STATS 330

Handout 16
Explanatory Regression Models

Department of Statistics, University of Auckland

16.1 Outline

16.1.1 Using Regression Models for Explanation

Regression models can be used to explore how a response is related to a set of explanatory variables. We will explore two ways regression models can be used to do this:

1. *Descriptive modeling* is when the regression model is used to describe the association (e.g., correlation) between the response and the explanatory variables.
2. *Causal modeling* is when the regression model is used to explore *causal relationships*.

Regression models indicate relationships that may or may not be causal. There is no test that can be done to establish that the association detected by a regression model is the result of causation. However, under certain circumstances it is reasonable to assume causation exists (this will be discussed in detail later).

16.2 Descriptive Modeling

The first type of explanatory modeling we will consider is *descriptive modeling*.

- The regression model is used to describe how one variable, designated as the response, is related to one or more explanatory variables.
- Can be used in conjunction with graphical techniques (exploratory data analysis) to discover interesting relationships between variables in a data set.
- Keep in mind that relationships, even if they are strong, are not necessarily causal.

Most of the regression models that you looked at in STATS 201 could be classified as descriptive models. We will look at a couple of examples of descriptive models to illustrate how regression models are used in this context.

16.2.1 Perching Birds Example I

A student (Amy Moore) at Grinnell College (Iowa, USA) collected data from readily available sources that pertains to North American *passerines* (perching birds).

16.2.1 Perching Birds Example II

- Data set is available in the Stat2Data package in **R**.
- This data can be used investigate relationships between characteristics related to the nesting habits of the different species.
- Regression models and suitable plots are effective tools in exploring these relationships.
- For example, we may be interested in investigating whether the amount of time spent incubating eggs is related to the amount of time the chicks remain in the nest.
- Causality is not an issue in this example as we wouldn't expect any causal relationships among the measured characteristics.

Bird Data Variables

Length	Mean body length (cm).
NestType	Type of nest built.
OorC	Is the nest open or closed?
Location	Location of the nest. Note that `decid` means that the nest is in a *deciduous* tree and `conif` means that it is in a *coniferous* tree (the remaining levels are self explanatory).
Eggs	Average number of eggs.
Marking	1 indicates eggs have markings and 0 indicates eggs have no markings.
Incubate	Mean length of time (in days) the eggs are incubated.
Nestling	Mean length of time (in days) the babies are cared for in the nest.
TotalCare	Total care time = Incubate + Nestling.

A Quick Look at the Data

```
summary(nest.df)
```

```
##        X                  Length          NestType      Location        Eggs
##  Length:84          Min.   : 9.0   burrow   : 2   decid  :24   Min.   : 1.00
##  Class :character   1st Qu.:14.0   cavity   :17   ground :19   1st Qu.: 3.50
##  Mode  :character   Median :17.0   crevice  : 3   shrub  :17   Median : 4.50
##                     Mean   :17.6   cup      :53   conif  :14   Mean   : 4.58
##                     3rd Qu.:20.0   pendant  : 2   bank   : 3   3rd Qu.: 5.00
##                     Max.   :31.5   saucer   : 4   snag   : 3   Max.   :12.50
##                                    spherical: 3   (Other): 4
##  Marking     Incubate        Nestling       TotalCare        OorC
##  no :14   Min.   :10.0   Min.   : 8.0   Min.   :19.0   closed:27
##  yes:70   1st Qu.:12.0   1st Qu.:11.4   1st Qu.:23.5   open  :57
##           Median :13.0   Median :14.0   Median :27.5
##           Mean   :13.3   Mean   :14.4   Mean   :27.7
##           3rd Qu.:14.0   3rd Qu.:17.0   3rd Qu.:31.0
##           Max.   :17.0   Max.   :22.5   Max.   :37.5
##           NA's   :1                     NA's   :1
```

Some Possible Questions

1. How are incubation time and nestling time related? Is this relationship the same for "closed nest" and "open nest" species?
2. Are any of the other variables related to the mean number of eggs for a species? If so what do these relationships look like.
3. Is the presence of markings on the eggs related to whether the nest is open or closed?
4. Is the presence of markings on the eggs related to the location of the nest?
5. Is the location of the nest related to whether the nest is open or closed?
6. Is the nestling time related to whether the nest is open or closed?

Answering the Questions

Regression models can be employed to help answer those questions.

- We need to be careful to make statements about associations between variables rather than statements that imply causality.
- We also need to be careful about effect modification: the relationship between two variables may be affected by the value of a third variable.
  - Effect modification is explored by adding appropriate interactions to the regression model and by using conditional plots.
- If we are trying to answer a specific question, then the model(s) we use are (to a large extent) determined by the question.

16.2.1.1 Answering (a Specific) Question 1

As an example of using regression to answer a specific question, we will consider question 1 in our list:

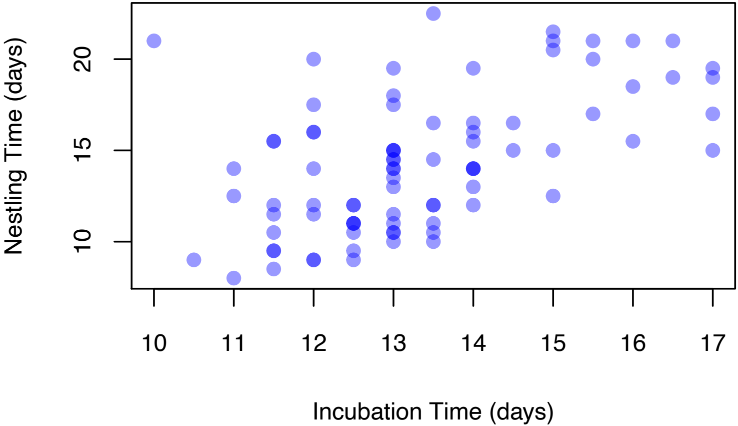
> *How are incubation time and nestling time related? Is this relationship the same for "closed nest" and "open nest" species?*

Our analysis strategy will consist of 3 key steps.

