Multiple Linear Regression & Model Selection

Written by Rafael Romero

2024-01-02

Contents

Multiple Linear Regression & Model Selection

This lab will apply data visualization skills (Lab 3), hypothesis testing (Lab 4 & 5), knowledge about residuals & how to interpret diagnostic plots (Lab 6), how to run correlation or linear regression models (Lab 7), and how to build multiple regression models and perform model selection (Lab 8).

Lab Questions

Question 1

Please answer **True or False** for 1a-c. a) If you build a linear regression model with 'mod1 <- `lm(y~x)` and then run the code `plot(mod1)`, you get four diagnostic plots that are showing the results of your model residuals **TRUE**

b) generally, you want to choose a model with the highest AIC value **FALSE**

c) When testing the assumptions of your linear regression model you are testing the normality & equality of variance assumptions of your residuals **TRUE**

Question 2

What is the difference between AIC and BIC?

ANSWER: AIC is used for prediction focused models and BIC is used for identify the true model in a dataset.

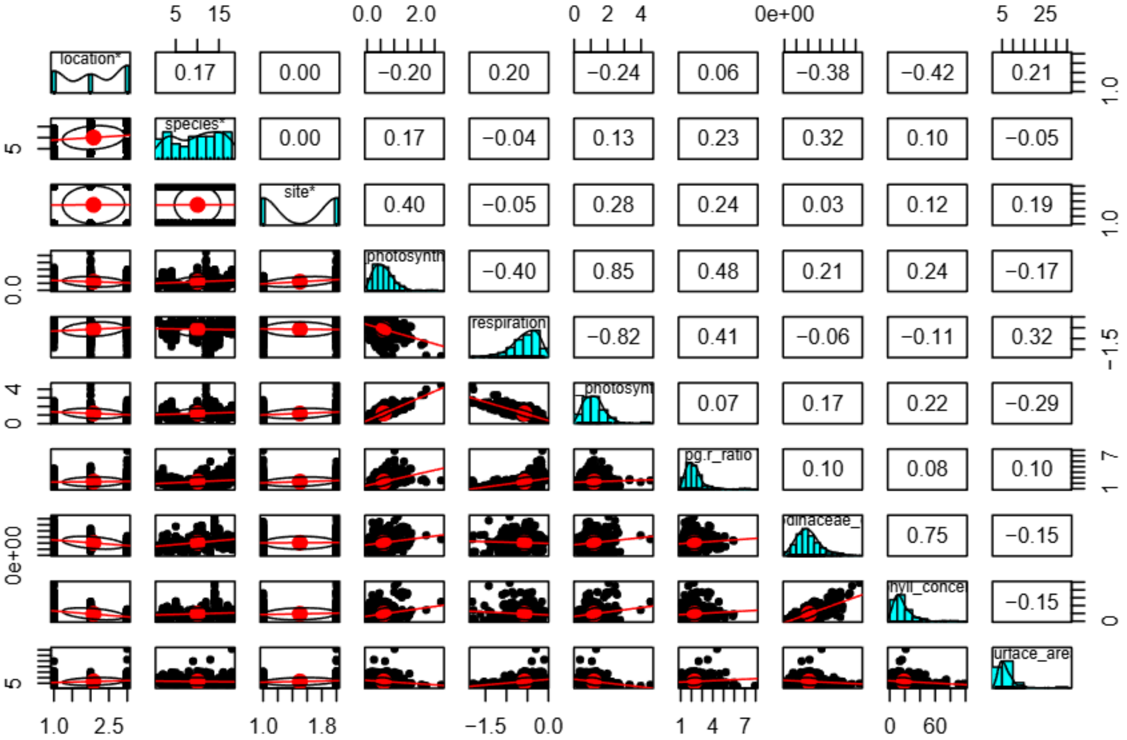
Question 3

For Question 3, we will be looking at the data set `coral` which is a subset of data from Biscéré et al 2019 (Royal Society Biology Letters). Ocean acidification (OA) is known to impact coral reefs calcification and metabolic processess (ie photosynthesis and respiration). The authors present data collected between 2016-2018 at hree natural C02 seeps in Papua New Guinea where they measured metabolic flexibility (ie in hospite photosynthesis and dark respiration). More details regarding the data collection can be found in the paper (doi: 10.1098/rsbl.2018.0777)

