Model Selection for Explanatory Models

Chapter 7, Lab 5: Solutions

OpenIntro Biostatistics

Topics

- Building explanatory models
- Transforming variables
- Model comparison with adjusted R^2

In previous labs, multiple regression modeling was shown in the context of estimating an association while adjusting for possible confounders. This lab introduces explanatory modeling, in which the goal is to construct a model that explains the observed variation in the response variable. Explanatory modeling is concerned with identifying predictors associated with the response; there is no pre-specified primary predictor of interest.

The material in this lab corresponds to Section 7.8 in *OpenIntro Biostatistics*.

Introduction

Approaches to model selection vary from those based on careful study of a relatively small set of predictors to purely algorithmic methods that screen a large set of predictors and choose a final model by optimizing a numerical criterion. This course discusses model selection in the context of a small set of potential predictors.

Model selection for explanatory modeling follows these general steps:

1. *Data exploration*. Examine both the distributions of individual variables and the relationships between variables.
2. *Initial model fitting*. Fit an initial model with the predictors that seem most highly associated with the response variable, based on the data exploration.
3. *Model comparison*. Work towards a model with the highest adjusted R^2.
4. *Model assessment*. Use residual plots to assess the fit of the final model.

The process behind model selection will be illustrated with a case study in which a regression model is built to examine the association between the abundance of forest birds in a habitat patch and features of a patch.

Background Information

Habitat fragmentation is the process by which a habitat in a large contiguous space is divided into smaller, isolated pieces. Smaller patches of habitat are only able to support limited populations of organisms, which reduces genetic diversity and overall population fitness. Ecologists study habitat fragmentation to understand its effect on species abundance.

The `forest.birds` dataset in the `oibiostat` package contains a subset of the variables from a 1987 study analyzing the effect of habitat fragmentation on bird abundance in the Latrobe Valley of southeastern Victoria, Australia.[1]

The dataset consists of the following variables, measured for each of the 57 patches.

- `abundance`: average number of forest birds observed in the patch, as calculated from several independent 20-minute counting sessions.
- `patch.area`: patch area, measured in hectares. 1 hectare is 10,000 square meters and approximately 2.47 acres.
- `dist.nearest`: distance to the nearest patch, measured in kilometers.
- `dist.larger`: distance to the nearest patch larger than the current patch, measured in kilometers.
- `altitude`: patch altitude, measured in meters above sea level.
- `grazing.intensity`: extent of livestock grazing, recorded as either "light", "less than average", "average", "moderately heavy", or "heavy".
- `year.of.isolation`: year in which the patch became isolated due to habitat fragmentation.
- `yrs.isolation`: number of years since patch became isolated due to habitat fragmentation.[2]

Data exploration

1. Identify the variables in the dataset relevant for modeling the relationship between species abundance and features of a habitat; that is, the response variable and the potential predictor variables.

   The response variable is `abundance`; this variable contains information on the average number of birds observed in a particular patch over several independent 20-minute counting sessions. There are six potential predictor variables: patch area, distance to the nearest patch, distance to the nearest patch larger than the current patch, patch altitude, grazing intensity, and number of years since isolation. The variable `year.of.isolation` is used to calculate the more informative variable, `yrs.isolation`, which is a measure of the number of years the patch became isolated as of the time the study was conducted.

2. Explore the distribution of each variable with numerical and graphical summaries.

   a) Briefly describe the distribution of each variable.

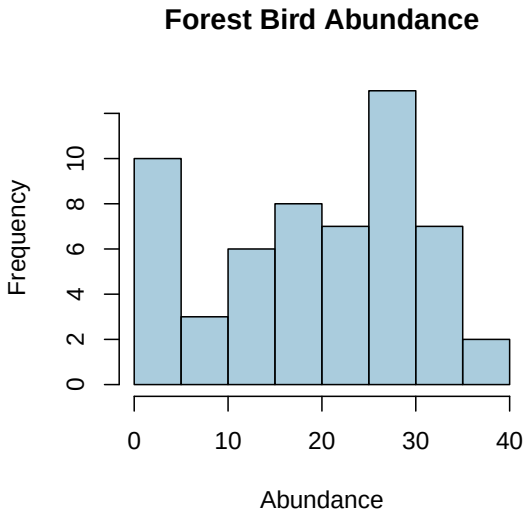
   Bird abundance ranges from 1.5 to 39.6, with a mean of 21.0. The distribution of bird abundance is bimodal, with modes at small values of abundance and at between 25 and 30 birds.

