Self-assess: Statistical Learning

Dr. Brenda Vo

To have necessary background for the course, you should have done undergraduate courses in statistics and Mathematics including multivariable data analysis, statistical models and basic matrix and linear algebra. It is also desirable to have some familiarity with computing programming in R. The following questions will help you assess your readiness for the course.

Question

Microplastics are found in almost all marine and fresh water environments, where they pose a potential risk to fish and crustaceans. Therefore, the effects of microplastics on aquatic organisms are currently the subject of intense research. Here we have a dataset containing 200 seawater samples collected at different sites around a bay in Sweden. The dataset can be downloaded from
`https://drive.google.com/file/d/1ktKF18eAJhGYUgW9n4Emtq0Prt9gxlfI/view?usp=sharing`

There are 5 variables included in the study:

- PE = polyethylene microplastics ($\mu g/m^3$)
- PP = polypropylene microplastics ($\mu g/m^3$)
- PS = polystyrene microplastics ($\mu g/m^3$)
- temp = water temperature at each site (Celsius)
- larvae = number of fish larvae of a single species per 100 m^3

(a) Plot the data and summarise the information available from the plot.

(b) Fit a model of the form

$$larvae = \beta_0 + \beta_1 PE + \beta_2 PP + \beta_3 PS + \beta_4 temp + \epsilon$$

Print the table of regression coefficients and write down the least squares regression equation.

(c) Which variables are significant predictors of fish larvae density in this model, at a 5% level?

Now drop the non-significant terms and refit the model using only the explanatory variables that are significant. This is referred to as the final model.

(d) Print the table of regression coefficients and write down the least squares regression equation for the final model.

(e) Produce the diagnostic plots for the final model and explain what can be understood from the plots.

(f) Write a concise, informative conclusion based on your analysis and results.

