Exercises

Exercise: Linear model with continuous and categorical explanatory variables

This exercise builds on the linear model with one continuous explanatory variable, and the linear model with one categorical explanatory variable, by adding these two sources of variation in the same model. The first part of the exercise will explore fitting an 'additive' model and the second part fitting a model with an interaction term.

Part 1: Linear model with additive explanatory variables

1. As in previous exercises, either create a new R script (perhaps call it linear_model_3) or continue with your previous R script in your RStudio Project. Again, make sure you include any metadata you feel is appropriate (title, description of task, date of creation etc) and don't forget to comment out your metadata with a `#` at the beginning of the line.

2. Import the data file 'loyn.txt' into R and take a look at the structure of this dataframe using the `str()` function. We know that the abundance of birds `ABUND` increases quickly with the area of the patch `LOGAREA`, and more slowly for the larger patches. We now also know that bird abundance changes in a non-linear way with the grazing intensity `FGRAZE`. But how do these effects combine together? Would a small patch with low grazing intensity have more birds than a larger patch with high grazing intensity? Could the poor fit of the `ABUND ~ LOGAREA` model for the large patches be improved if we accounted for grazing intensity in the patches?

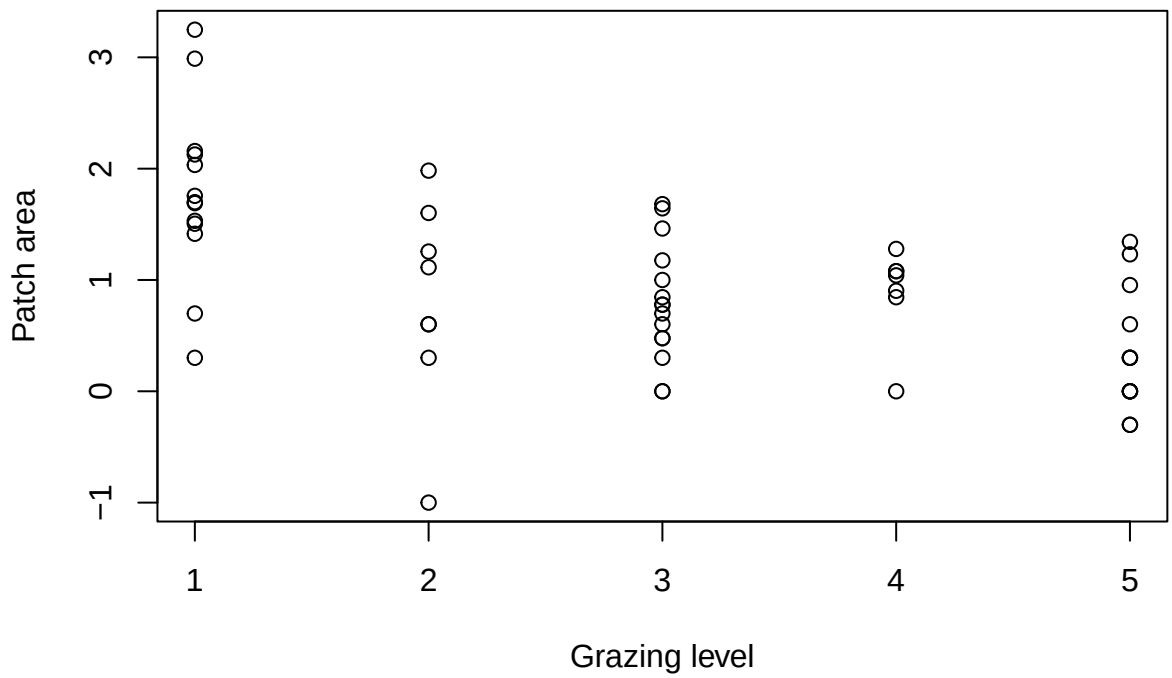
```
loyn <- read.table("data/loyn.txt", header = TRUE, stringsAsFactors = TRUE)
str(loyn)
## 'data.frame':    56 obs. of  8 variables:
##  $ Site   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ABUND  : num  5.3 2 1.5 17.1 13.8 14.1 3.8 2.2 3.3 3 ...
##  $ AREA   : num  0.1 0.5 0.5 1 1 1 1 1 1 1 ...
##  $ DIST   : int  39 234 104 66 246 234 467 284 156 311 ...
##  $ LDIST  : int  39 234 311 66 246 285 467 1829 156 571 ...
##  $ YR.ISOL: int  1968 1920 1900 1966 1918 1965 1955 1920 1965 1900 ...
##  $ GRAZE  : int  2 5 5 3 5 3 5 5 4 5 ...
##  $ ALT    : int  160 60 140 160 140 130 90 60 130 130 ...
```

3. As previously we want to treat `AREA` as a log-transformed area to limit the influence of the few disproportionately large patches, and `GRAZE` as a categorical variable with five levels. So the first thing we need to do is create the corresponding variables in the loyn dataframe, called `LOGAREA` and `FGRAZE`.

```
loyn$LOGAREA <- log10(loyn$AREA)
# create factor GRAZE as it was originally coded as an integer
loyn$FGRAZE <- factor(loyn$GRAZE)
```

4. Explore the relationship between grazing and patch area, using a scatterplot. Hint: you may want to use `GRAZE` rather than `FGRAZE` for this. Is there any variability in patch area within each grazing level? Is the sampling design balanced, i.e., is the whole range of patch areas evenly represented at each grazing level?

```
plot(loyn$LOGAREA ~ loyn$GRAZE, xlab = "Grazing level", ylab = "Patch area")
```