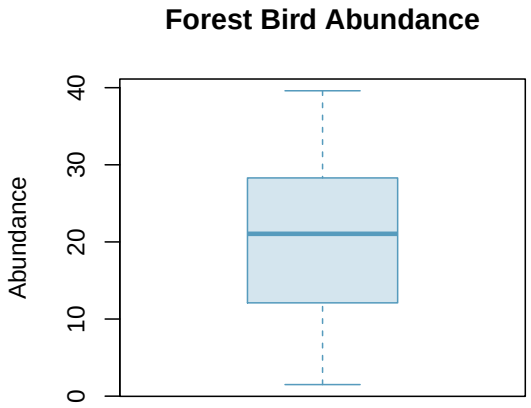
   There is right skewing visible in patch area, distance to nearest patch, distance to larger patch, and years of isolation. The right skewing is particularly notable for patch area;

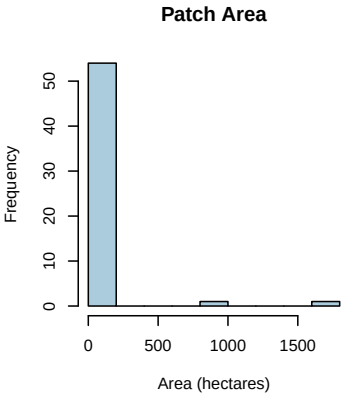
[1] Loyn, R.H. 1987. "Effects of patch area and habitat on bird abundances, species numbers and tree health in fragmented Victorian forests." Printed in Nature Conservation: The Role of Remnants of Native Vegetation. Saunders DA, Arnold GW, Burbridge AA, and Hopkins AJM eds. Surrey Beatty and Sons, Chipping Norton, NSW, 65-77, 1987.

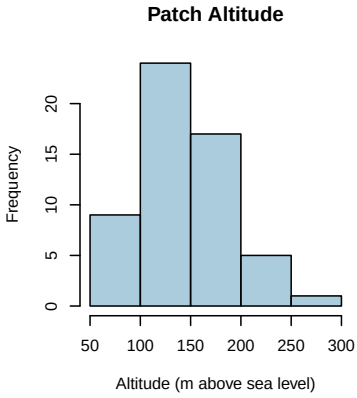
[2] The Loyn study completed data collection in 1983; `yrs.isolation` = 1983 – `year.of.isolation`.

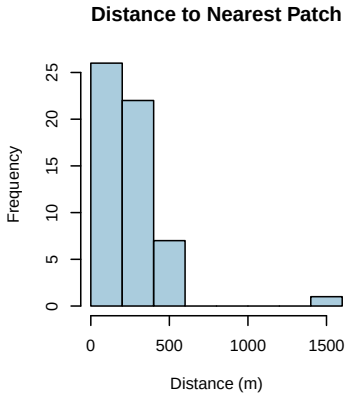
75% of patches have area lower than 30 hectares, but a few patches have area well past 500 hectares. The altitude variable is roughly symmetric, centered around a mean of 146.2 meters above sea level. The grazing intensity variable indicates that the three largest categories are light grazing, average grazing, and heavy grazing.