Solutions

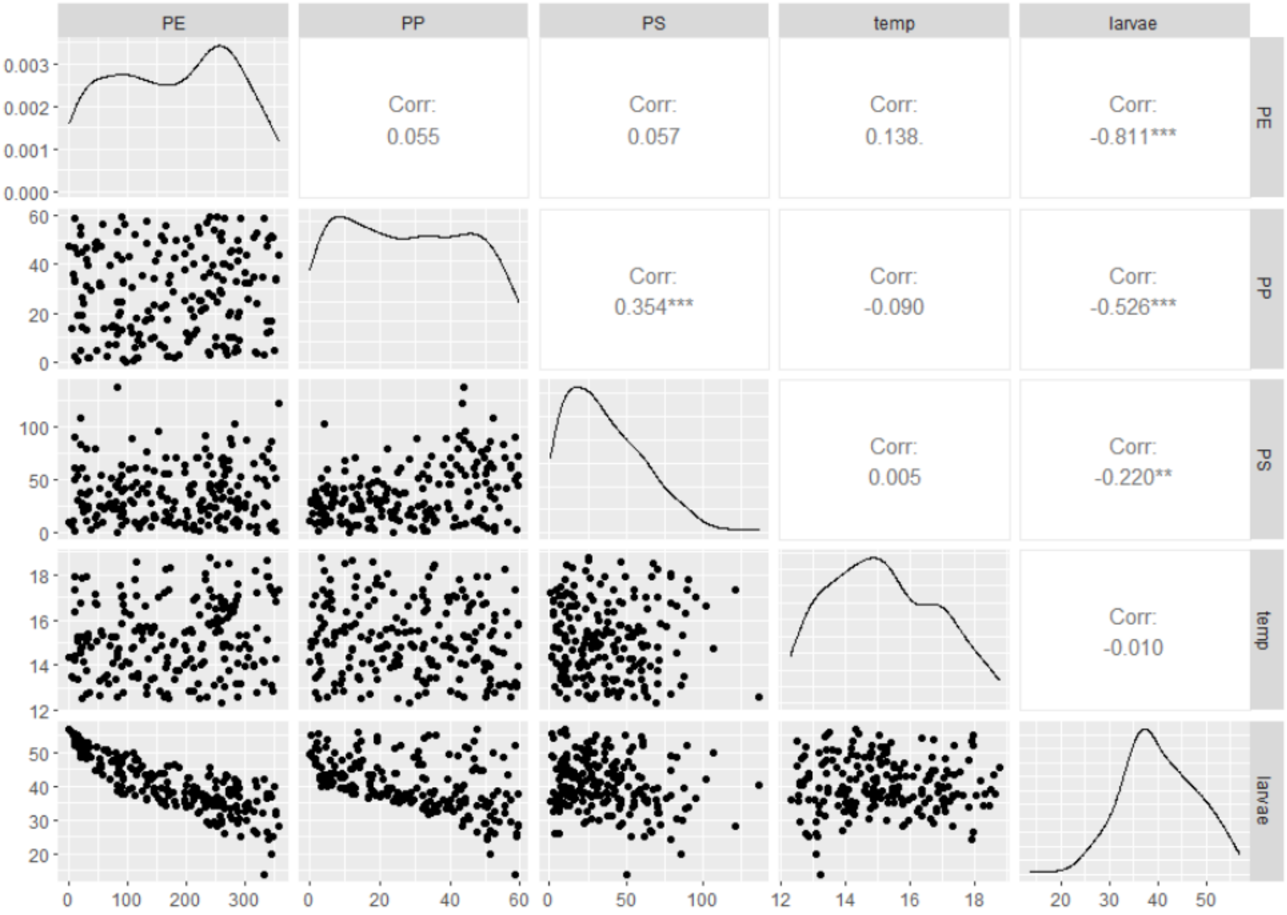

Figure 1: Pairs plot for the microplastics data

(a) From the pairs plot (Figure 1), we see that the number of larvae has a moderate to strong negative correlation with PP and PE, $r = -0.526$ and $r = -0.811$ respectively. However, there appears a weak correlation between the number of larvae with both PS and temperature ($r = -0.220$ and $r = -0.010$, respectively).

Among the 4 predictors, it's noticed that PP is moderately positively correlated with PS ($r = 0.354$). There is little or no linear relationship between other pairs of predictors.

(b) The table of regression coefficients is given in Table 1.

The least squares regression equation is:
E(larvae) = 52.550 - 0.060PE - 0.207PP - 0.002PS + 0.274temp

Table 1: Table of regression coefficients for model with all 4 predictors

```
Coefficients:
             Estimate  Std. Error   t value   Pr(>|t|)
(Intercept) 52.54973    1.75625     29.92     <2e-16
PE          -0.05982    0.00179    -33.44     <2e-16
PP          -0.20693    0.01099    -18.83     <2e-16
PS          -0.00237    0.00746    -0.32      0.751
temp         0.27380    0.11304     2.42      0.016

Residual standard error: 2.57 on 195 degrees of freedom
Multiple R-squared:  0.893,Adjusted R-squared:  0.891
F-statistic:  408 on 4 and 195 DF,  p-value: <2e-16
```

(c) We'll use the t statistics and p-values in Table 1 to test the significance of individual partial regression coefficients.

Only PS is the non-significant predictor ($p = 0.751 > 0.05$), and it can be removed from the model. All three predictors PE, PP and temp are significantly useful to predict the number of larvae and so they should be retained.

(d) Results from the final model with only three predictors: PE, PP and temp are given in Tables 2.

The final model is:

$$E(larvae) \quad = \quad 52.52 - 0.06PE - 0.21PP + 0.27temp$$

Table 2: Table of regression coefficients for the final model

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.51909    1.74958   30.02   <2e-16
PE          -0.05984    0.00178  -33.54   <2e-16
PP          -0.20817    0.01026  -20.29   <2e-16
temp         0.27260    0.11272    2.42    0.017