For this lab, we will focus on our outcome of interest `respiration` and try to build the best model to predict coral `respiration` with the variables in the `coral` data set.

a) Check collinearity between potential Xs with a pairs.panel() figure

```
pairs.panels(coral,
             density = TRUE,
             cor = TRUE,
             lm = TRUE) #Looking for potential Xs, respiration & gross_photosynthesis (-0.82), symbiodi
```

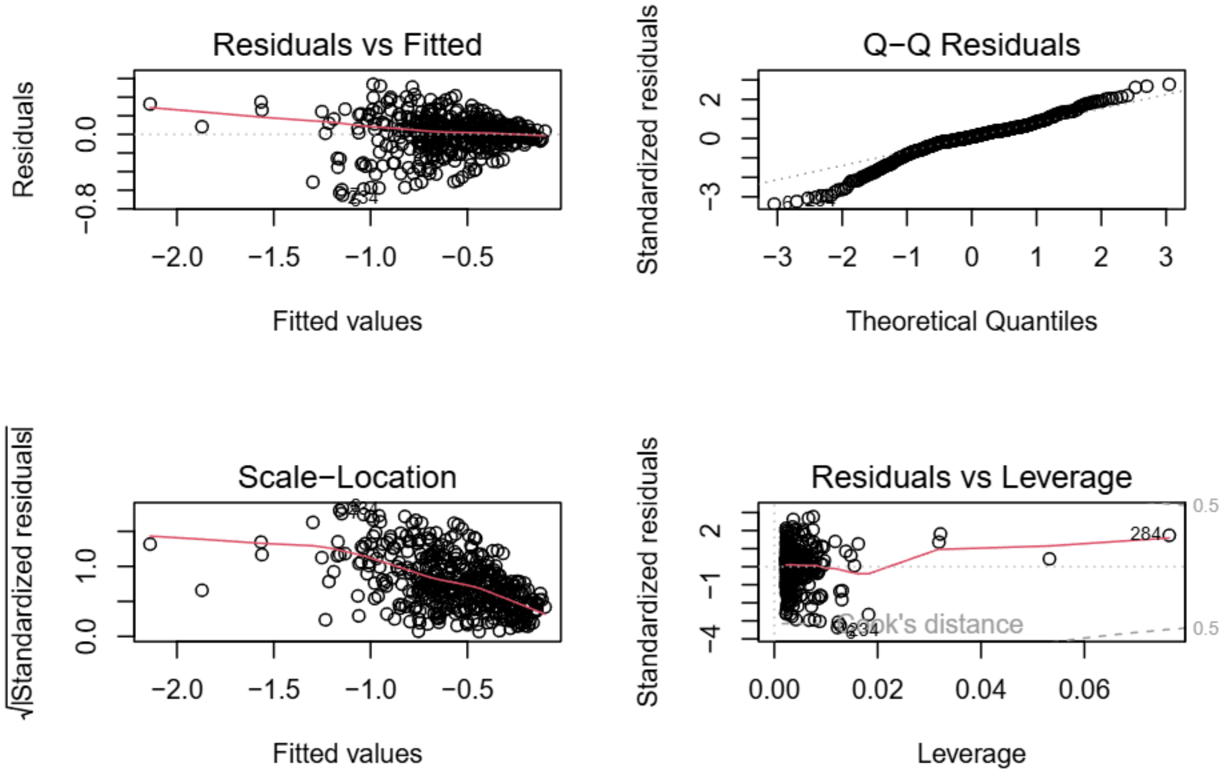


b) Build three separate regression models

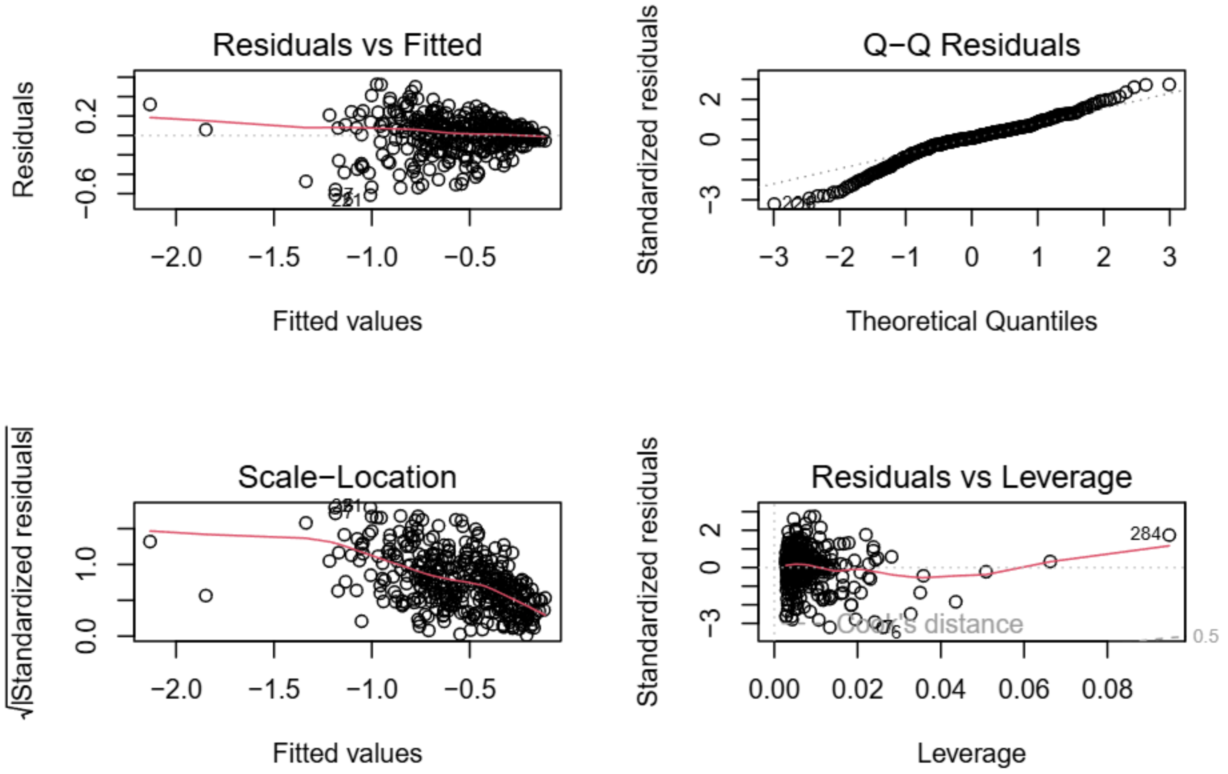
```
fit_1var <- lm(respiration ~ gross_photosynthesis, data = coral)
fit_2var <- lm(respiration ~ gross_photosynthesis+symbiodinaceae_density, data = coral)
fit_3var <- lm(respiration ~ gross_photosynthesis+symbiodinaceae_density+chlorophyll_concentration , da
```

c) Create model diagnostic plots for each model you created in step b. What can you say about the normality and equal variance assumptions? Do you need to transform any of your variables? Don't worry if not all of your candidate models pass all of the assumptions. What is important for this exercise is that you try to accomplish meeting the assumptions for the multiple regressions and interpret the output.