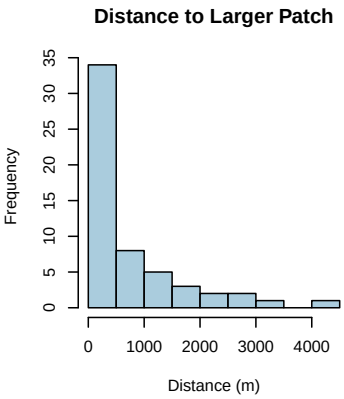


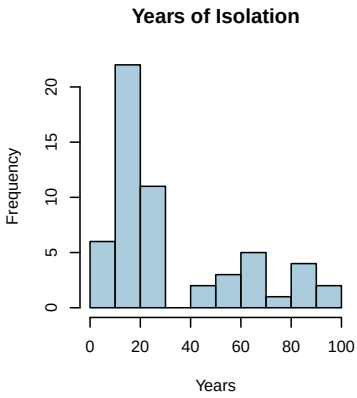


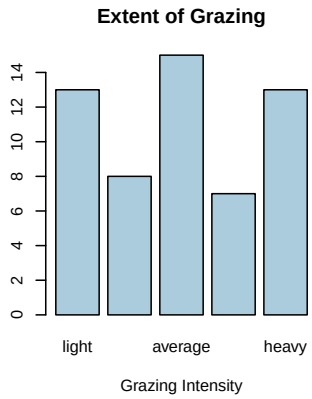












```r
#numerical summaries
summary(forest.birds$abundance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.50   12.40   21.05   19.51   28.30   39.60
```

```r
summary(forest.birds$patch.area)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.10    2.00    7.50   69.27   29.75 1771.00
```

```r
summary(forest.birds$altitude)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    60.0   120.0   140.0   146.2   182.5   260.0
```

```r
summary(forest.birds$dist.nearest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    26.0    93.0   234.0   240.4   333.2  1427.0
```

```r
summary(forest.birds$dist.larger)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    26.0   158.2   338.5   733.3   913.8  4426.0
```

```r
summary(forest.birds$yrs.isolation)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   17.00   20.50   33.25   55.50   93.00
```

```r
table(forest.birds$grazing.intensity)
```

```
##
##             light less than average           average  moderately heavy
##                13                 8                15                 7
##             heavy
##                13
```

b) A common technique to improve model fit in linear regression (particularly in regards to achieving approximate linearity) is to transform variables that exhibit skew. A natural log transformation can help induce symmetry in right-skewed variables.

Identify which variables could benefit from a natural log transformation. Apply the transformation and use the transformed version going forward.

The variables patch area, distance to nearest patch, distance to larger patch, and years of isolation should be log-transformed.

```r
#transform variables
forest.birds$log.area = log(forest.birds$patch.area)
forest.birds$log.dist.nearest = log(forest.birds$dist.nearest)
forest.birds$log.dist.larger = log(forest.birds$dist.larger)
forest.birds$log.yrs.isolation = log(forest.birds$yrs.isolation)
```

c) Examine the relationships between the predictor and response variables, as well as the relationships between predictor variables.

i. Run the code in the template to create a scatterplot matrix. Each subplot in the matrix is a simple scatterplot; the variable names are listed along the diagonal of the matrix and the diagonal divides the matrix into symmetric plots.

Describe what you see. Which variables seem to be strongly associated with the response? Do any predictor variables seem strongly associated with each other?