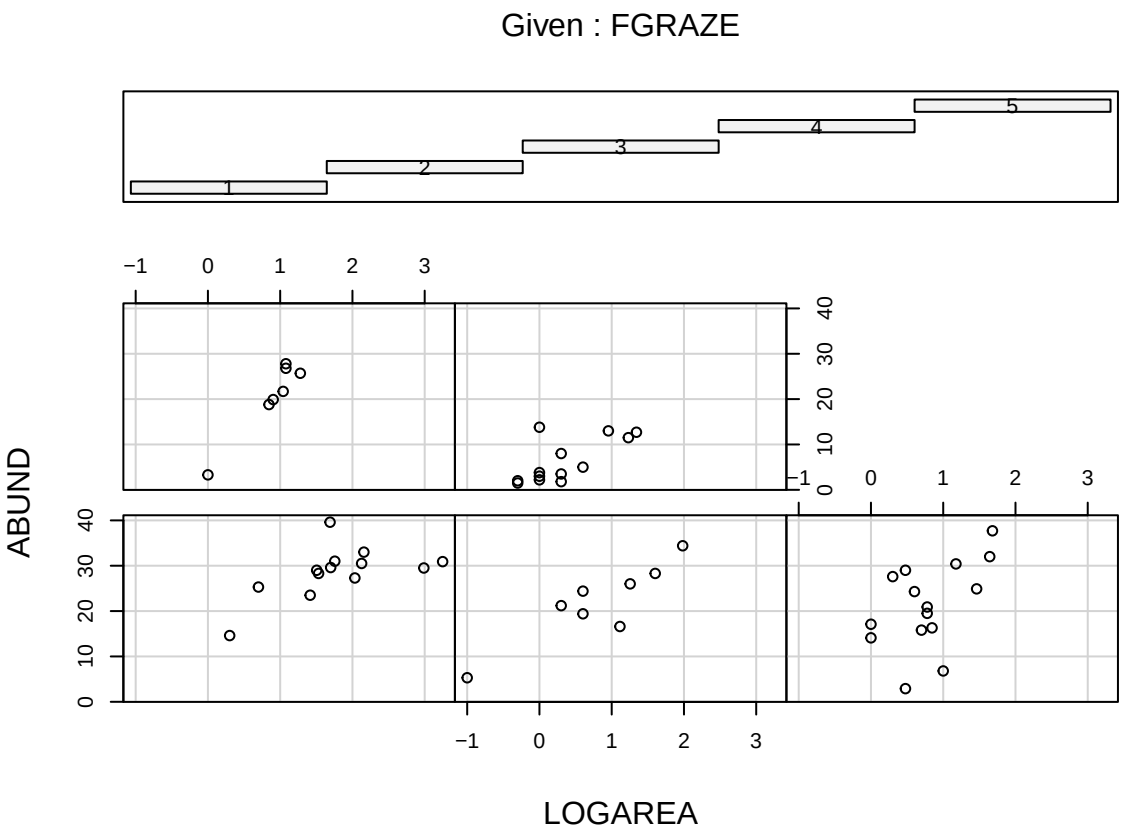



```
# There is a good spread of patch areas within each grazing level overall,
# although there is a trend for more grazing the smaller the patch is.
# the lowest level of grazing intensity happens to be predominantly in
# larger patches (including the two monster patches)

# How would we expect adding grazing level to the LOGAREA model to affect
# the predictions of the model? Think particularly of the largest two patches
# which were previously overestimated by the model (negative residuals -
# see the linear model 1 exercise)?
# Since the lowest grazing levels appear to be associated with the highest
```

```
# bird abundances, we could expect a model combining area and grazing level
# to predict an even higher abundance for these patches
# this would not improve the situation for these patches, at least.
# But let's find out if that's the case!
```

5. You could explore the joint effect of `FGRAZE` and `LOGAREA` on `ABUND`, using panel plots. Hint: See the function `coplot` in the Data exploration lecture, and/or the help page for `coplot`. Factor levels increase from the bottom-left panel to the top-right panel. What pattern do you see? What do you expect your model results to look like?

```
coplot(ABUND ~ LOGAREA | FGRAZE, data = loyn)
```




```
# There is a lot of variation in there, but:
# The mean abundance seems to decrease as grazing levels increase.
# Most noticeable in the highest grazing level.
# Within a grazing level, abundance seems to increase with the log-patch area.
# It is unclear from this if the slope of the log-area effect is
# different between grazing levels
```

6. Fit an appropriate linear model in R to explain the variation in the response variable, `ABUND` with the explanatory variables `LOGAREA` and `FGRAZE` acting additively. Hint: combine explanatory variables using a `+`

symbol. Remember to use the `data =` argument. Assign this linear model to an appropriately named object, like `birds.add.1`.

```
birds.add.1 <- lm(ABUND ~ LOGAREA + FGRAZE, data = loyn)
```

7. Produce the ANOVA table using the `anova()` function on the model object. What null hypotheses are being tested? Hint: the `anova()` function performs sequential tests. Do you reject or fail to reject the null hypotheses?

```
anova(birds.add.1)
## Analysis of Variance Table
##
## Response: ABUND
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## LOGAREA    1 3471.0  3471.0 100.2944 1.530e-13 ***
## FGRAZE     4 1136.5   284.1   8.2101 3.598e-05 ***
## Residuals 50 1730.4    34.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# null hypothesis 1: There is no effect of LOGAREA on ABUND
# (the coefficient for LOGAREA is zero)
# null hypothesis 2: There is no effect of FGRAZE on ABUND
# (no difference between grazing levels *after* the effect of LOGAREA)

# the p values are all very small therefore reject both null hypotheses.
```

