Regression Analysis and Lack of Fit

February 24, 2009

Christensen Chapters 4 & 6

Coral Reef Example

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.

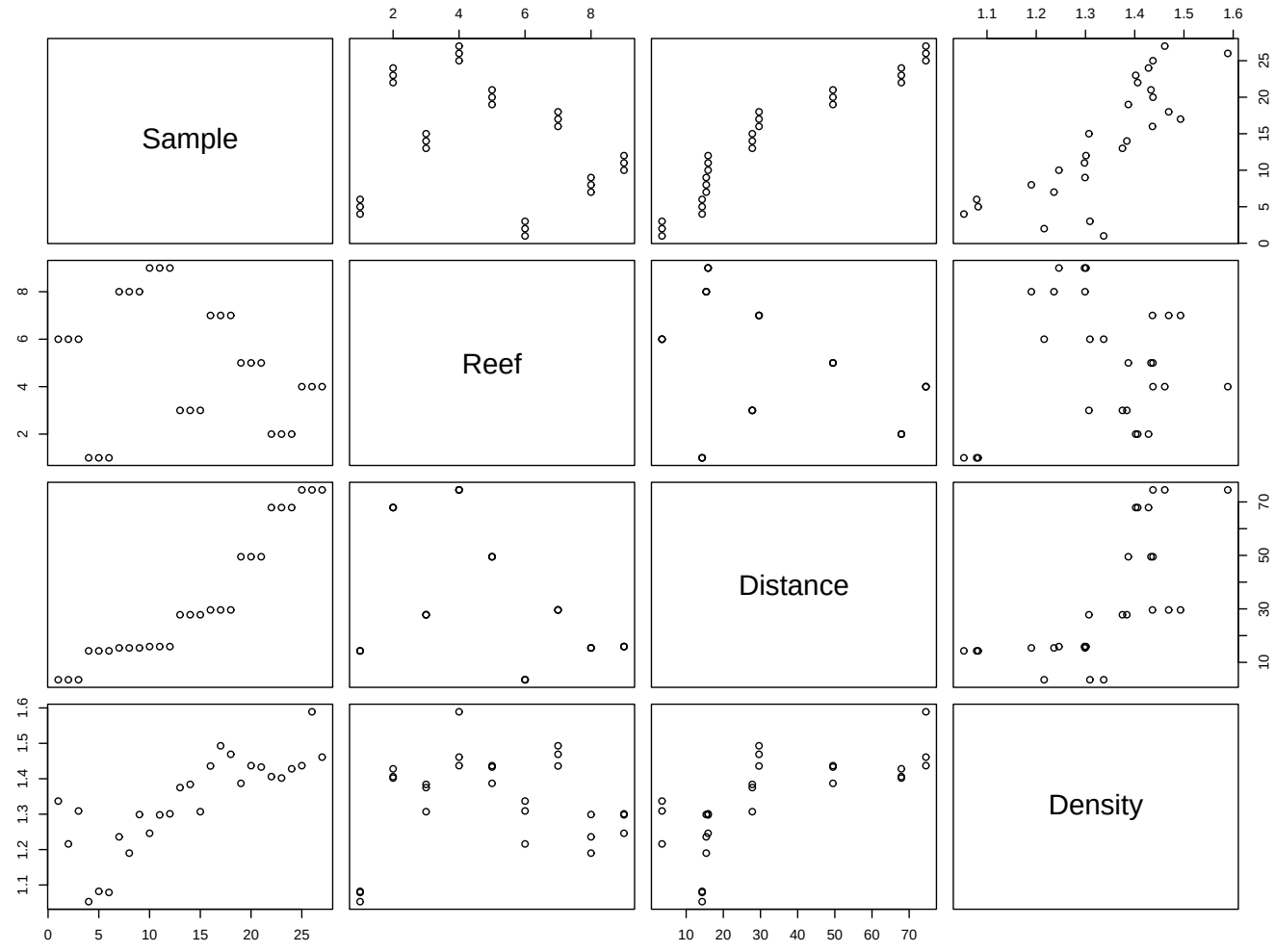
Statistical Models?

Summaries

```
> coral <- read.table("http://www.stat.duke.edu/
           courses/Spring09/sta244/Handouts/rwg.179.txt",
           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
```

Response: `density`; `Reef`: gives the name of each reef, and is a categorical, `distance`: (continuous)

Pairs Plot



Statistical Models

RS summarized the relstionship between density and distance with a second order polynomial,

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

where Y_i is the density and D_i is the distance. This is still a linear (regression) model as it is linear in the parameters.

