STAT210/410: Assessment 2 Solutions

The results for this assignment have been generally quite good. It was a good start to the unit - well done! I hope that most of you are now feeling a little more confident writing your own script/command files to analyse data. Increased confidence will come with practice!

Please go through the comments on your assignment and these solutions carefully. Contact myself for clarification if required.

Some of the more common mistakes/ misunderstandings in Assessment 2 related to the interpretation of confidence intervals for the regression coefficients

- Incorrect interpretation of confidence intervals. Please revise this fundamental concept (refer to STAT100 material or another introductory statistics reference). A CI is formulated using a sample estimate but it is **an interval estimate for the unknown *population parameter.***
- When interpreting the CI for a regression coefficient in multiple regression, note that you must acknowledge the presence of multicollinearity. The regression coefficient is the change in the response for a unit increase in the predictor, ***provided that the other predictors are held constant.***
- In general, the table label goes above a table and figure label is placed below a figure. Preferably have the figure or table follow the discussion of those results.

Solutions

(a) [15 marks]

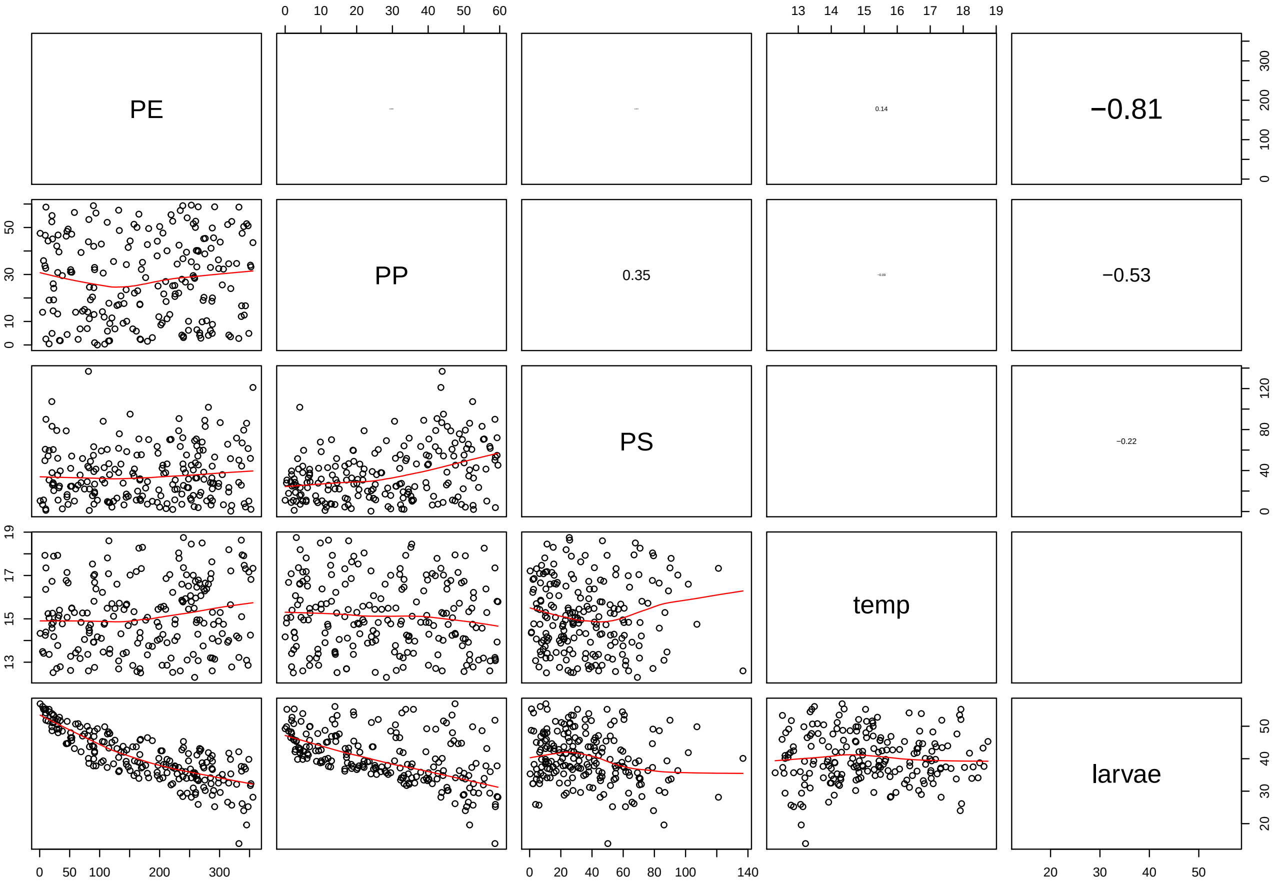

Figure 1: Pairs plot for the plastics data

From the pairs plot (Figure 1), we see that the number of larvae has a moderate to strong negative correlation with PP and PE, $r = -0.53$ and $r = -0.81$ respectively. However, there appears a weak correlation between the number of larvae with both PS and temperature ($r = -0.22$ and $r = -0.0097$ (Table 1), respectively).

Among the 4 predictors, it's noticed that PP is moderately positively correlated with PS ($r = 0.35$). There is little or no linear relationship between other pairs of predictors.

Producing Table 1 is optional.

Table 1: Correlation matrix for all 5 variables, with corresponding p-values

```
             PE       PP         PS      temp    larvae
PE      1.00000  0.44081  4.256e-01  0.050595 0.000e+00
PP      0.05481  1.00000  2.689e-07  0.207187 1.332e-15
PS      0.05665  0.35410  1.000e+00  0.948220 1.705e-03
temp    0.13843 -0.08957  4.621e-03  1.000000 8.913e-01
larvae -0.81102 -0.52555 -2.205e-01 -0.009725 1.000e+00
```

(b) [10 marks]

The table of regression coefficients is given in Table 2.

Table 2: 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
```