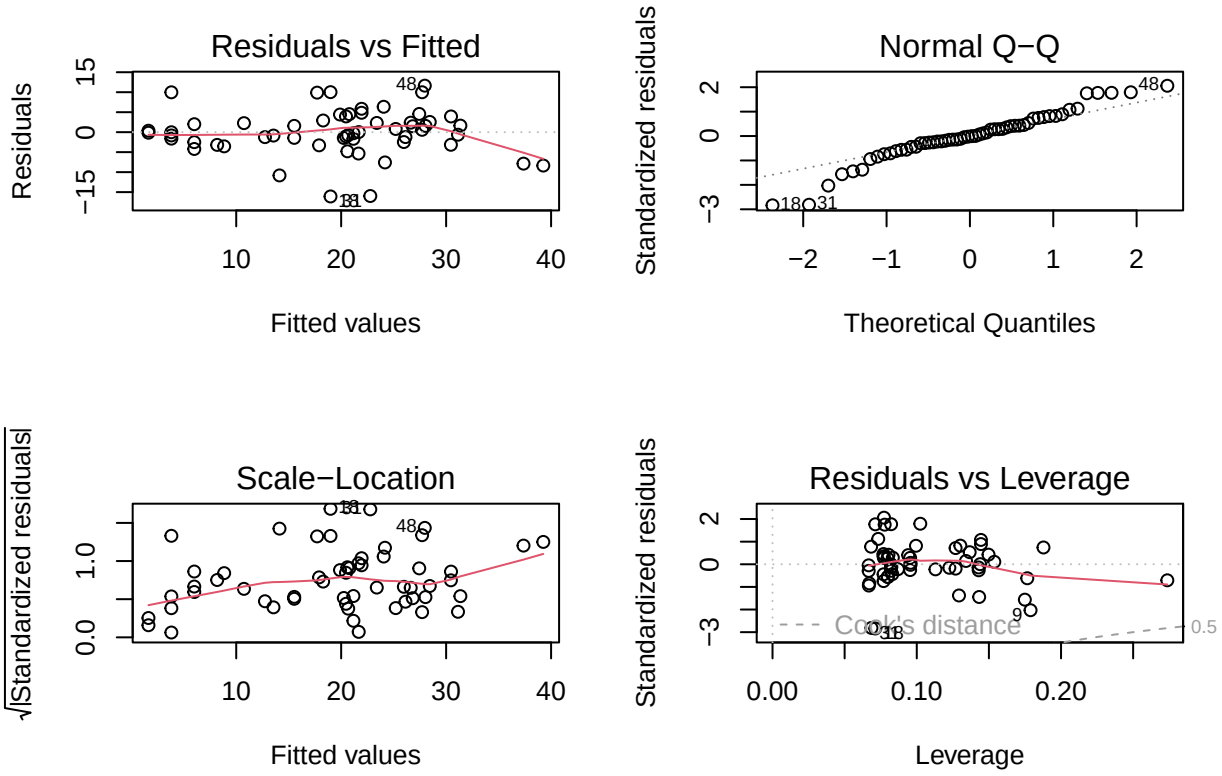
8. Use the `summary()` function on the model object to produce the table of parameter estimates. Using this output take each line in turn and answer the following questions: (A) what does this parameter represent? (B) What is the biological interpretation of the corresponding estimate? (C) What is the null hypothesis associated with it? (D) Do you reject or fail to reject this hypothesis?

```
summary(birds.add.1)
##
## Call:
## lm(formula = ABUND ~ LOGAREA + FGRAZE, data = loyn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0849  -2.4793  -0.0817   2.6486  11.6344
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.7164     2.7674   5.679 6.87e-07 ***
## LOGAREA       7.2472     1.2551   5.774 4.90e-07 ***
```

