Statistical Learning self-assessment quiz
(For the AMSI ACE network 2024)

Robert Cope

To have the 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 this link: https://www.dropbox.com/s/1vmbcg3euahjcup/Mplastics.csv?dl=0

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 coeffcients and write down the least squares regression equation.

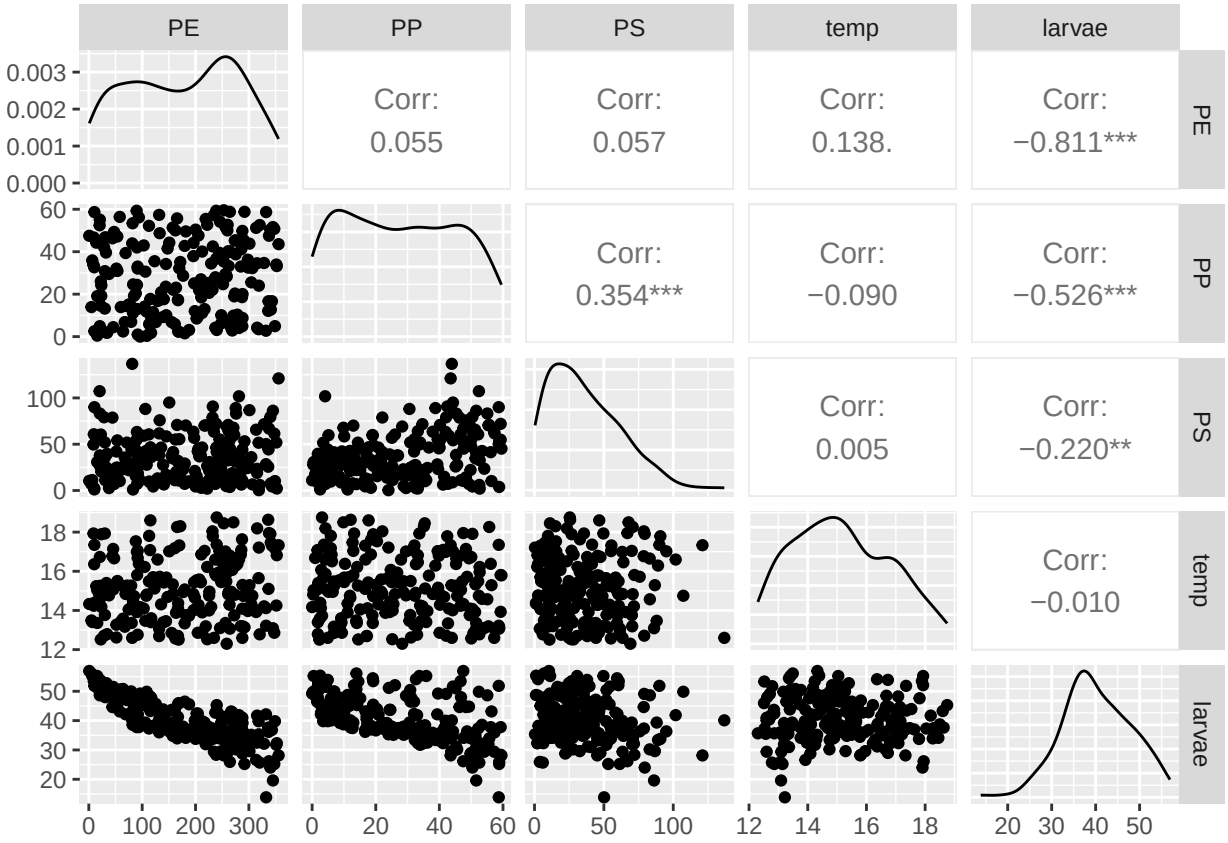
(c) Which variables are signiffcant predictors of fish larvae density in this model, at a 5% level? Now drop the non-significant terms and refit the model using only the explanatory variables that are significant. This is referred to as the final model.
(d) Print the table of regression coeffcients and 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.

Example solutions

(you can use a variety of different packages or base R tools, the choices here are one example only)

```
library(readr)
library(GGally)
library(ggResidpanel)
Mplastics <- read_csv('Mplastics.csv')
```

```
ggpairs(Mplastics, columns = 2:6) # we leave out the "sample" column
```



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.

fitting models

```
mod1<-lm(larvae ~ PE + PP + PS + temp, data = Mplastics)
summary(mod1)
```

```
##
## Call:
## lm(formula = larvae ~ PE + PP + PS + temp, data = Mplastics)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1717  -1.8030  -0.3335   1.3810  10.3551
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 52.549730   1.756252  29.922   <2e-16 ***
## PE          -0.059817   0.001789 -33.436   <2e-16 ***
## PP          -0.206932   0.010990 -18.828   <2e-16 ***
## PS          -0.002371   0.007457  -0.318   0.7509
## temp         0.273804   0.113041   2.422   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 195 degrees of freedom
## Multiple R-squared:  0.8932, Adjusted R-squared:  0.891
## F-statistic: 407.5 on 4 and 195 DF,  p-value: < 2.2e-16
```

