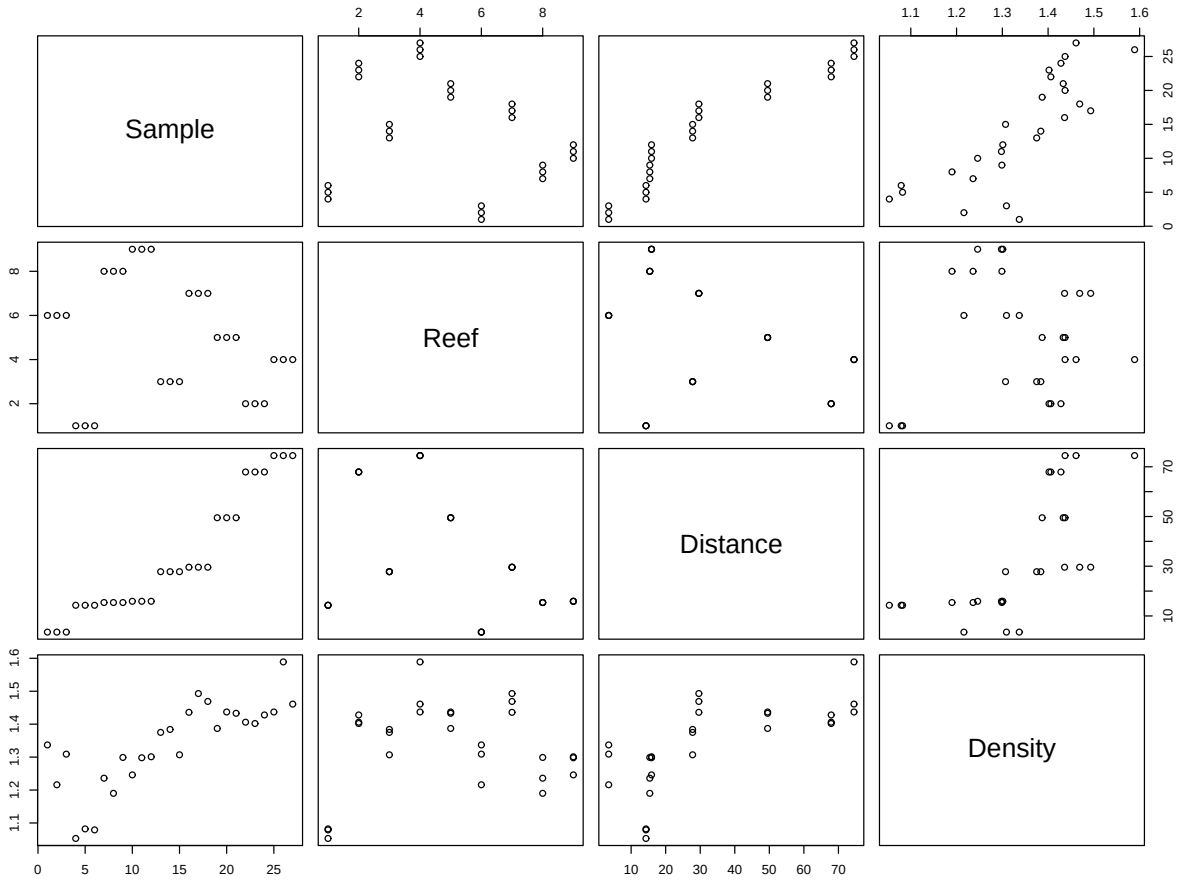

Figure 1: Pairs plot of coral reef data

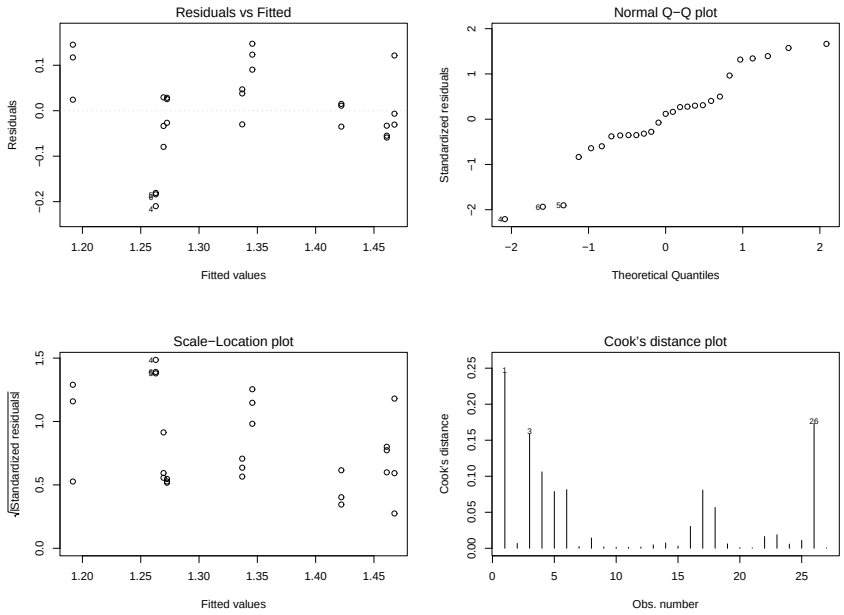

Figure 2: Residual and diagnostic plot of coral reef data with the linear regression model using Distance

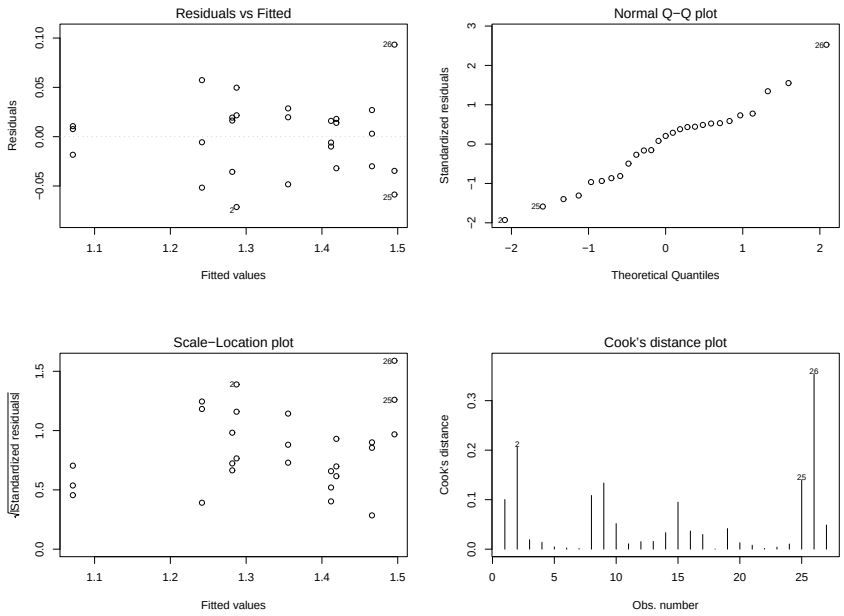

Figure 3: Residual and diagnostic plot of coral reef data with the AOV model using Reef