1. Data exploration.
2. Model building.
3. Model interpretation.

Data Exploration

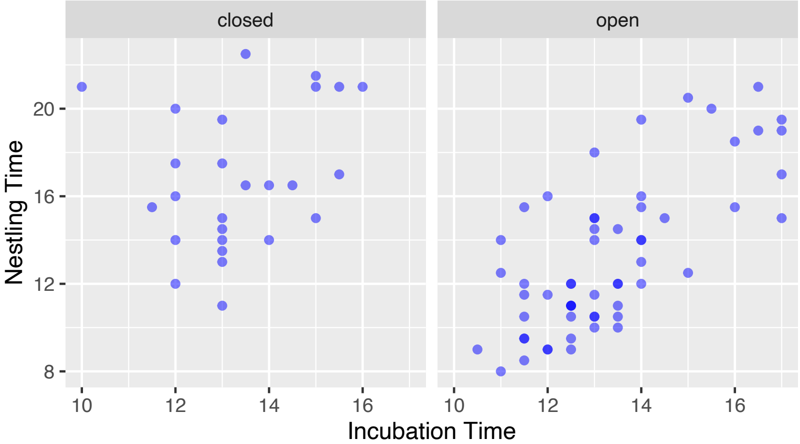
We start by plotting nestling time versus incubation time.



- A positive relationship with quite a bit of scatter around the trend.
- One unusual observation (short incubation time but long nestling time)

Conditioning on an Open or Closed Nest

We can use a conditional plot to see how the incubation/nestling relationship depends on nest type.



- A positive relationship is apparent for open nest species but not so clear for closed nest species.

Initial Assessment

- The scatter plot of nestling time versus incubation does appear to indicate a positive correlation.
- There is considerable scatter in these plots.
- When we condition on whether the nest is open or closed we see a clear difference between the relationships.
  - The relationship for open nests seems much stronger than for closed nests.
  - In fact for closed nests it isn't clear that there is a relationship.
- We should fit a regression model that contains incubation time, nesting time and their interaction.

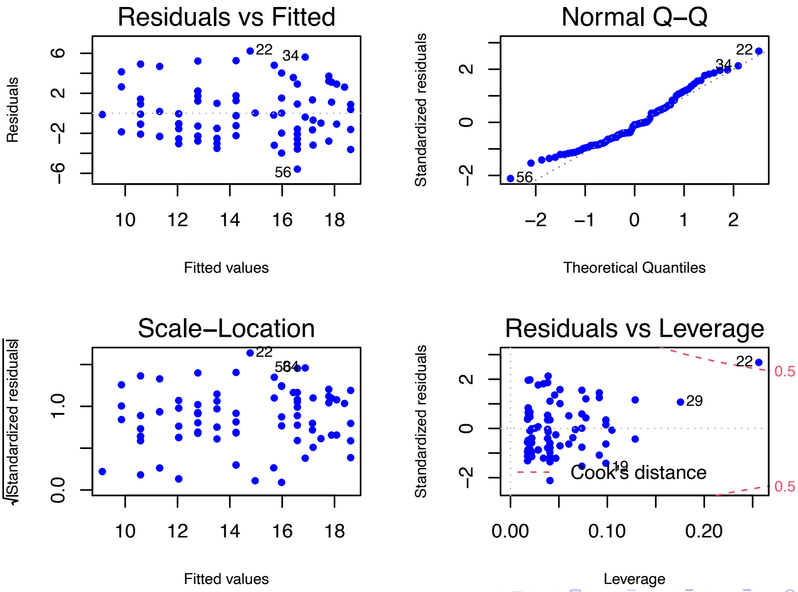
A First Model

Now we enter the model building phase of our analysis. As a starting point:

```
fit1.lm = lm(Nestling ~ Incubate * OorC, data=nest.df)
summary(fit1.lm)
```

```
##
## Call:
## lm(formula = Nestling ~ Incubate * OorC, data = nest.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.580 -2.083 -0.237  1.622  6.224
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8.762      5.038    1.74    0.086 .
## Incubate             0.601      0.375    1.60    0.113
## OorCopen           -14.970      5.776   -2.59    0.011 *
## Incubate:OorCopen    0.859      0.430    2.00    0.049 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.69 on 79 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.507,Adjusted R-squared:  0.488
## F-statistic: 27.1 on 3 and 79 DF,  p-value: 3.86e-12
```