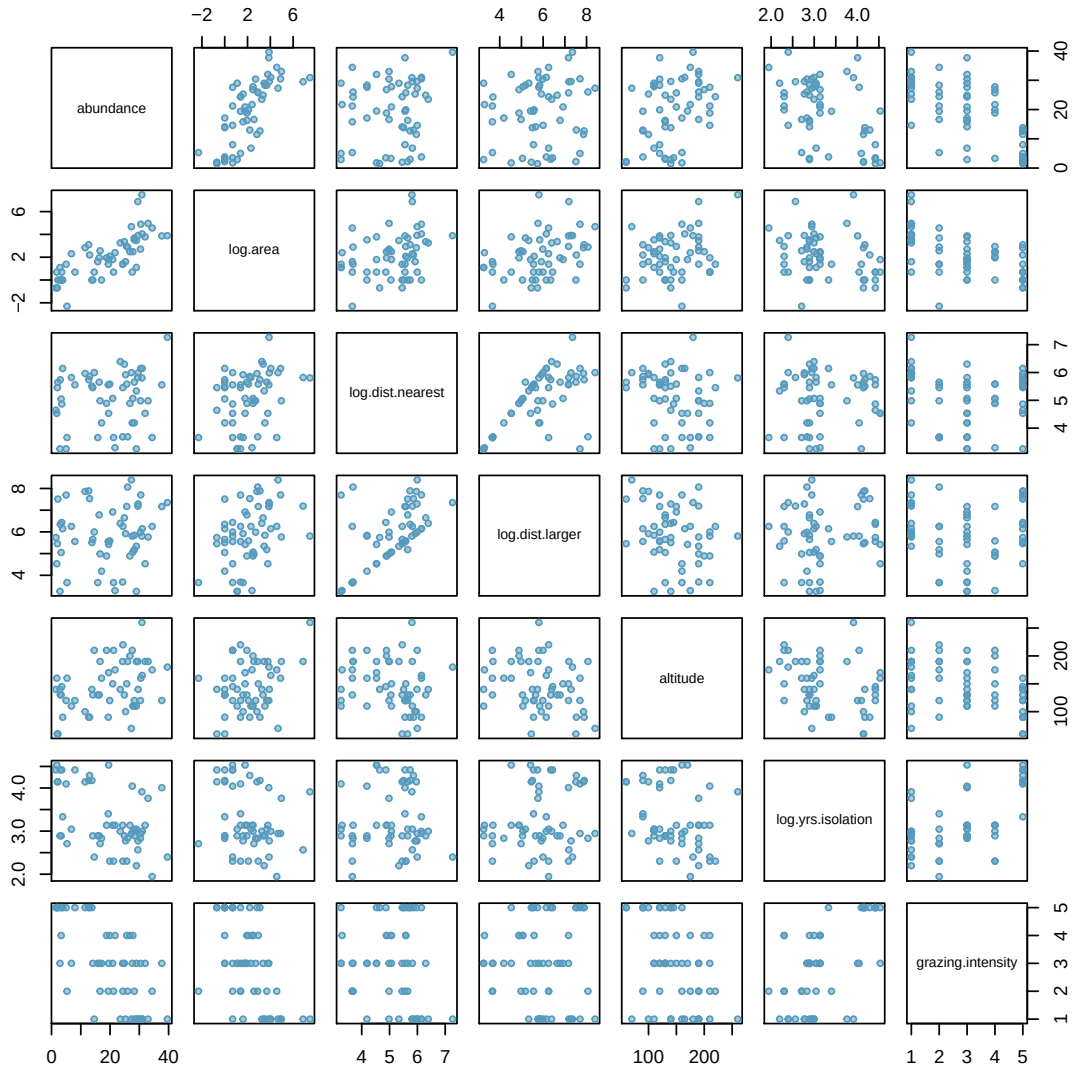
The plots in the first row show the relationships between abundance and the predictors. There is a strong positive association between abundance and `log.area` and a moderate negative association between abundance and `log.yrs.isolation`. Both distance variables seem weakly positively associated with abundance. There is high variance of abundance and somewhat similar centers for the first for grazing intensity categories, but abundance does clearly tend to be lower in the highest category than the others.

The distance variables appear strongly associated; a model may only need one of the two, as they may be essentially "redundant" in explaining variability in the response variable. In this case, however, since both are only weakly associated with the response, both may be unnecessary in a model.

ii. Run the code in the template to create a correlation matrix. Confirm that the numerical summaries cohere with what you observed from the graphical summaries.