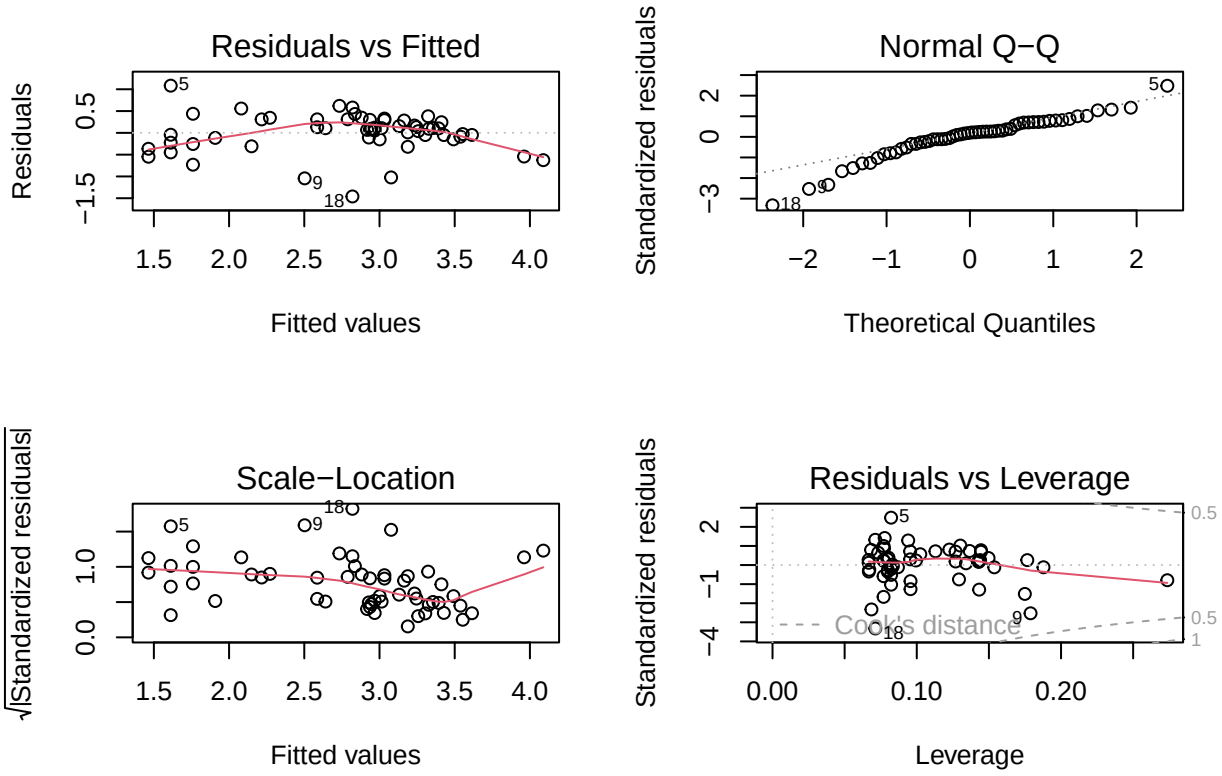
```
## FGRAZE2       0.3826     2.9123   0.131 0.895993
## FGRAZE3      -0.1893     2.5498  -0.074 0.941119
## FGRAZE4      -1.5916     2.9762  -0.535 0.595182
## FGRAZE5     -11.8938     2.9311  -4.058 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.883 on 50 degrees of freedom
## Multiple R-squared:  0.727,  Adjusted R-squared:  0.6997
## F-statistic: 26.63 on 5 and 50 DF,  p-value: 5.148e-13

# Here the intercept (baseline) is *NOT* the mean abundance of birds for
# FGRAZE level 1. It is the predicted `ABUND` for LOGAREA = 0 & FGRAZE level 1.
# the null hypothesis for the intercept is that the intercept = 0.
# As the p value (p < 2e-16) is very small we reject this null hypothesis
# and conclude that the intercept is significantly different from 0.
# Not a biologically relevant hypothesis test, in this context
# (in fact totally arbitrary, as the location of the zero is determined
# by the transformation we chose)

# For LOGAREA, the null hypothesis is that the slope of the relationship
# between LOGAREA and ABUND = 0 (no relationship)

# The remaining estimates are differences (contrasts) between each level
# and the reference level, FGRAZE1.
# For example the FGRAZE2 estimate is 0.38, so there are 0.38 more birds on
# average in graze level 2 compared to graze level 1, *for a given patch area*.
# This difference is however not significantly different from zero (p = 0.89).

# The difference between graze level 5 (FGRAZE5) and the reference FGRAZE1 is
# -11.89 (11.89 fewer birds in graze 5 compared to graze 1),
# for *an identical patch area*.
# This difference is significantly different from 0 (p = 0.00017) and therefore
# the mean abundance of birds in graze level 5 is significantly lower than in
# graze level 1, for the same patch area.

# The Multiple R-square value is as we calculated from the anova table
```

9. Now let's check the assumptions of your linear model by creating plots of the residuals from the model. Remember, that you can split your plotting device into 2 rows and 2 columns using the `par()` function before you create the plots. Check each of the assumptions using these plots and report whether your model meets these assumptions.

```
# first split the plotting device into 2 rows and 2 columns
par(mfrow = c(2,2))

# now create the residuals plots
plot(birds.add.1)
```




```
# To test the normality of residuals assumption we use the Normal Q-Q plot.
# The central residuals are not too far from the Q-Q line but the extremes
# are too extreme (the tails of the distribution are too long). Some
# observations, both high and low, are poorly explained by the model.

# The plot of the residuals against the fitted values suggests these
# extreme residuals happen for intermediate fitted values.

# Looking at the homogeneity of variance assumption (Residuals vs
# Fitted and Scale-Location plot),
# the graphs are mostly messy, with no clear pattern emerging. There is
# a hint of smaller variance with the lowest fitted values, which is not ideal.
# This could mean that the homogeneity of variance assumption is not met
# (i.e. the variances are not the same).

# The observations with the highest leverage don't appear to be overly
# influential, according to the Cook's distances in the Residuals vs
# Leverage plot.

# ABUND being bounded by zero, it wouldn't be too surprising that the variance increases with the mean
# This is often improved by log-transforming the response
loyn$logABUND<- log(loyn$ABUND + 1) # here the natural log
birds.add.3 <- lm(logABUND ~ LOGAREA + FGRAZE, data = loyn)
par(mfrow = c(2, 2))
plot(birds.add.3)
```