Diagnostic Plots



Box-Cox Plots

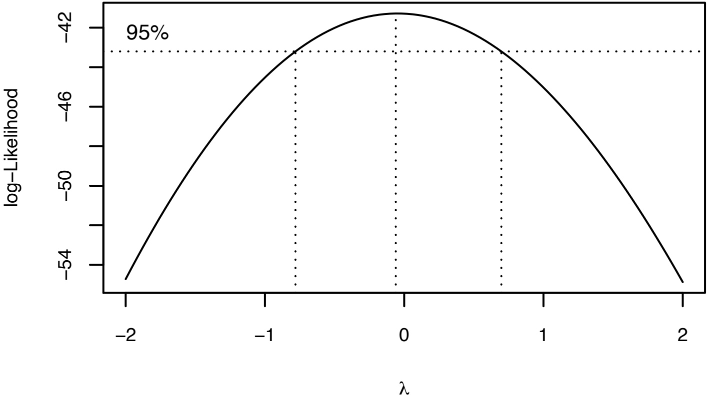
A Box-Cox plot is a convenient method of assessing whether or not transforming the reponse will be useful for an ordinary regression model. Think of a more general model:

$$Y^{\lambda} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

- For different values of λ fit the model and calculate the log-likelihood.
- Plot the values of the log-likehood versus λ—the maximum value of the log-likehood corresponds to the value of λ that results in the model that best fits the data.
- Choose a value of λ that corresponds to a value of the log-likehood that is close to the maximum vlaue.
- $\lambda = 0$ represents a log transformation.

A Box-Cox Plot

We might consider a transformation of the response as Normal Q–Q indicates non-normality.



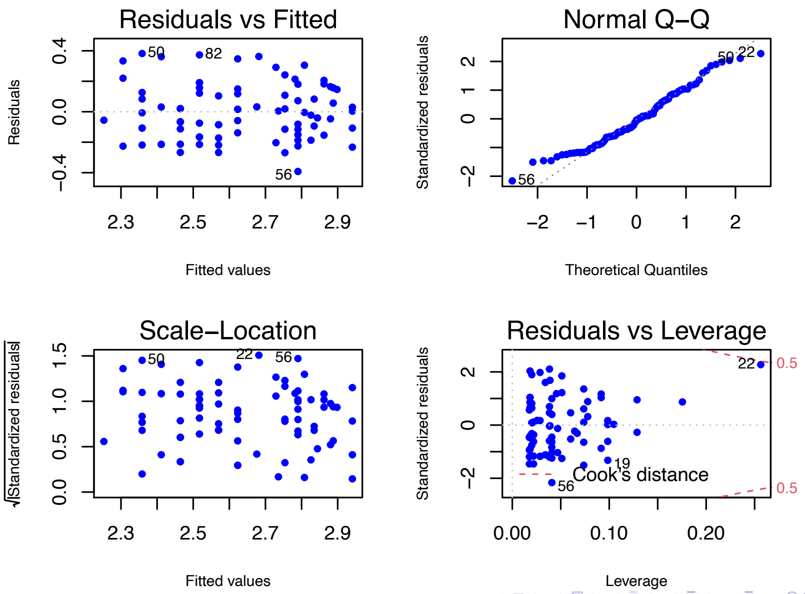
- The Box-Cox plot suggests a log transformation.

The Log Model

If we try logging the response:

```
fit2.lm = lm(log(Nestling) ~ Incubate*OorC, data=nest.df)
summary(fit2.lm)


##
## Call:
## lm(formula = log(Nestling) ~ Incubate * OorC, data = nest.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3921 -0.1446 -0.0072  0.1363  0.3823
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.3211     0.3464    6.70  2.8e-09 ***
## Incubate              0.0361     0.0258    1.40   0.1662
## OorCopen             -1.1798     0.3971   -2.97   0.0039 **
## Incubate:OorCopen     0.0698     0.0296    2.36   0.0208 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.185 on 79 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.527,Adjusted R-squared:  0.509
## F-statistic: 29.3 on 3 and 79 DF,  p-value: 7.83e-13
```

More Diagnostic Plots



Some Conclusions I

Based on the two sets of diagnostic plots, the logged model is slightly better than the original model.

- Observation 22 shows up as having a high value of Cook's Distance.

```
options(width=100)
nest.df[22, ]

##         X Length  NestType Location Eggs Marking Incubate Nestling TotalCare   OorC
## 22 Verdin   10.5 spherical    shrub  4.5     yes       10       21        31 closed
```

This is the observation we noted as having an usual combination of incubation and nestling times. Checking the "Audubon Guide to North American Birds" indicates that the given times are correct.

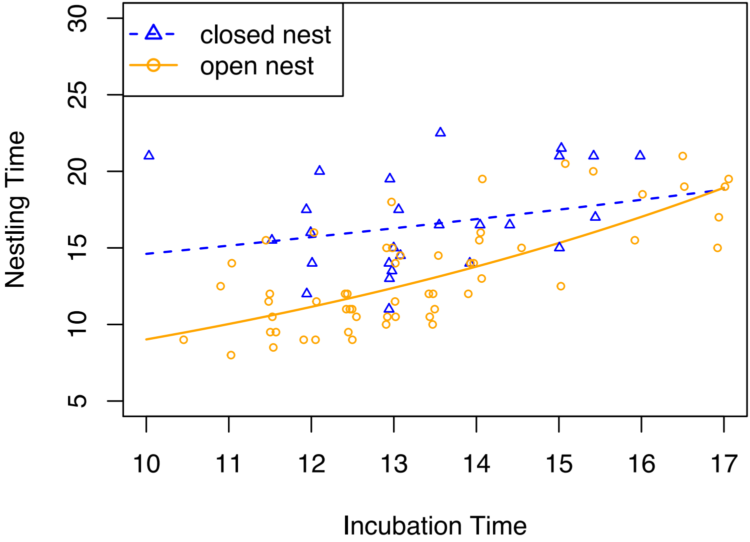
Some Conclusions II

So we retain observation 22 and base our conclusions on the coefficients for the logged model.