A numerical approach confirms some of the features observable from the graphical summaries. Correlations between abundance and `log.area` and between abundance and `log.yrs.isolation` are relatively high, at 0.74 and -0.48, respectively. In contrast, the correlations between bird abundance and the distance variables are much smaller, at 0.13 and 0.12. Additionally, the two distance variables have correlation of 0.60.

```
#create a scatterplot matrix
pairs(~ abundance + log.area + log.dist.nearest
  + log.dist.larger  + altitude + log.yrs.isolation +
    grazing.intensity, data = forest.birds,
  pch = 21, cex = 0.7, bg = COL[1, 3], col = COL[1])
```




```
#subset numerical variables
forest.subset = subset(forest.birds, select = c(abundance, log.area, log.dist.nearest,
                                    log.dist.larger, altitude, log.yrs.isolation))

#create a correlation matrix
cor(forest.subset)
```

```
##                   abundance   log.area log.dist.nearest log.dist.larger
## abundance         1.0000000  0.7400358       0.12672333       0.1181245
## log.area          0.7400358  1.0000000       0.30216662       0.3824795
## log.dist.nearest  0.1267233  0.3021666       1.00000000       0.6038664
## log.dist.larger   0.1181245  0.3824795       0.60386637       1.0000000
## altitude          0.3858362  0.2751428      -0.21900701      -0.2740438
## log.yrs.isolation -0.4796380 -0.2506109       0.02274064       0.1491448
##                    altitude log.yrs.isolation
```

```
## abundance          0.3858362       -0.47963800
## log.area           0.2751428       -0.25061092
## log.dist.nearest  -0.2190070        0.02274064
## log.dist.larger   -0.2740438        0.14914481
## altitude           1.0000000       -0.28759892
## log.yrs.isolation -0.2875989        1.00000000
```

Initial model fitting

3. Based on the data exploration, which predictor variables should be included in an initial model?

   Based on the data exploration, the initial model should include the variables `log.area`, `altitude`, `log.yrs.isolation`, and `grazing.intensity`.

4. Fit the initial model.

```r
#fit the model
model0 = lm(abundance ~ log.area + altitude + log.yrs.isolation +
                grazing.intensity, data = forest.birds)

#print model summary
summary(model0)
```

```
##
## Call:
## lm(formula = abundance ~ log.area + altitude + log.yrs.isolation +
##     grazing.intensity, data = forest.birds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0135  -2.3512  -0.2195   2.5416  11.5484
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        14.15087    6.30057   2.246  0.02935
## log.area                            3.12216    0.56484   5.528 1.31e-06
## altitude                            0.00799    0.02156   0.371  0.71262
## log.yrs.isolation                   0.12997    1.91931   0.068  0.94629
## grazing.intensityless than average  0.29666    2.99211   0.099  0.92143
## grazing.intensityaverage           -0.16168    2.75351  -0.059  0.95342
## grazing.intensitymoderately heavy  -1.59361    3.03497  -0.525  0.60194
## grazing.intensityheavy            -11.74351    4.33702  -2.708  0.00936
##
## (Intercept)                        *
## log.area                           ***
## altitude
## log.yrs.isolation
## grazing.intensityless than average
```

```
## grazing.intensityaverage
## grazing.intensitymoderately heavy
## grazing.intensityheavy             **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.995 on 48 degrees of freedom
## Multiple R-squared:  0.7279, Adjusted R-squared:  0.6882
## F-statistic: 18.34 on 7 and 48 DF,  p-value: 1.293e-11
```

a) Report the R^2 and adjusted R^2 of the model.

The R^2 of the model is 0.73. The adjusted R^2 of the model is 0.69.

b) Identify which variables are statistically significant at the $\alpha = 0.05$ level.

The variable area (transformed) and one of the categories of grazing intensity.

Model comparison

5. Fit models excluding the predictors that were not statistically significant. Based on comparing the adjusted R^2 values, consider whether any of these models are an improvement from the initial model.