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

The `I()` inhibits the formula interpretation of mathematical operations

PartialOutput

```
> summary(coral.lm)

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(Distanceˆ2) -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
```

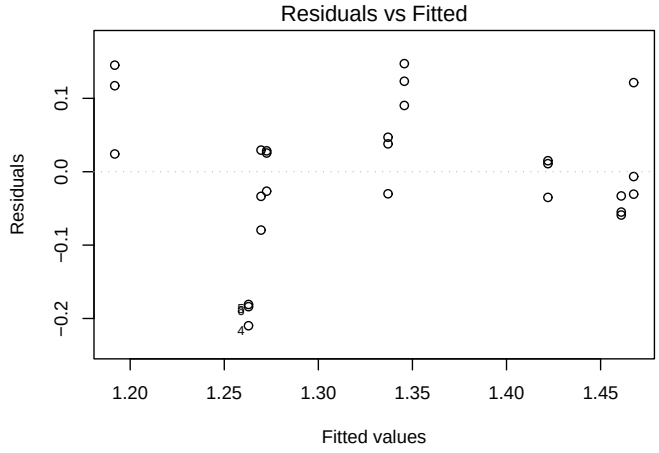
Interpretations/Conclusions?

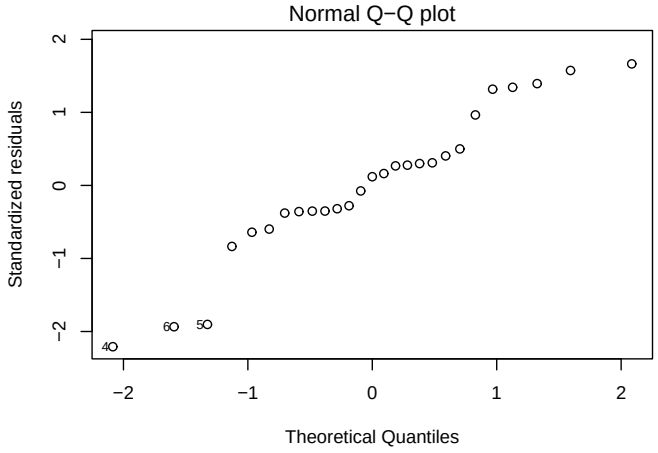
ANOVA Table

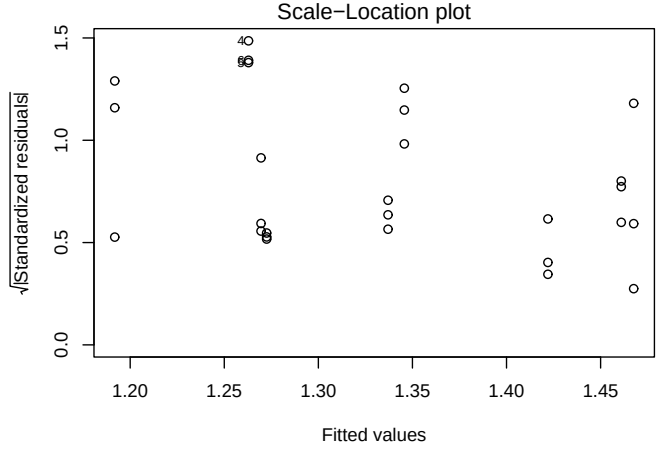
```
> anova(coral.lm)
Analysis of Variance Table

Response: Density
              Df   Sum Sq  Mean Sq F value    Pr(>F)
Distance       1 0.215260 0.215260 22.3681 8.261e-05 ***
I(Distanceˆ2)  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
```

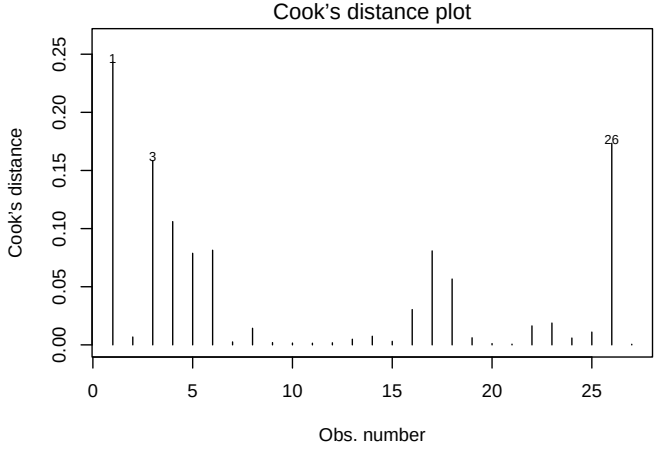
Is this model adequate?

Residuals for Quadratic Regression Model









Models

Models for the expected density for observation i at reef j:

- $E(Y_{ij}) = \mu$ (all observations have same mean, independent of distance or reef)
- $E(Y_{ij}) = \beta_0 + \beta_1 D_j + \beta_1 D_j^2$ (regression model, density changes with distance)
- $E(Y_{ij}) = \mu_j$ (all observations in the same reef have the same mean, not necessarily as quadratic function of distance)

We can compare the regression model to the model that assumes that each location has its own mean by fitting an one-way AOV model and carrying out an F-test, based on the extra SS.