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.3211     0.3464   6.701   0.0000
## Incubate             0.0361     0.0258   1.397   0.1662
## OorCopen            -1.1798     0.3971  -2.971   0.0039
## Incubate:OorCopen    0.0698     0.0296   2.360   0.0208
```

- The estimated coefficient for incubate is not significant. This indicates that for closed nest there is no clear evidence of a relationship between nestling time and incubation time.
- The estimated coefficient for the interaction is significant. In this case this indicates evidence that the coefficient for incubation time is different for open and closed nest species.

Scatter Plot with Trend Lines



Summary

From our analysis we would conclude the following about the relationship between incubation time and nestling time.

- The relationship between incubation time and nestling time is different for closed nest species and open nest species.
  - For open nest species there is a clear increase in nestling time as incubation time increases.
  - For closed nest species there is no clear indication of an increase in nestling time as incubation time increases.
- For low incubation times, open nest species tend to have lower nestling times than open nest species. But as incubation time increases, the difference between nestling times for open and closed nest species decreases.
- There is quite a bit of scatter around the trend lines for both open nest and closed nest species.

16.2.1.2 Exploratory Modeling for Question 2

We have just looked at an example where predictive modeling was used to answer a specific question:

> *How are incubation time and nestling time related? Is this relationship the same for "closed nest" and "open nest" species?*

In other cases the question may be more exploratory in nature. For example question 2 on our list:

> *Are any of the other variables related to the mean number of eggs for a species? If so what do these relationships look like?*

- We will now look at how we can approach an exploratory type of question.

Data Management I

Before we start our analysis we should take a look at the data and decide which variables we want to consider as potential regressors.

To start, the factor `NestType` has several levels with small numbers.

```
summary(nest.df$NestType)
```

```
##     burrow    cavity   crevice       cup   pendant    saucer spherical
##          2        17         3        53         2         4         3
```

Instead we will use the factor `OorC` which was created by collapsing the levels "cup" and "saucer" into "open" and the rest of the levels into "closed".

```
summary(nest.df$OorC)
```

```
## closed   open
##     27     57
```

Data Management II

The factor `Location` also has levels with small numbers.

```
summary(nest.df$Location)

##     bank   bridge building    cliff    conif    decid
##        3        1        2        1       14       24
##   ground    shrub     snag
##       19       17        3
```

Instead, we will create a factor `Ground` with levels "ground" and "other".

```
nest.df <- transform(nest.df,
              Ground = ifelse(Location == "ground", "ground", "other"))
nest.df$Ground <- factor(nest.df$Ground)
with(nest.df, summary(Ground))

## ground  other
##     19     65
```

Data Management III

The variable `TotalCare` is the sum of `Incubate` and `Nestling` and thus is redundant.

So we will consider `Length`, `Marking`, `OorC`, `Ground`, `Incubate` and `Nestling` as possible regressors. A new data frame containing these variables plus `Eggs` is created.

```
head(new.df)
```

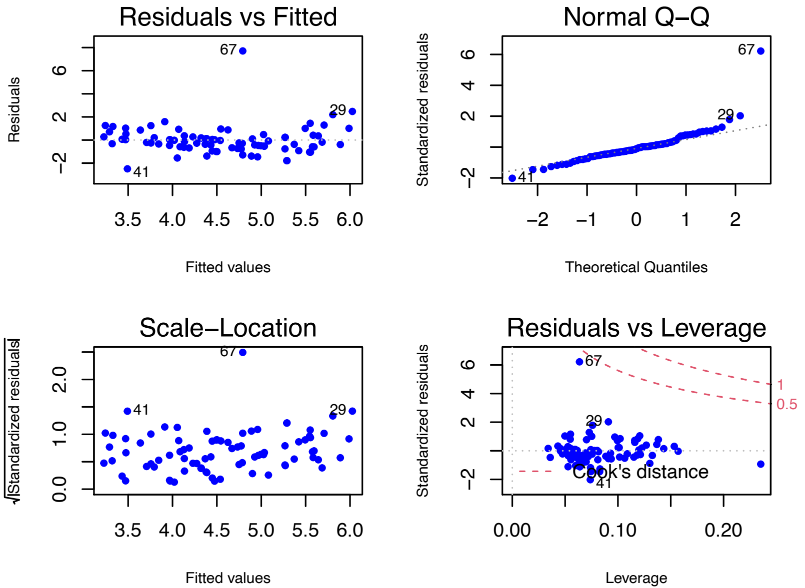
```
##                              X Length Eggs Marking Incubate Nestling   OorC Ground
## 1            Eastern Kingbird   20.0  3.5     yes     17.0     17.0   open  other
## 2 Sulphur-bellied Flycatcher    20.0  3.5     yes     15.5     17.0 closed  other
## 3     Ash-thoated Flycatcher    20.0  4.5     yes     15.0     15.0 closed  other
## 4   Brown-crested Flycatcher    22.5  4.5     yes     14.0     16.5 closed  other
## 5    Dusky-capped Flycatcher    17.0  4.5     yes     14.0     14.0 closed  other
## 6              Eastern Phoebe   17.0  4.5      no     16.0     15.5   open  other
```

Egg Model

The full model using the average number of eggs as the response:

```
##
## Call:
## lm(formula = Eggs ~ Length + Marking + OorC + Ground + Incubate +
##     Nestling, data = new.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.492 -0.590 -0.135  0.362  7.708
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.07965    1.24032    4.90  5.3e-06
## Length      -0.07355    0.03305   -2.23   0.0290
## Markingyes   0.21354    0.39030    0.55   0.5859
## OorCopen    -1.21099    0.38928   -3.11   0.0026
## Groundother -0.53806    0.39284   -1.37   0.1748
## Incubate     0.07409    0.12151    0.61   0.5439
## Nestling    -0.00853    0.05525   -0.15   0.8777
##
## Residual standard error: 1.28 on 76 degrees of freedom
## Multiple R-squared:  0.27,Adjusted R-squared:  0.212
## F-statistic: 4.68 on 6 and 76 DF,  p-value: 0.000425
```

Egg Model Diagnostic Plots



Diagnostic Plots Interpretation

- Observation 67 is clearly an outlier and has moderately high influence . . . this point should be investigated.
- The "Residuals vs Fitted Values Plot" indicates a non-linear relationship.

First investigate observation 67:

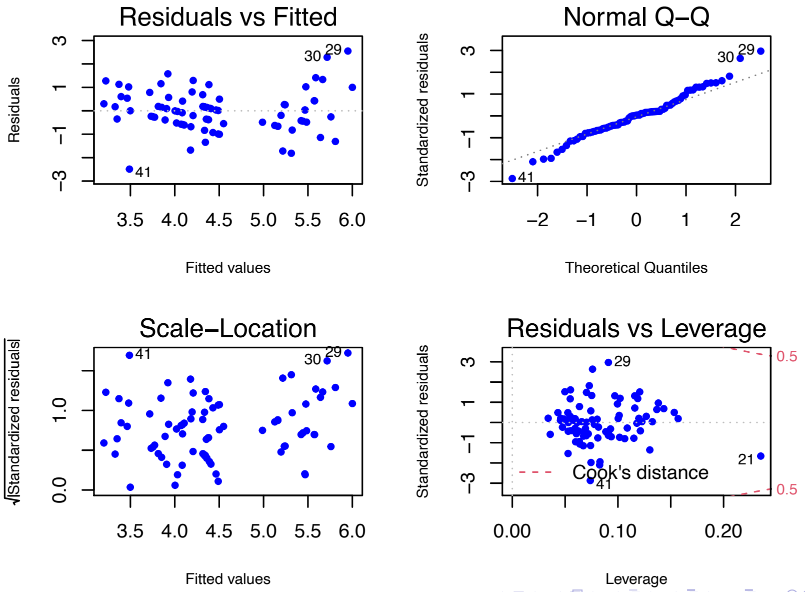
```
new.df[67, ]
```

```
##                           X Length Eggs Marking Incubate Nestling OorC Ground
## 67 American Tree Sparrow   15.5 12.5     yes     12.5        9 open ground
```

The average number of eggs (12.5) seems very high. I checked the internet and most sources say 4 to 6 eggs. So we will replace this value with 5.

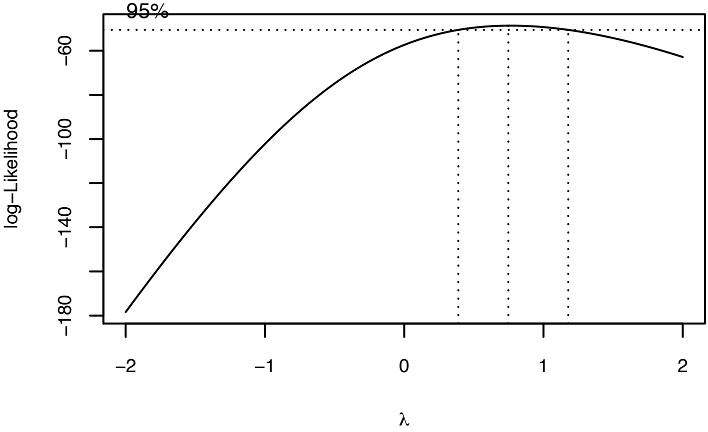
```
new.df$Eggs[67]=5
```

Diagnostic Plots for Adjusted Data



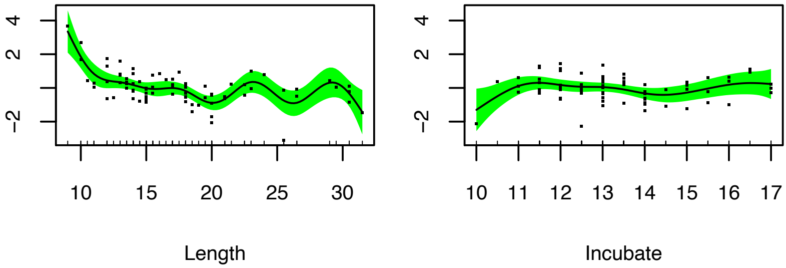
Box-Cox Plot

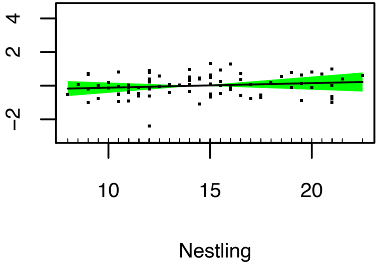
Given that we are still seeing curvature in the "Residuals vs Fitted Values" plot, we might consider a transformation of the response.



- The Box-Cox plot indicates no transformation is needed.

GAM Plots





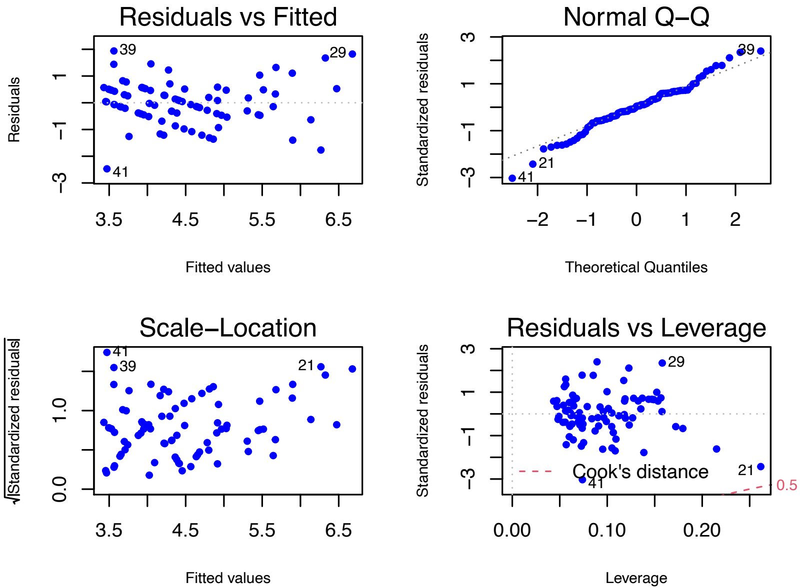
Add quadratic term for `Length`?

Quadratic Model