Models excluding either variable have adjusted R^2 of 0.69, while the model excluding both variables has an adjusted R^2 of 0.70. This is a small but noticeable increase from the initial model (adjusted R^2 of 0.69) which suggests both variables can be dropped from the model. At this point, the working model contains only `log.area` and `grazing.intensity`.

```
#model excluding altitude
model1 = lm(abundance ~ log.area + grazing.intensity + log.yrs.isolation,
        data = forest.birds)

summary(model1)$adj.r.squared
```

```
## [1] 0.6936528
```

```
#model excluding log.yrs.isolation
model2 = lm(abundance ~ log.area + grazing.intensity + altitude,
        data = forest.birds)

summary(model2)$adj.r.squared
```

```
## [1] 0.6944973
```

```
#model excluding both
model3 = lm(abundance ~ log.area + grazing.intensity,
        data = forest.birds)

summary(model3)$adj.r.squared
```

```
## [1] 0.699675
```

6. The working model contains the grazing intensity variable. Only one of the coefficients associated with grazing intensity is statistically significant: heavy grazing. Individual categories of a categorical variable cannot simply be dropped, so a data analyst has the choice of leaving the variable as is, or collapsing it into fewer categories.

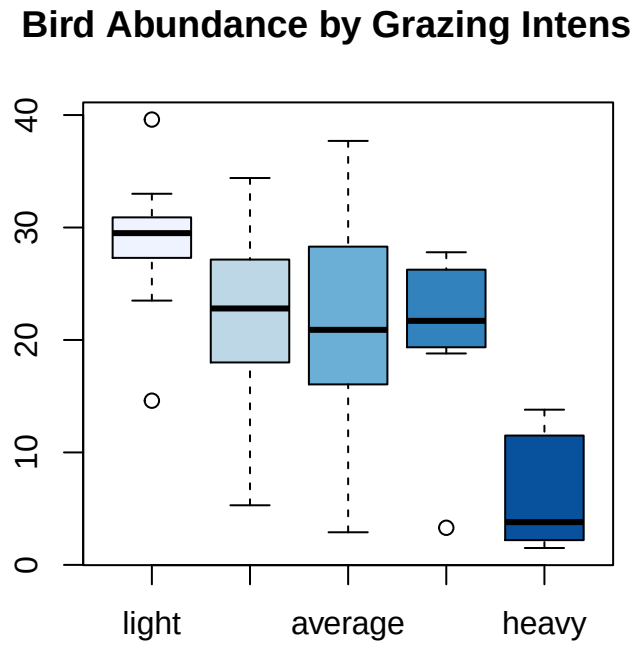
   For this model, it might be useful to collapse grazing-intensity into a two-level variable, with one category corresponding to the original classification of heavy, and another category corresponding to the other four categories.

   a) Create a plot of abundance versus grazing intensity. Does it seem that the distribution of abundance within the lowest four grazing intensity categories is roughly similar, relative to that within the highest category?

      Yes, the centers of the distributions of bird abundance in the four lowest categories are roughly similar, with median bird abundance in the 20 to 30 range. The median bird abundance in the heavy grazing patches is much lower, at about 3 birds; as shown in the plot, the entire distribution is shifted lower than the IQRs of the other grazing categories.

```
#load color package
library(RColorBrewer)

boxplot(abundance ~ grazing.intensity, data = forest.birds,
    main = "Bird Abundance by Grazing Intensity",
    col = brewer.pal(5, "Blues"))
```



   b) Run the code in the template to create `grazing.binary`, which has levels `NotHeavy` and

Heavy.

```
#create the grazing.binary variable
forest.birds$grazing.binary = forest.birds$grazing.intensity

#redefine the factor levels of grazing.binary
levels(forest.birds$grazing.binary) = list(NotHeavy = c("light",
                                                        "less than average",
                                                        "average",
                                                        "moderately heavy"),
                                           Heavy = c("heavy"))
```

c) Fit a model with the binary version of grazing intensity. Is this model an improvement over the model with the original version of grazing intensity?

The model with the binary version of grazing intensity is an improvement over the previous model; the adjusted R^2 increases to 0.71.

```
#fit model with grazing.binary
model4 = lm(abundance ~ log.area + grazing.binary, data = forest.birds)
summary(model4)$adj.r.squared
```

```
## [1] 0.7139773
```

7. Check whether incorporating an interaction term improves the model.

Incorporating an interaction term does not improve the model; adding an interaction term decreases the adjusted R^2 to 0.709.

```
#fit model with interaction term
model5 = lm(abundance ~ log.area*grazing.binary, data = forest.birds)
summary(model5)$adj.r.squared
```

```
## [1] 0.7094178
```

8. Report the variables in the final model and the model R^2.

The final model is `model4`, the model with `log.area` and `grazing.binary` as predictors. The model R^2 is 0.724.