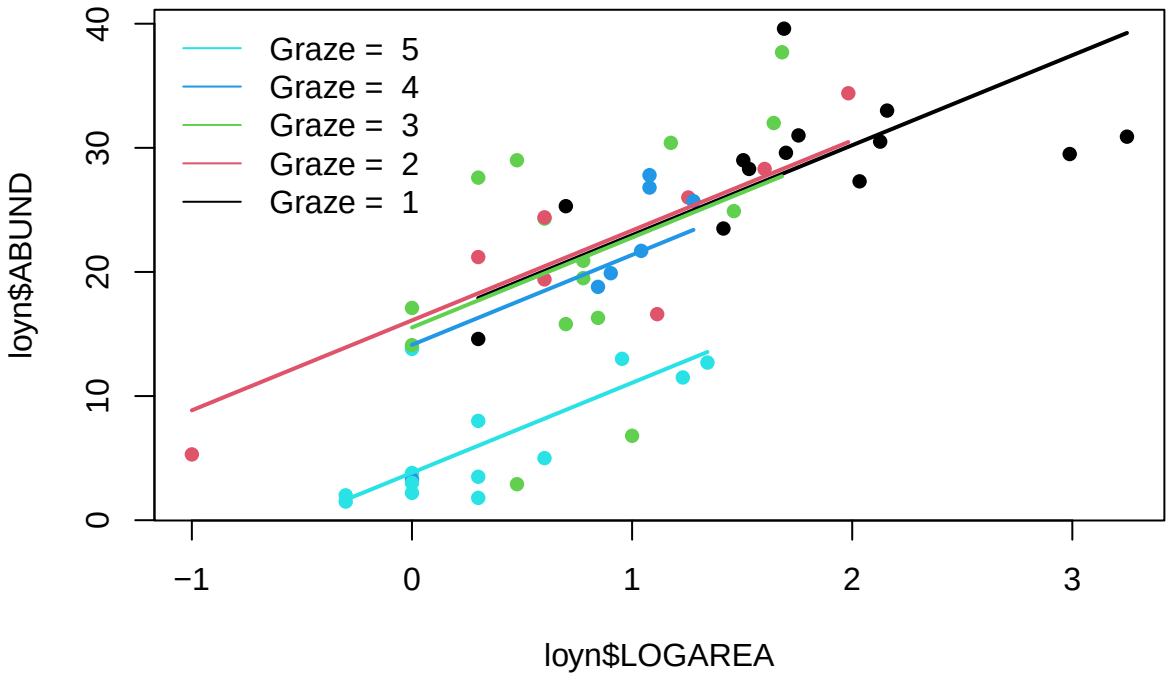

```
# Not this time! Lots of extreme negative residuals generated.

# Back to `birds.add.1` the other issue was the extreme residuals.
# This could be due to missing important explanatory variables from the model, either
# new explanatory variables altogether, or interactions: is it okay to assume
# the effect of LOGAREA to be the same for all grazing levels?
```

10. Let's plot the predictions of your initial model to figure out how it really fits the data. Here's a recipe, using the `predict()` function. + plot the raw data, using a different colour per `FGRAZE` level + for each `FGRAZE` level in turn, + create a sequence of `LOGAREA` from the minimum value to the maximum within the grazing level (unless you wish to predict outside the range of observed values) + store it in a data frame (e.g. `dat4pred`) containing the variables `FGRAZE` and `LOGAREA`. Remember that `FGRAZE` is a factor, so it requires quotes. + add a predicted column containing the explanatory variables from the model for the new data frame, using `predict()` + plot the predictions with the appropriate colours

See the script below, for one of many ways of doing this.

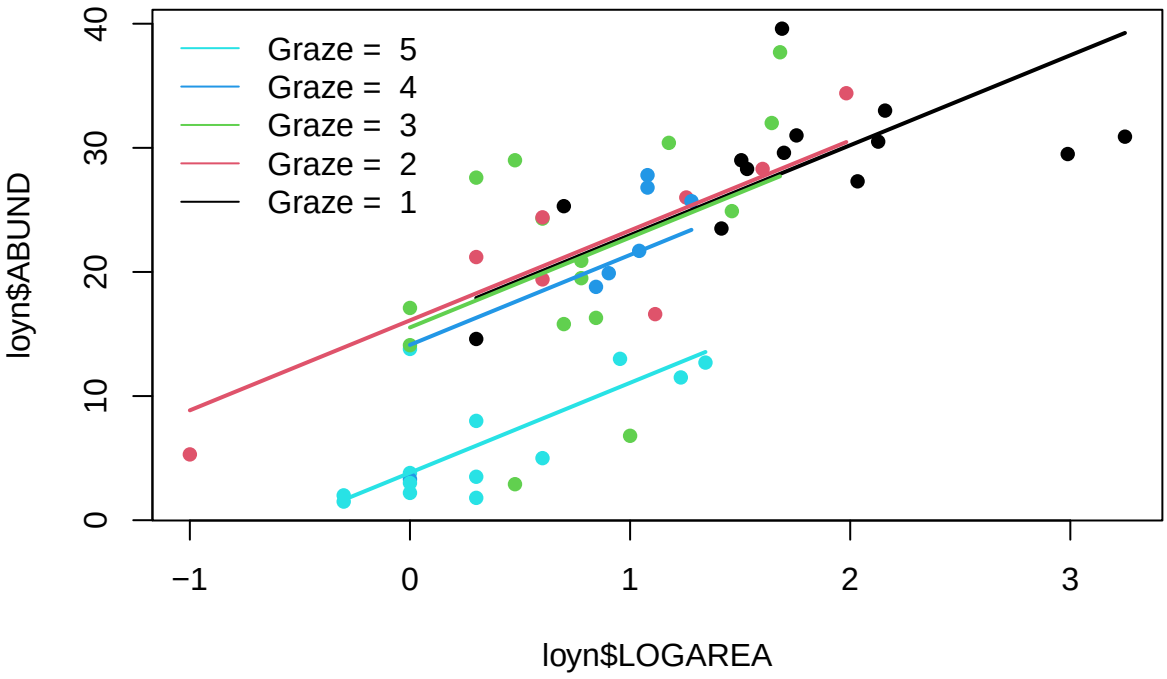
```
par(mfrow= c(1, 1))
plot(loyn$ABUND ~ loyn$LOGAREA, col = loyn$GRAZE, pch = 16)
# Note: # colour 1 means black in R
# colour 2 means red in R
# colour 3 means green in R
# colour 4 means blue in R
# colour 5 means cyan in R
```

```r
# FGRAZE1
# create a sequence of increasing Biomass within the observed range
LOGAREA.seq <- seq(from = min(loyn$LOGAREA[loyn$FGRAZE == 1]),
                   to = max(loyn$LOGAREA[loyn$FGRAZE == 1]),
                   length = 20)
# create data frame for prediction
dat4pred <- data.frame(FGRAZE= "1", LOGAREA = LOGAREA.seq)
# predict for new data
dat4pred$predicted <- predict(birds.add.1, newdata = dat4pred)
# add the predictions to the plot of the data
lines(predicted ~ LOGAREA, data = dat4pred, col = 1, lwd = 2)

# FGRAZE2
LOGAREA.seq <- seq(from = min(loyn$LOGAREA[loyn$FGRAZE == 2]),
                   to = max(loyn$LOGAREA[loyn$FGRAZE == 2]),
                   length = 20)
dat4pred <- data.frame(FGRAZE = "2", LOGAREA = LOGAREA.seq)
dat4pred$predicted <- predict(birds.add.1, newdata= dat4pred)
lines(predicted ~ LOGAREA, data = dat4pred, col = 2, lwd = 2)

# FGRAZE3
LOGAREA.seq <- seq(from = min(loyn$LOGAREA[loyn$FGRAZE == 3]),
                   to = max(loyn$LOGAREA[loyn$FGRAZE == 3]),
                   length = 20)
dat4pred <- data.frame(FGRAZE = "3", LOGAREA = LOGAREA.seq)
dat4pred$predicted <- predict(birds.add.1, newdata = dat4pred)
lines(predicted ~ LOGAREA, data = dat4pred, col = 3, lwd = 2)

# FGRAZE4
LOGAREA.seq <- seq(from = min(loyn$LOGAREA[loyn$FGRAZE == 4]),
                   to = max(loyn$LOGAREA[loyn$FGRAZE == 4]),
                   length = 20)
dat4pred <- data.frame(FGRAZE = "4", LOGAREA = LOGAREA.seq)
dat4pred$predicted <- predict(birds.add.1, newdata= dat4pred)
lines(predicted ~ LOGAREA, data = dat4pred, col = 4, lwd = 2)

# FGRAZE5
LOGAREA.seq <- seq(from = min(loyn$LOGAREA[loyn$FGRAZE == 5]),
                   to = max(loyn$LOGAREA[loyn$FGRAZE == 5]),
                   length = 20)
dat4pred <- data.frame(FGRAZE = "5", LOGAREA = LOGAREA.seq)
dat4pred$predicted <- predict(birds.add.1, newdata = dat4pred)
lines(predicted ~ LOGAREA, data = dat4pred, col = 5, lwd = 2)

legend("topleft",
 legend= paste("Graze = ", 5:1),
 col = c(5:1), bty = "n",
 lty = c(1, 1, 1),
 lwd = c(1, 1, 1))
```