```
egg3.lm = lm(Eggs~Length+I(Length^2)+Marking+OorC+Ground+Incubate+Nestling, data=new.df)
summary(egg3.lm)


##
## Call:
## lm(formula = Eggs ~ Length + I(Length^2) + Marking + OorC + Ground +
##     Incubate + Nestling, data = new.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4714 -0.4321  0.0344  0.5015  1.9377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.89597    1.46858    6.74  2.9e-09 ***
## Length      -0.45220    0.11695   -3.87  0.00023 ***
## I(Length^2)  0.00968    0.00291    3.32  0.00137 **
## Markingyes  -0.01903    0.26427   -0.07  0.94280
## OorCopen    -1.17616    0.25752   -4.57  1.9e-05 ***
## Groundother -0.29968    0.26206   -1.14  0.25646
## Incubate     0.01438    0.08047    0.18  0.85862
## Nestling     0.01086    0.03665    0.30  0.76788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.847 on 75 degrees of freedom
## Multiple R-squared:  0.499,Adjusted R-squared:  0.453
## F-statistic: 10.7 on 7 and 75 DF,  p-value: 2.95e-09
```

- Adding squared term for `Length` has improved the model. (How?)

Diagnostic Plots for Quadratic Model



Best Subset Models I

Now we will look for the best subset model using the `dredge()` function.

```
options(na.action = "na.fail", width=110)
egg.fits <- dredge(egg3.lm)

## Fixed term is "(Intercept)"

head(egg.fits, n=10)

## Global model call: lm(formula = Eggs ~ Length + I(Length^2) + Marking + OorC + Ground +
##     Incubate + Nestling, data = new.df)
## ---
## Model selection table
##     (Intrc) Grond    Incbt   Lngth  Lngth^2 Mrkng    Nstln OorC df  logLik  AICc delta weight
## 77    9.852               -0.4377 0.009244                    +  5 -100.53 211.8  0.00  0.307
## 78   10.190     +         -0.4521 0.009770                    +  6  -99.93 213.0  1.14  0.174
## 109   9.769               -0.4366 0.009189       0.004781     +  6 -100.51 214.1  2.30  0.097
## 93    9.900               -0.4404 0.009310     +              +  6 -100.52 214.1  2.31  0.097
## 79    9.765      0.006571 -0.4370 0.009201                    +  6 -100.52 214.1  2.32  0.096
## 110   9.985     +         -0.4507 0.009673       0.014470     +  7  -99.81 215.1  3.28  0.060
## 80    9.879     + 0.026110 -0.4509 0.009650                   +  7  -99.85 215.2  3.37  0.057
## 94   10.260     +         -0.4558 0.009861     +              +  7  -99.92 215.3  3.50  0.053
## 125   9.813               -0.4386 0.009243     + 0.004207     +  7 -100.51 216.5  4.68  0.030
## 111   9.757      0.001281 -0.4366 0.009184       0.004489     +  7 -100.51 216.5  4.69  0.029
## Models ranked by AICc(x)
```

Best Subset Models II

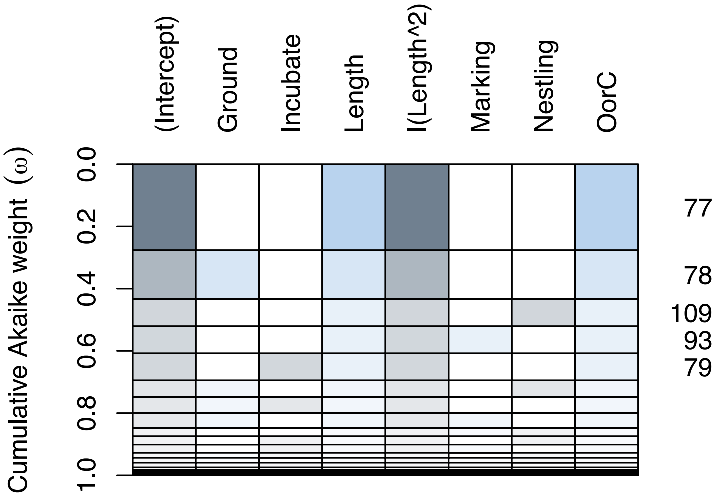
Note the following:

- All of the 10 best models (as ranked by AICc) contain `Length`, $\texttt{Length}^2$ and `OorC`.
- The estimated coefficients for `Length` and $\texttt{Length}^2$ are quite consistent over these models (the same is true for `OorC`).
- The variables `Ground`, `Incubation`, `Marking` and `Nestling` occur sporadically over this list.
- Notice the "weight" columns: these values sum to 1 and can be interpreted as the "weight" that should be attached to each model.

We conclude that the average number of eggs is related to `Length` and `OorC`.

A Useful Plot