Residual standard error: 2.56 on 196 degrees of freedom
Multiple R-squared:  0.893,Adjusted R-squared:  0.891
F-statistic:  546 on 3 and 196 DF,  p-value: <2e-16
```

(e) The assumptions of the linear model are that the residuals are independent, normally distributed, centred around 0 and have constant variance: $\epsilon \sim N(0, \sigma^2)$. You should also check for potential outliers.

From the residuals vs fitted plot (LHS of Figure 2)it appears that the variance is not constant and two observations have very large residuals (observations 131 and 176). These observations are identified as extreme outliers in the Normal QQ plot (RHS of Figure 2), and there are a number of other observations having a std. residual > 2.

The residuals in the Normal QQ plot (RHS of Figure 2) do not follow a straight line. Even though a large number of residuals are in the diagonal line, there are clear bends/deviations in the tails. A Shapiro-Wilk's test produces a very small p-value, 9×10^{-4} (Table 3), suggesting that the normality assumption for the residuals is violated.

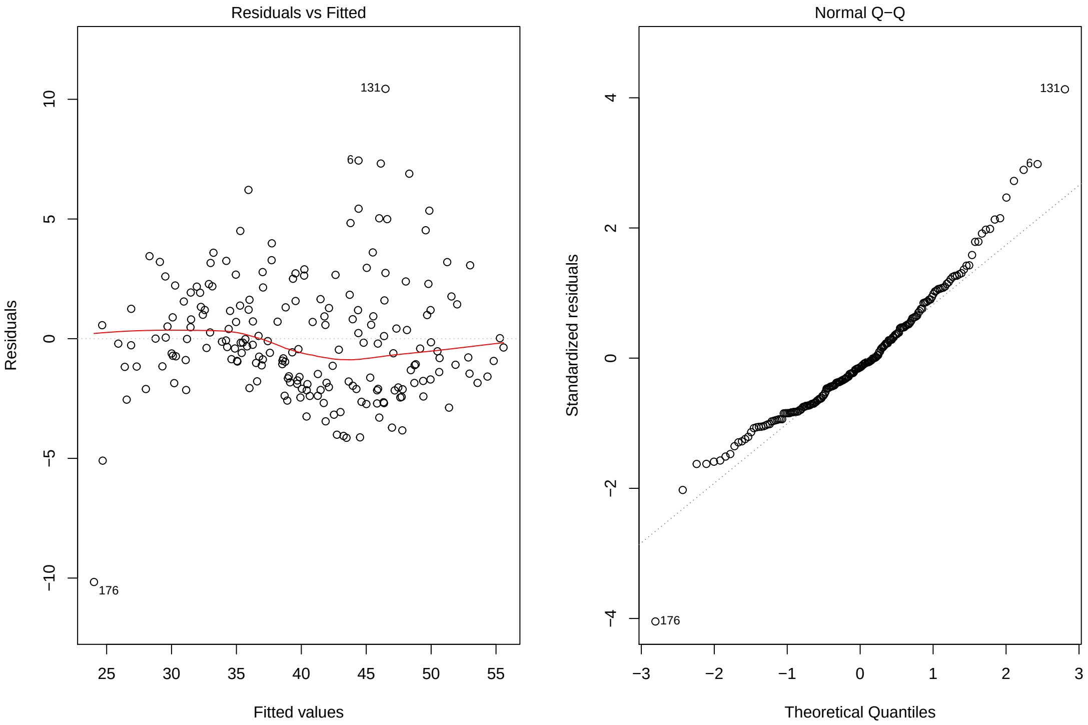

Figure 2: Residual plots

Table 3: Shapiro-Wilk normality test

```
Shapiro-Wilk normality test

data:  mod2$residuals
W = 0.96, p-value = 8e-06
```

(f) **Conclusion/Summary**

After the regression analysis, it was found that the microplastics (PE & PP) have strong negative correlation with the number of fish larvae. The amount of PE, PP and temperature are useful in predicting the number of fish larvae, whereas the polystyrene microplastics (PS) are not having significant effects in the fish larvae.

The final main effects model is:

E(larvae) = 52.52 - 0.06PE - 0.21PP + 0.27temp

The average number of fish larvae will decrease when PE increases, given the amount of PP and temp are held constant. Similarly, the higher the amount of PP the lower the average number of larvae, given the amount of PE and temp are held constant. However, for the same amount of PE and PP, the average number of larvae is higher for higher temperature. Since the model assumptions are violated, making prediction using this final model should be taken with care.

R **Code**

```
## remove all variables currently in the working environment
rm(list=ls())
options(digits=3, show.signif.stars=F)
install.packages("ggplot2")           # Packages need to be installed only once
install.packages("GGally")

library("ggplot2")                    # Load ggplot2 package
library("GGally")
## read in data and store in object named microP
microP<-read.table("Mplastics.txt", header=TRUE)
##summary of each variable
summary(microP)
## Generate a pairs plot
ggpairs(microP[, c(2:6)])
## Fit linear regression model
mod1<-lm(larvae ~ PE + PP + PS + temp, data = microP)
## table of parameter estimates, and t-tests
summary(mod1)
## Refit the model using only variables PE, PP & temp
mod2 <- lm(larvae ~ PE + PP + temp, data = microP)
summary(mod2)

## Residuals plots
par(mfrow =c(1,2))
plot(mod2, which=1:2)
shapiro.test(mod2$residuals)
```