AOV ANOVA

```
> 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
```

Interpretation?

Comparing two models

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
```

Lack of Fit F-test

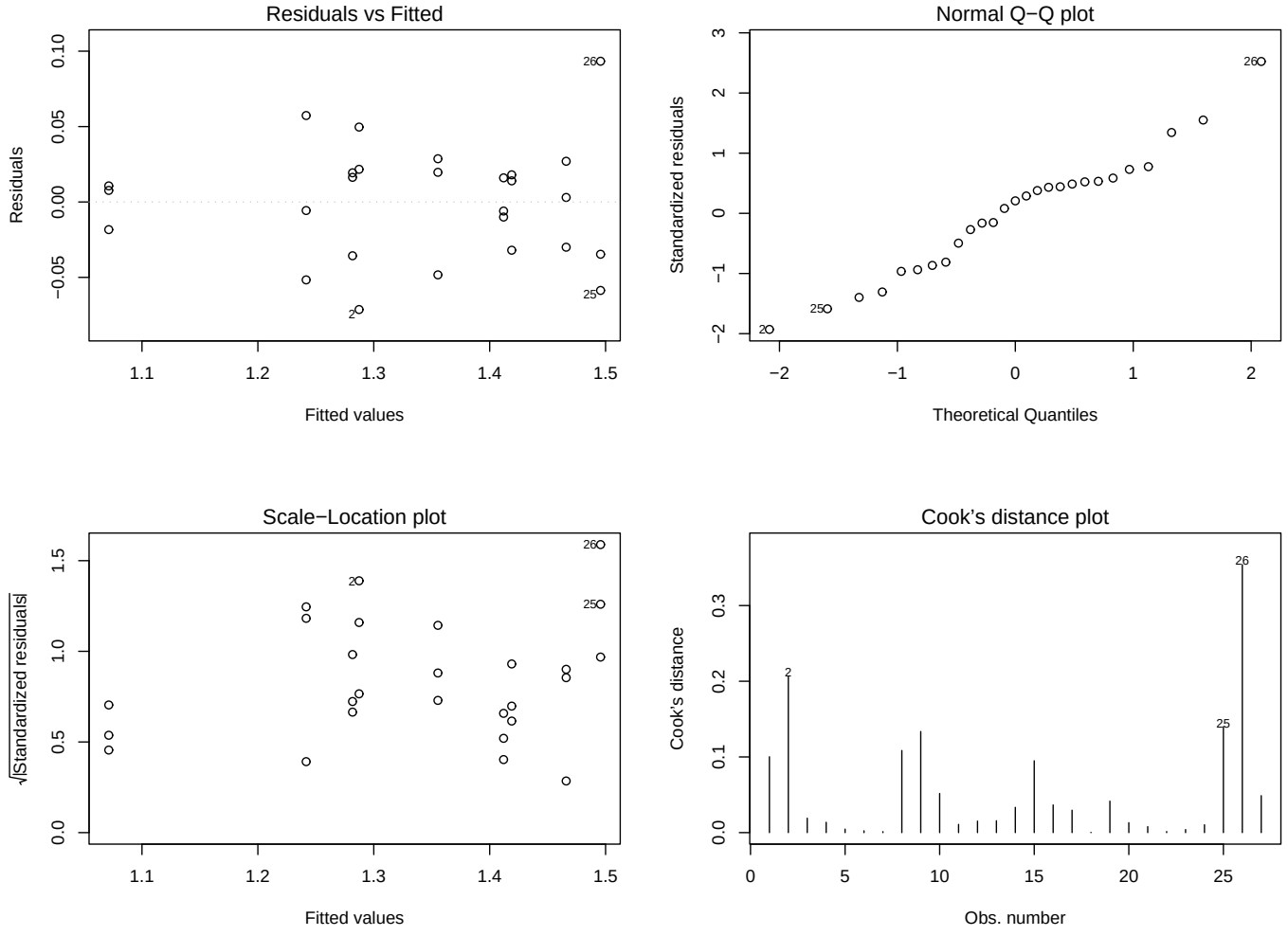
Source	df	SS	MS	F	p-value
Between Groups	8	0.4191			
Regression	2	0.2250			
Lack of Fit	6	0.1941	0.03235	15.78	2.740e-06
Within Group	18	0.0369	0.00205		

Two views of the decomposition:

Between Groups = Regression + Lack of Fit of Regression Model

Error in Regression Model = Lack of Fit + Within Group (Pure error)

Residuals from AOV Model



Conclusions ?