```r
par(mar=c(4, 6, 6, 4))
plot(egg.fits)
```



Interaction(s)?

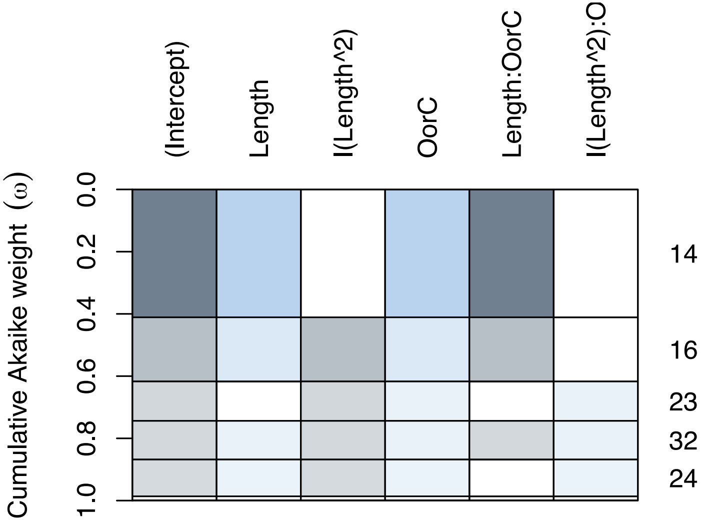
It is prudent to check for possible interactions.

```
egg4.lm = lm(Eggs~(Length+I(Length^2))*OorC, data=new.df)
options(na.action = "na.fail", width=110)
egg.fits2 <- dredge(egg4.lm)


## Fixed term is "(Intercept)"

head(egg.fits2, n=10)


## Global model call: lm(formula = Eggs ~ (Length + I(Length^2)) * OorC, data = new.df)
## ---
## Model selection table
##      (Int)      Lng       Lng^2 OrC Lng:OrC I(Lng^2):OrC df  logLik  AICc delta weight
## 14  9.003 -0.23280                +       +               5  -97.11 205.0  0.00  0.411
## 16  9.730 -0.33260  0.0032230     +       +               6  -96.64 206.4  1.38  0.206
## 23  7.234          -0.0072250     +                    +  5  -98.29 207.4  2.36  0.127
## 32 12.710 -0.74210  0.0164500     +       +            +  7  -95.95 207.4  2.39  0.124
## 24  8.741 -0.20410 -0.0007129     +                    +  6  -97.19 207.5  2.49  0.118
## 8   9.852 -0.43770  0.0092440     +                       5 -100.53 211.8  6.83  0.014
## 6   6.460 -0.06567                +                       4 -106.15 220.8 15.80  0.000
## 7   5.789          -0.0013370     +                       4 -107.93 224.4 19.37  0.000
## 5   5.462                         +                       3 -111.64 229.6 24.58  0.000
## 4  10.090 -0.51890  0.0104900                             4 -113.55 235.6 30.60  0.000
## Models ranked by AICc(x)
```

The top 2 models seem to stand out from the rest.

A Plot of the Best Models

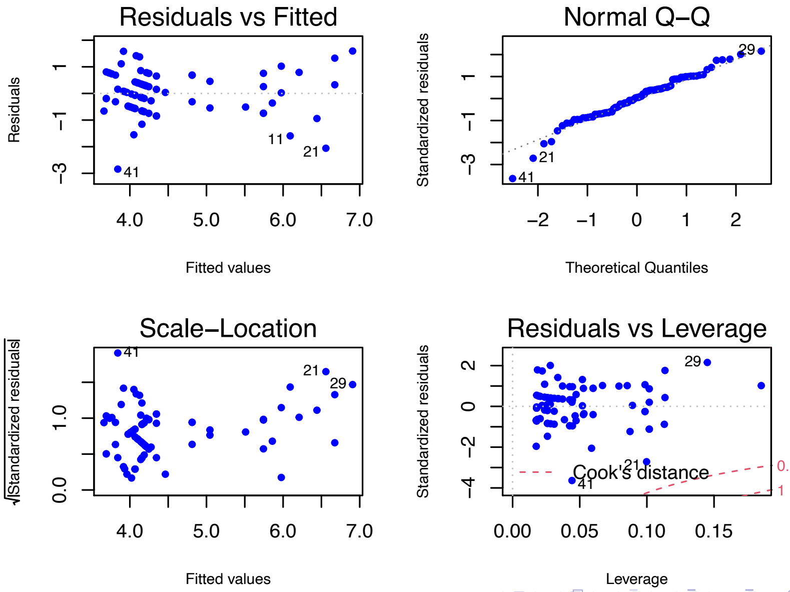
```r
par(mar=c(4, 6, 6, 4))
plot(egg.fits2)
```



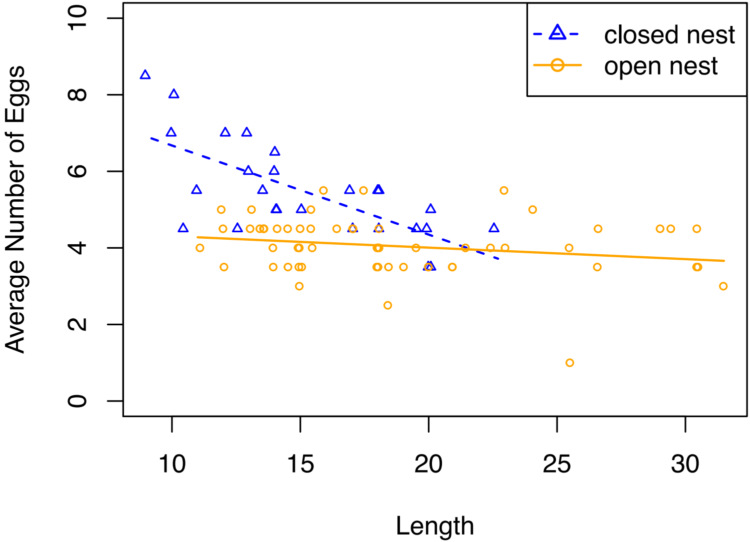
The Best Model