Model assessment

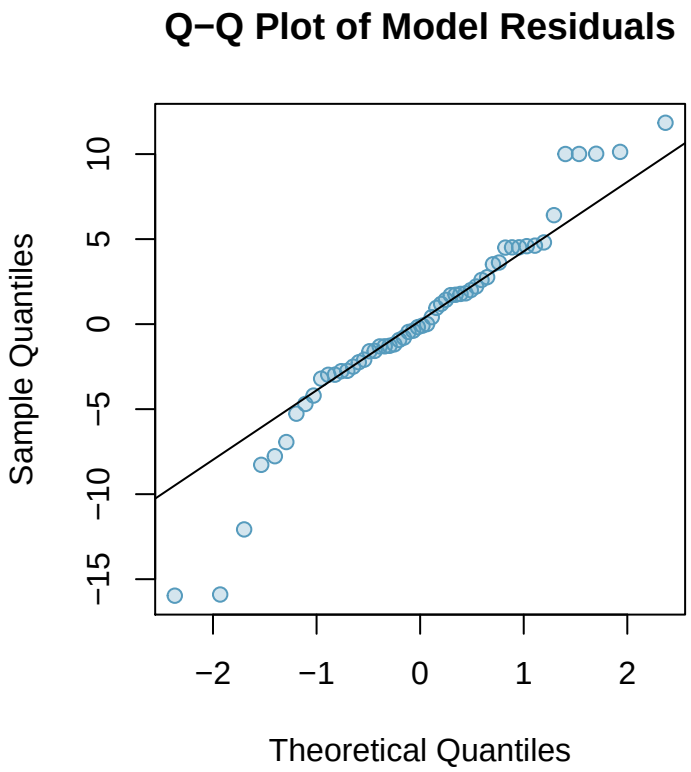
9. Assess whether the residuals are normally distributed.

The residuals follow a normal distribution in the center, but fit less well to a normal curve in the tails. There are too many large positive and large negative values, relative to a normal distribution.

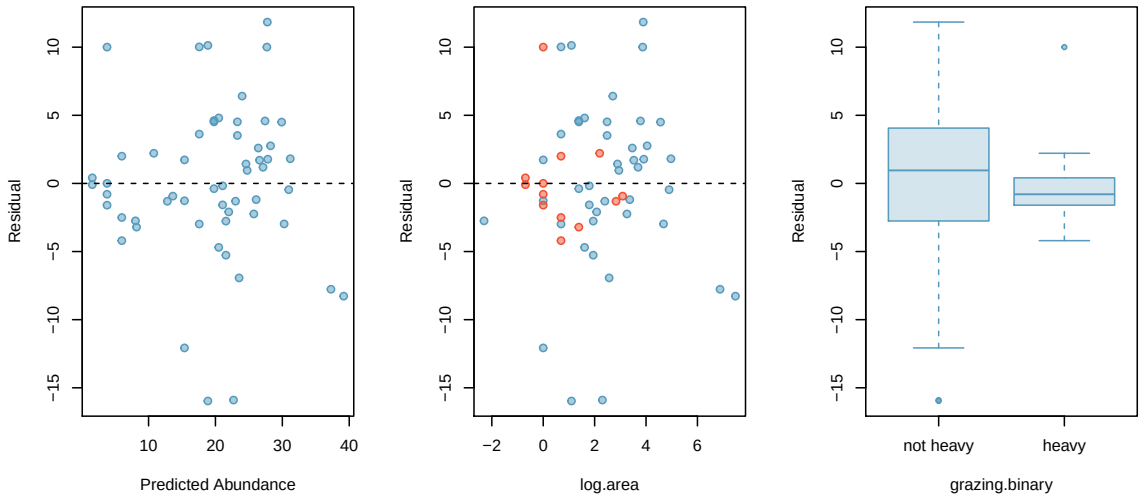
```
#define final model
final.model = model4

#create q-q plot
qqnorm(resid(final.model),
```

```
   pch = 21, col = COL[1], bg = COL[1, 4],
   main = "Q-Q Plot of Model Residuals")
qqline(resid(final.model))
```