[Optional] Alternative method:

```
# Okay, that was a long-winded way of doing this.
# If, like me, you prefer more compact code and less risks of errors,
# you can use a loop, to save repeating the sequence 5 times:
par(mfrow = c(1, 1))
plot(loyn$ABUND ~ loyn$LOGAREA, col = loyn$GRAZE, pch = 16)

for(g in levels(loyn$FGRAZE)){# `g` will take the values "1", "2",..., "5" in turn
    LOGAREA.seq <- seq(from = min(loyn$LOGAREA[loyn$FGRAZE == g]),
                                        to = max(loyn$LOGAREA[loyn$FGRAZE == g]),
                                                    length = 20)
    dat4pred <- data.frame(FGRAZE= g, LOGAREA= LOGAREA.seq)
    dat4pred$predicted <- predict(birds.add.1, newdata= dat4pred)
    lines(predicted ~ LOGAREA, data = dat4pred, col = as.numeric(g), lwd = 2)
}
legend("topleft",
 legend = paste("Graze = ", 5:1),
 col = c(5:1), bty = "n",
 lty = c(1, 1, 1),
 lwd = c(1, 1, 1))
```



11. What have we learned from this analysis so far?

```
# There is a significant effect of grazing levels, especially the highest
# level with a negative effect on bird abundance

# There is a significant positive effect of patch area, too.

# The relative importance of patch area and grazing is not clear, as these
# are correlated, with smaller patches also having higher grazing intensity
# on average, and larger patches lower grazing intensity.

# Some observations are poorly predicted (fitted) using the current set
# of explanatory variables.
```

Part 2: Linear model with interactive explanatory variables

Let's now explore this question left unanswered: "is it okay to assume the effect of `LOGAREA` to be the same for all grazing levels?" This is effectively asking if we should let the slope of `LOGAREA` vary across `FGRAZE` levels, which is the definition of an interactive effect.

12. Fit the corresponding linear model in R to explain the variation in the response variable, `ABUND` with the explanatory variables `LOGAREA` and `FGRAZE` and the interaction between these variables. Hint: this time, the interaction is included using the `*` symbol instead of `+`. Remember to use the `data =` argument. Assign this linear model to an appropriately named object, like `birds.inter.1`.

```r
birds.inter.1 <- lm(ABUND ~ FGRAZE * LOGAREA , data = loyn)
```