```
##
## Call:
## lm(formula = Eggs ~ Length + OorC + Length:OorC + 1, data = new.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8419 -0.5239  0.0368  0.4439  1.5925
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.0026     0.6578   13.69  < 2e-16 ***
## Length           -0.2328     0.0420   -5.54  3.8e-07 ***
## OorCopen         -4.3953     0.7593   -5.79  1.4e-07 ***
## Length:OorCopen   0.2028     0.0463    4.38  3.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.799 on 79 degrees of freedom
## Multiple R-squared:  0.531,Adjusted R-squared:  0.513
## F-statistic: 29.8 on 3 and 79 DF,  p-value: 5.58e-13
```

Diagnostic Plots for Best Model



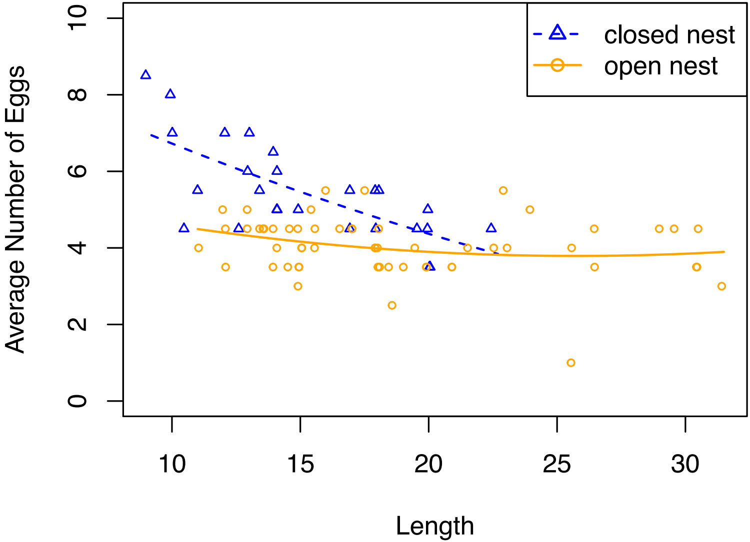
Best Model: Scatter Plot with Trend Lines



The Second Best Model

```
##
## Call:
## lm(formula = Eggs ~ Length + I(Length^2) + OorC + Length:OorC +
##     1, data = new.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7903 -0.4634  0.0832  0.5039  1.6852
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.72957    1.01197    9.61    7e-15 ***
## Length          -0.33256    0.11355   -2.93  0.00446 **
## I(Length^2)      0.00322    0.00341    0.95  0.34716
## OorCopen        -3.79921    0.98718   -3.85  0.00024 ***
## Length:OorCopen  0.16645    0.06015    2.77  0.00706 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8 on 78 degrees of freedom
## Multiple R-squared:  0.536,Adjusted R-squared:  0.512
## F-statistic: 22.5 on 4 and 78 DF,  p-value: 2.17e-12
```

Second Model: Scatter Plot with Trend Lines



Summary

Both of the top two models are giving much the same message about how the average number of eggs laid is related to the species length and whether the nest is open or closed.

- The relationship between eggs laid and length is different for closed nest species and open nest species.
  - For open nest species the average number of eggs laid is close to 4 for species of all lengths.
  - For closed nest species the average number of eggs laid decreases as species length increases.
- For smaller lengths, closed nest species tend to lay more eggs than open nest species. But as length increases, the difference between number of eggs laid for open and closed nest species decreases.
- There are no large closed nest species.