10. Run the code in the template to generate three plots that allow for a closer look at the residuals: a plot of residuals versus predicted abundance, and plots of residuals versus the two predictors.



a) Recall that the definition of a residual is $e_i = y_i - \hat{y}_i$. Residual values closer to 0 are indicative of a more accurate prediction. In terms of comparing an observed value and a value predicted from a model, what does a large positive residual indicate? What does a large negative residual indicate?

A large positive residual occurs when the predicted value from the model is much smaller than the observed value; i.e., the model is underpredicting the value. A large negative residual indicates the model is overpredicting the value, and $\hat{y}_i \gg y_i$.

b) Examine the left and middle plot. For what predicted values of bird abundance do large positive residuals tend to occur, versus large negative residuals? For what values of area do large positive residuals versus large negative residuals tend to occur?

In the left plot, the large positive residuals occur across the range of predicted values, while the large negative residuals occur around 20 (predicted birds). The middle plot shows that the large positive and negative residuals occur at intermediate values of `log.area`; they occur for values of `log.area` between 0 and 4 or equivalently for values of area between exp(0) = 1 and exp(4) = 54.5 hectares. In the same range, there are also relatively accurate predictions with most residuals being between -5 and 5.

c) In the middle plot, patches with heavy grazing are represented with red points. From the middle plot and right plot, assess how prediction error varies between patches where grazing intensity was between "light" and "moderately heavy" versus patches where grazing intensity was heavy.

The prediction error is smaller for patches with heavy grazing than for patches where grazing intensity was between "light" and "moderately heavy". Patches with heavy grazing mostly cluster around the $y = 0$ line, with the exception of one point with a residual value of about 10.

Conclusions

11. Summarize the final model; interpret the model coefficients and R^2 value.

    The R^2 indicates that the final model explains 72% of the observed variability in bird abundance, which suggests that patch area and extent of grazing (either heavy or not) are important features associated with bird abundance. Larger area is associated with an increaes in abundance; when grazing intensity does not change, the model predicts an increase in average bird abundance by 3.18 birds for every one unit increase in log area (or equivalently, every exp(1) = 2.7 hectares increase in area). A patch with heavy grazing is estimated to have a mean abundance of about 11.58 birds lower than a patch that has not been heavily grazed, assuming the patches are the same size.

```
#print model summary of final model
summary(final.model)
```

```
##
## Call:
## lm(formula = abundance ~ log.area + grazing.binary, data = forest.birds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9696  -2.5623  -0.1324   2.9502  11.8419
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          15.3736     1.4507  10.597 1.06e-14 ***
## log.area              3.1822     0.4523   7.035 3.96e-09 ***
## grazing.binaryHeavy -11.5783     1.9862  -5.829 3.38e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.741 on 53 degrees of freedom
## Multiple R-squared:  0.7244, Adjusted R-squared:  0.714
## F-statistic: 69.65 on 2 and 53 DF,  p-value: 1.473e-15
```

12. Ecologists might be interested in using the model to predict bird abundance based on features of a forest patch. Summarize the model accuracy, in terms accessible to a non-statistician.

    The final model may not be particularly accurate. For most observations, the predictions made by the model area accurate between ±5 birds, but there are several instances of overpredictions as high as around 10 birds and underpredictions of about 15 birds. A major weakness of the model is that the accurate and inaccurate predictions occur at similar ranges of area; if the model only tended to be inaccurate at a specific range, such as for small patches, it would be possible to provide clearer advice about when the model should be used. The model does seem to be more reliable for patches with heavy grazing.

Refer to Section 7.8 in *OpenIntro Biostatistics* for a discussion of reasons to avoid using a model with all potential predictor variables in this setting.