13. Produce the ANOVA table using the `anova()` function on the model object. What null hypotheses are being tested? Hint: the `anova()` function performs sequential tests. Which of these hypotheses are relevant to us? Do you reject or fail to reject the null hypotheses?

```r
anova(birds.inter.1)
## Analysis of Variance Table
##
## Response: ABUND
##                Df Sum Sq Mean Sq F value    Pr(>F)
## FGRAZE          4 3453.7  863.42 26.8974 1.550e-11 ***
## LOGAREA         1 1153.8 1153.85 35.9447 2.942e-07 ***
## FGRAZE:LOGAREA  4  253.8   63.44  1.9764    0.1139
## Residuals      46 1476.6   32.10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# null hypothesis 1: There is no effect of LOGAREA on ABUND
# (the coefficient for LOGAREA is zero)

# null hypothesis 2: There is no effect of FGRAZE on ABUND
# (no difference between grazing levels *after* the effect of LOGAREA)

# null hypothesis 3: There is no effect of an FGRAZE by ABUND interaction
# *after* the effects of LOGAREA and FGRAZE combined).
# A couple of equivalent ways to say this: the effect of LOGAREA doesn't differ
# among FGRAZE levels or: the difference in bird abundance between grazing
# levels is the same for all patch areas.

# As long as there is an interaction in the model, the null hypotheses 1 and 2
# ("main effects") are not relevant to us
# the p value for the interaction is large, therefore we fail to reject the
# null hypothesis: there is no evidence supporting this interaction.
```

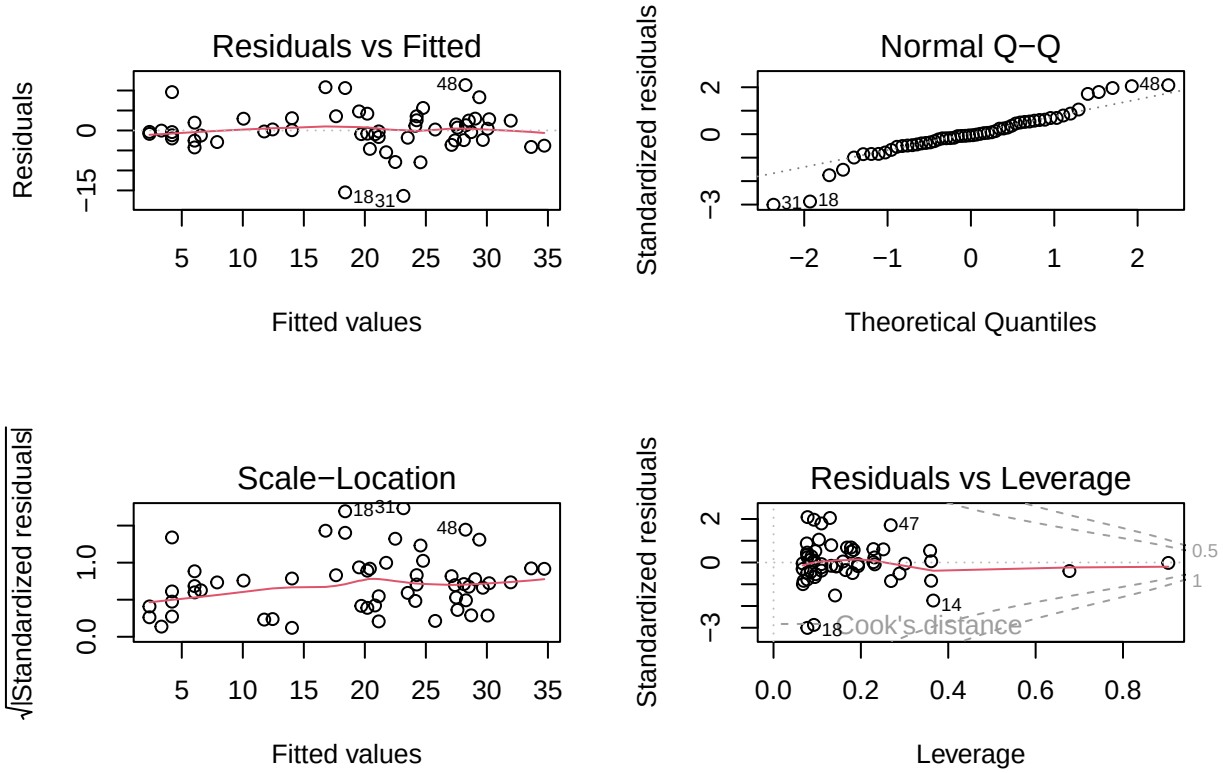
14. Use the `summary()` function on the model object to produce the table of parameter estimates. Using this output take each line in turn and answer the following questions: (A) what does this parameter represent, mathematically? (B) What is the biological interpretation of the corresponding estimate? (C) What is the null hypothesis associated with it? (D) Do you reject or fail to reject this hypothesis?

```r
summary(birds.inter.1)
##
## Call:
## lm(formula = ABUND ~ FGRAZE * LOGAREA, data = loyn)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3615  -2.3807  -0.2449   2.6181  11.3529
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       21.243      3.987   5.329 2.89e-06 ***
## FGRAZE2           -6.165      4.842  -1.273 0.209285
## FGRAZE3           -7.215      4.820  -1.497 0.141271
## FGRAZE4          -17.910      6.701  -2.673 0.010375 *
## FGRAZE5          -17.043      4.406  -3.868 0.000344 ***
## LOGAREA            4.144      2.057   2.014 0.049843 *
## FGRAZE2:LOGAREA    4.368      3.108   1.405 0.166593
## FGRAZE3:LOGAREA    4.989      3.531   1.413 0.164374
## FGRAZE4:LOGAREA   15.235      5.925   2.572 0.013422 *
## FGRAZE5:LOGAREA    1.996      3.650   0.547 0.587148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.666 on 46 degrees of freedom
## Multiple R-squared:  0.767,  Adjusted R-squared:  0.7214
## F-statistic: 16.83 on 9 and 46 DF,  p-value: 8.05e-12

# Here the intercept (baseline) is the predicted `ABUND` for LOGAREA = 0,
# for FGRAZE level 1.
# the null hypothesis for the intercept is that the intercept = 0
# (not biologically relevant).

# LOGAREA represents the slope for LOGAREA, specific to level FGRAZE = 1.
# The null hypothesis is that the slope of the relationship
# between LOGAREA and ABUND = 0, for level FGRAZE = 1 only.

# FGRAZE2...5 estimate differences (contrasts) between the *intercept* of
# each level and the *intercept* of the reference level, FGRAZE = 1.

# FGRAZE2...5:LOGAREA estimate differences (contrasts) between the *slope*
# of LOGAREA for each level and the *slope* of LOGAREA for the reference
# level, FGRAZE = 1.


# The Multiple R-square value is 0.76, slightly up from the purely additive
# model (but not much, given that we have added a whole 4 parameters to the
# model, i.e. nearly doubled its complexity)
```