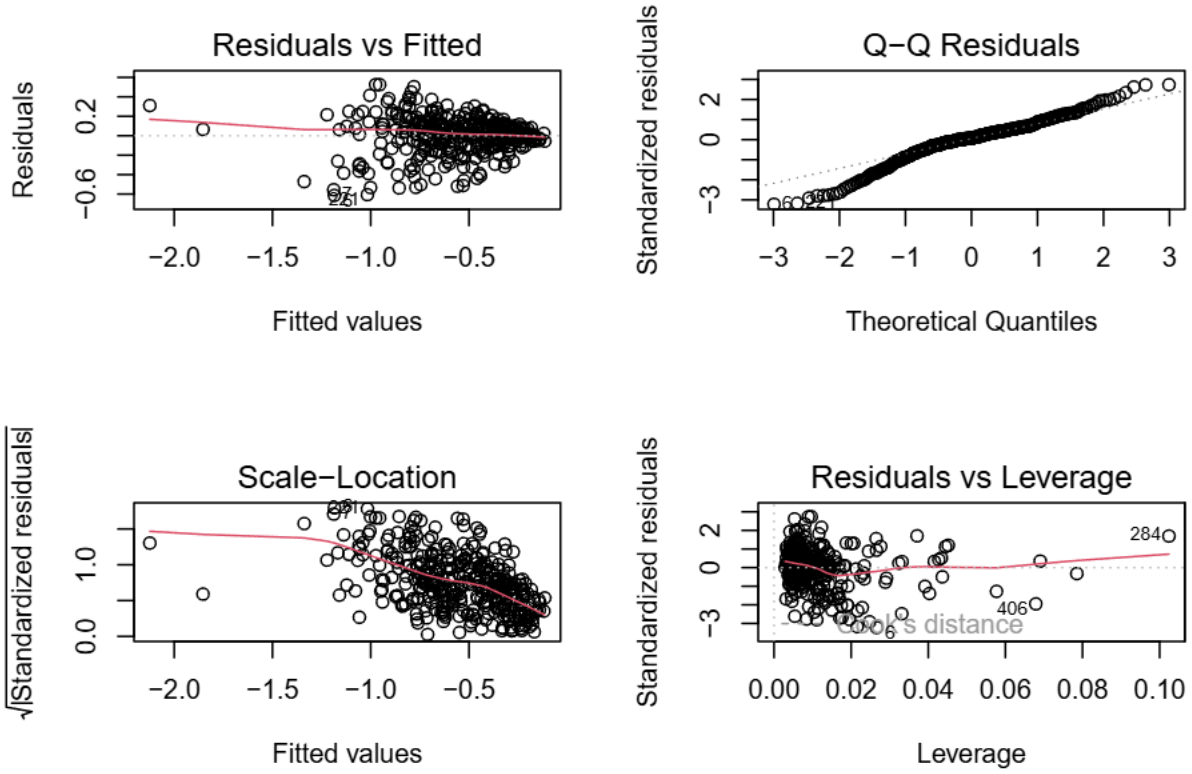
```
# 1 var
par(mfrow = c(2,2))
plot(fit_1var)
```



```r
# 2 var
par(mfrow = c(2,2))
plot(fit_2var)
```



```
# 3 var
par(mfrow = c(2,2))
plot(fit_3var)
```



Conclusion: The plots showcase a pattern of where the residuals lie vs the fitted as they are heavily concentrated around the higher fitted values. The Q-Q Residuals plot also does not follow the line very close which can indicate non-normality for the residuals. This means transformations may be needed in order to meet the linear regression model more effectively.

d) Calculate the AIC of each model and save as the variable 'results'

```
results <- AIC(fit_1var,fit_2var,fit_3var)
```

e) Using the list() make a variable called 'models' with your 3 regression models you built

```
models <- list(fit_1var,fit_2var,fit_3var)
```

f) Create a column on 'results' for BIC

```
results$BIC <- sapply(models, BIC)
```

g) Using the lappy() perform a summary for each of your models

```
model_summary <- lapply(models, summary)
```

h) Create a forloop to extract relevant information from model summaries

```
for(i in 1:length(models)){ #this creates a variable i that starts with the value i=1
  results$rsq[i] <- model_summary[[i]]$r.squared #we assign the rsq value from model i to the i'th row
  results$adj_rsq[i] <- model_summary[[i]]$adj.r.squared #same for adjusted rsq
} #now we go back to the beginning of the for-loop, add 1 to the value of i, and do everything again
```

i) Create a nice kable with model results.

Table 1: Model Results

	df	AIC	BIC	rsq	adj_rsq
fit_1var	3	-183.61	-171.38	0.67	0.67
fit_2var	4	-157.94	-142.39	0.67	0.67
fit_3var	5	-156.22	-136.79	0.67	0.67

```r
results %>%
  kbl(digits =2, align="c",caption = "Model Results") %>%
  kable_styling(bootstrap_options = c("striped", "hover")) %>%
  row_spec(0, bold = TRUE, background = "#f2f2f2")
```

j) Of your 3 models, which do you think is the best fit model? Must include AIC, BIC, R^2 in your answer.