The least squares regression equation is:

$$E(larvae) = 52.550 - 0.060PE - 0.207PP - 0.002PS + 0.274temp$$

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. (Note that this is one (simple) way to select coefficients; it is not necessarily a good way to do so.)

```
## Refit the model using only variables PE, PP & temp
mod2 <- lm(larvae ~ PE + PP + temp, data = Mplastics)
summary(mod2)
```

```
##
## Call:
## lm(formula = larvae ~ PE + PP + temp, data = Mplastics)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1650  -1.7976  -0.3581   1.3403  10.4370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 52.519088   1.749580  30.018   <2e-16 ***
## PE          -0.059837   0.001784 -33.544   <2e-16 ***
## PP          -0.208167   0.010257 -20.294   <2e-16 ***
## temp         0.272595   0.112717   2.418   0.0165 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.562 on 196 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.8915
## F-statistic: 545.9 on 3 and 196 DF,  p-value: < 2.2e-16
```

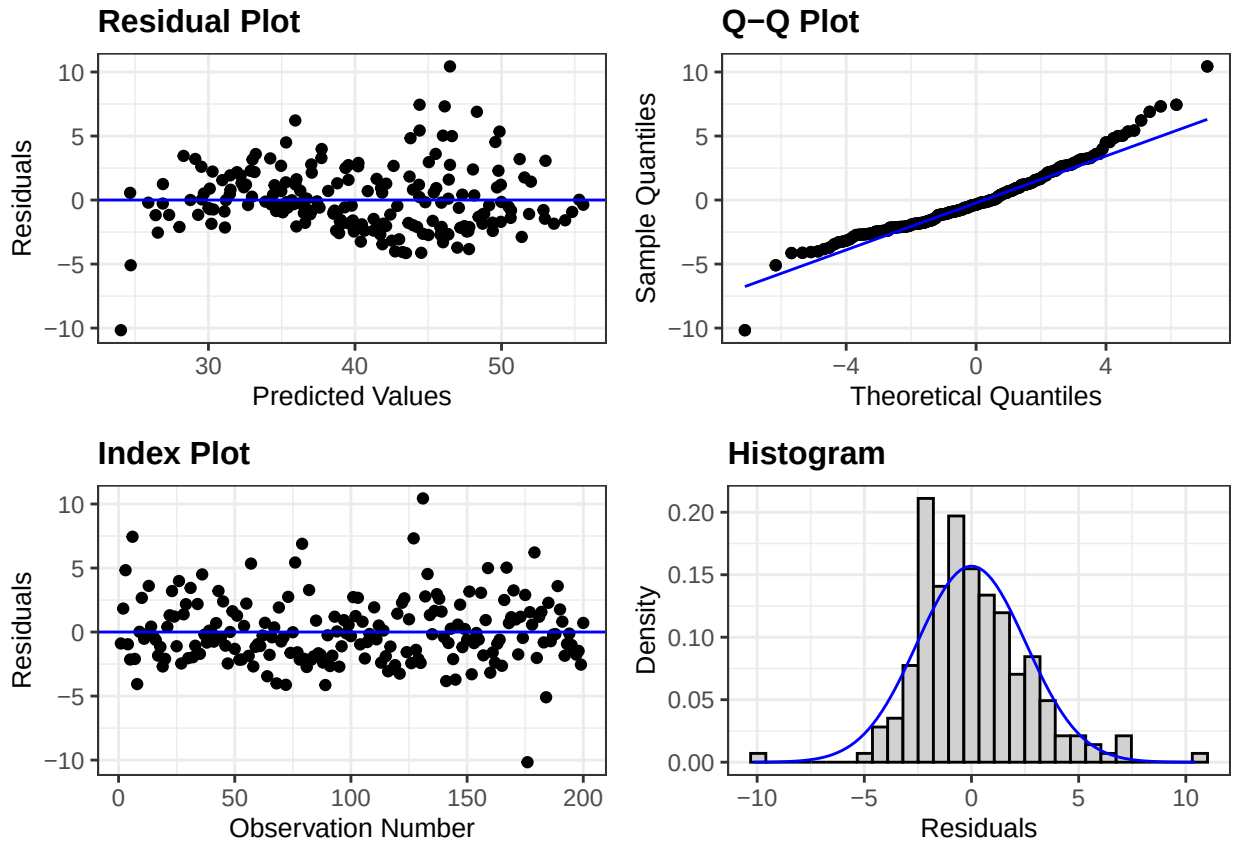
So the final model is:

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

analysing residuals

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.

```
resid_panel(mod2)
```



From the residuals vs fitted plot it appears that the variance is not constant and two observations have very large residuals.

```
Mplastics[which(abs(Mplastics$larvae - predict(mod2)) > 8),]
```

```
## # A tibble: 2 x 6
##   sample     PE    PP    PS  temp larvae
##    <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1    131   0.84  47.5  10.4  14.3   56.9
## 2    176 332.    58.7  50.2  13.2   13.9
```

It would be useful to understand the science of what is happening with these observations, but for now we should just proceed with caution. The QQ plot also suggests that there appears to be some deviation from normality in the tails of the data.

conclusions

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 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.