15. Now let's check the assumptions of your linear model by creating plots of the residuals from the model. Remember, that you can split your plotting device into 2 rows and 2 columns using the `par()` function before you create the plots. Check each of the assumptions using these plots and report whether your model meets these assumptions.

```r
# first split the plotting device into 2 rows and 2 columns
par(mfrow = c(2,2))

# now create the residuals plots
plot(birds.inter.1)
```

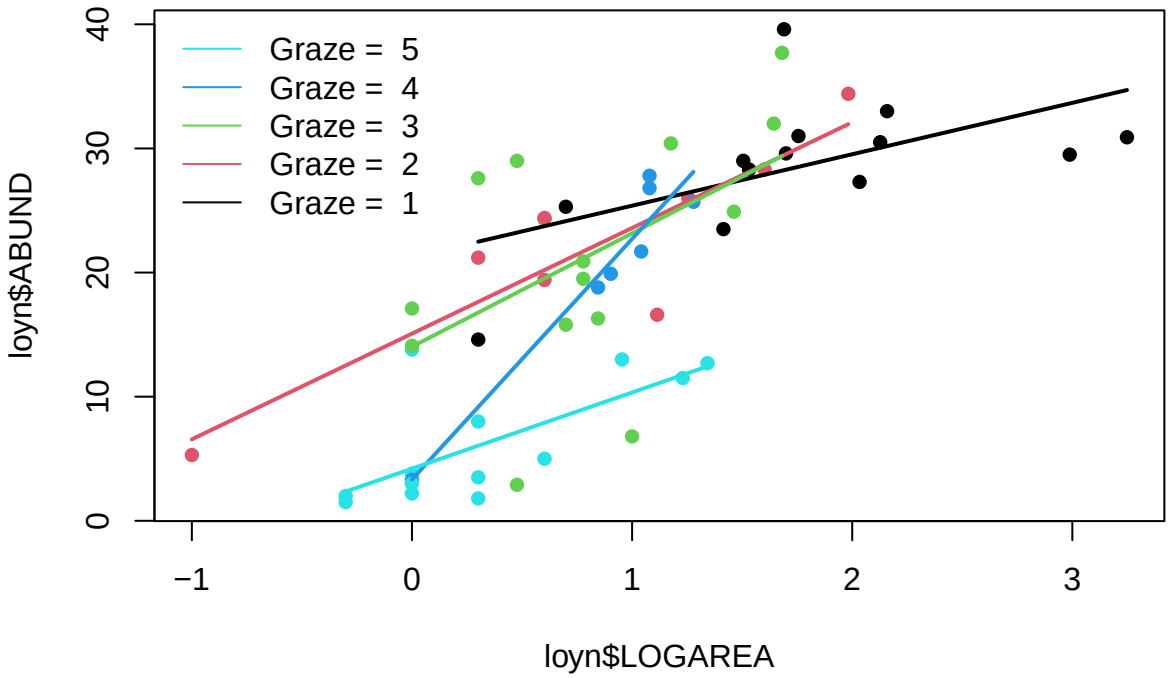



```r
# Not a great deal of an improvement! Just marginally better in every respect,
# thanks to increasing the fit slightly (by throwing lots of new model
# parameters at the data).
```

16. Let's now plot the predictions of the interactive model to figure out how it really fits the data. Hint: the script from question 11 should work all the same, since the predictor variables involved in the equation are the same. Don't forget to update the name of the model!

```r
# NOTE: I'm using the loop version of the plot, here.
# If you don't like it, refer to the long-hand code version at Question 11

par(mfrow = c(1, 1))
plot(loyn$ABUND ~ loyn$LOGAREA, col = loyn$GRAZE, pch = 16)

for(g in levels(loyn$FGRAZE)){# `g` will take the values "1", "2",..., "5" in turn
    LOGAREA.seq <- seq(from = min(loyn$LOGAREA[loyn$FGRAZE == g]),
                                        to = max(loyn$LOGAREA[loyn$FGRAZE == g]),
                                                    length = 20)
```

```
    dat4pred <- data.frame(FGRAZE = g, LOGAREA = LOGAREA.seq)
    dat4pred$predicted <- predict(birds.inter.1, newdata = dat4pred)
    lines(predicted ~ LOGAREA, data = dat4pred, col = as.numeric(g), lwd = 2)
}
legend("topleft",
 legend = paste("Graze = ", 5:1),
 col = c(5:1), bty = "n",
 lty = c(1, 1, 1),
 lwd = c(1, 1, 1))
```



17. Do you think the model is biologically plausible? Is it supported statistically?

```
# The slopes of the LOGAREA effect across grazing levels are all over the
# place, without any coherent pattern (for instance, they could have been
# increasing or decreasing gradually from low to high grazing intensity)

# The interaction is non-significant, so isn't supported statistically either.

# Time to revert to the simpler, or a different model? Decisions, decisions!
```

End of the Linear model with continuous and categorical explanatory variables exercise