ANSWER: 'fit_3var' is the best model since it has the lowest AIC and BIC (-156.22,-136.79), and highest R squared since they are all 0.67.

k) With your best fit model from j, separate your data into training and testing sets

- *Hint: Look at Exercise 3 Step 9*

```r
splitter <- sample(1:nrow(coral),15,replace = F)
coral_train <- coral[-splitter,]
coral_test <- coral[splitter,]
```

l) Use the best fit model on your TRAINING data set

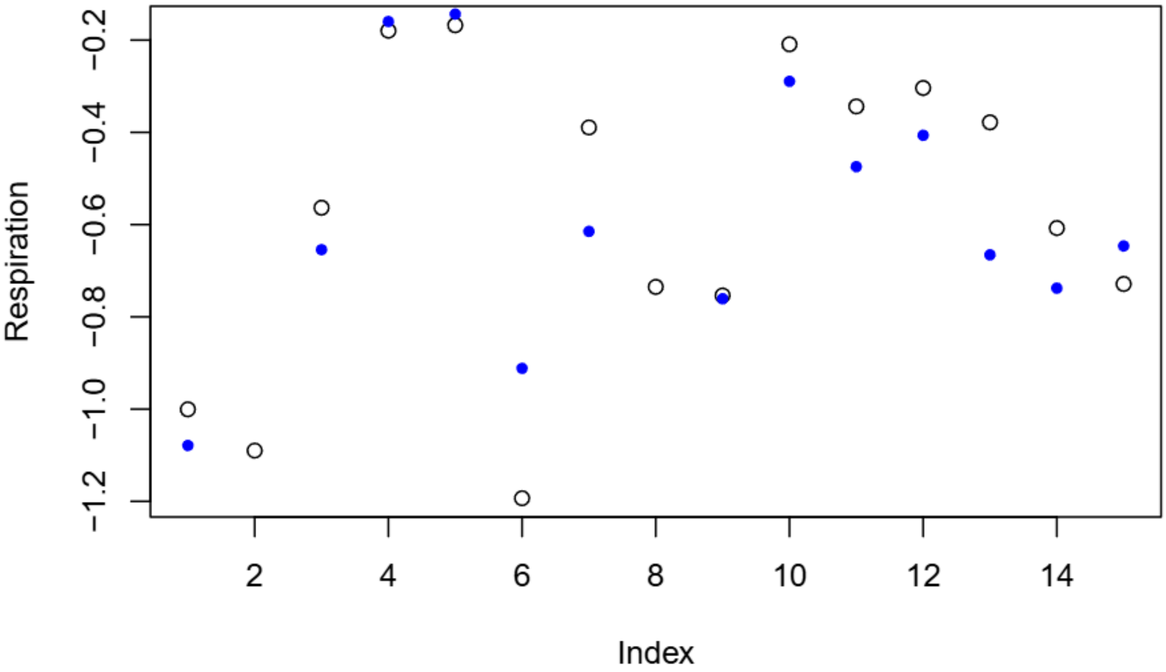
```r
fit_3var_train <- lm(respiration ~ gross_photosynthesis + symbiodinaceae_density + chlorophyll_concentra
                     data = coral_train)
```

m) Use the fitted model with the training data to make predictions of your Y outcome

```r
fit_3var_train_predict <- predict(fit_3var_train,coral_test)
```

n) Visualize how your best fit model did with predictions

```r
plot(coral_test$respiration, pch = 1, ylab = "Respiration") # plot actual test data values
points(fit_3var_train_predict, pch = 20, col = "blue") # plot the model predictions for those points
```



o) Comment on how your best fit model performed at predicting `respiration`your model prediction. What would be some ways to improve your model performance if you had unlimited money/ethics to collect all the data you wanted?

CONCLUSION: We notice that there is not much overlap between the blue and open cirles which means we did not have a perfect model. With more time and money I would increase the sample in order to improve accuracy. I also would have transformed the variables since the residuals were not normally distributed and they also were not randomly scattered.

End of Lab 8