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

(c) [15 marks]

We'll use the t statistics and p-values in Table 2 to test the significance of individual partial regression coefficients.

PE

$H_0 : \beta_{PE} = 0$ or PE is not a significant predictor, given PP, PS and temp are already in the model.
$t = -33.44, p < 2^{-16}$(closer to 0). Since p-value is less than the threshold 0.05, we reject the null hypothesis.

PP

$H_0 : \beta_{PP} = 0$ or PP is not a significant predictor, given PE, PS and temp are already in the model.

$t = -18.83, p < 2^{-16}$(closer to 0). Since p-value is less than the threshold 0.05, we reject the null hypothesis.

PS

$H_0 : \beta_{PS} = 0$ or PS is not a significant predictor, given PE, PP and temp are already in the model.

$t = -0.32, p = 0.751$. Since p-value is greater than the threshold 0.05, we fail to reject the null hypothesis.

temp

$H_0 : \beta_{temp} = 0$ or temp is not a significant predictor, given PE, PP and PS are already in the model.

$t = 2.42, p = 0.016$. Since p-value is less than the threshold 0.05, we reject the null hypothesis.

In summary, only PS is the non-significant predictor, 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) [10 marks]

Results from the final model with only three predictors: PE, PP and temp are given in Tables 3.

The final model is:

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

Table 3: 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) [10 marks]

Table 4: 95% CI for the regression coefficients

```
            Estimate Std. Error    2.5 %   97.5 %
(Intercept) 52.51909   1.749580 49.06867 55.96951
PE          -0.05984   0.001784 -0.06335 -0.05632
PP          -0.20817   0.010257 -0.22840 -0.18794
temp         0.27260   0.112717  0.05030  0.49489
```

From Table 4

- The 95% CI for β_{PE} is (-0.063, -0.056) suggests that if PP and temp are held constant then for each $\mu g/m^3$ increase in polyethylene microplastics, the average number of fish larvae will decrease between 0.056 and 0.063.
- The 95% CI for β_{PP} is (-0.228, -0.188) suggests that if PE and temp are held constant then for each $\mu g/m^3$ increase in polypropylene microplastics, the average number of fish larvae will decrease between 0.188 and 0.228.
- The 95% CI for β_{temp} is (0.05, 0.49) suggests that if PE and PP are held constant then for each degree (Celsius) increase in temperature, the average number of fish larvae will increase between 0.05 to 0.49.

(f) [10 marks]

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 5), suggesting that the normality assumption for the residuals is violated.

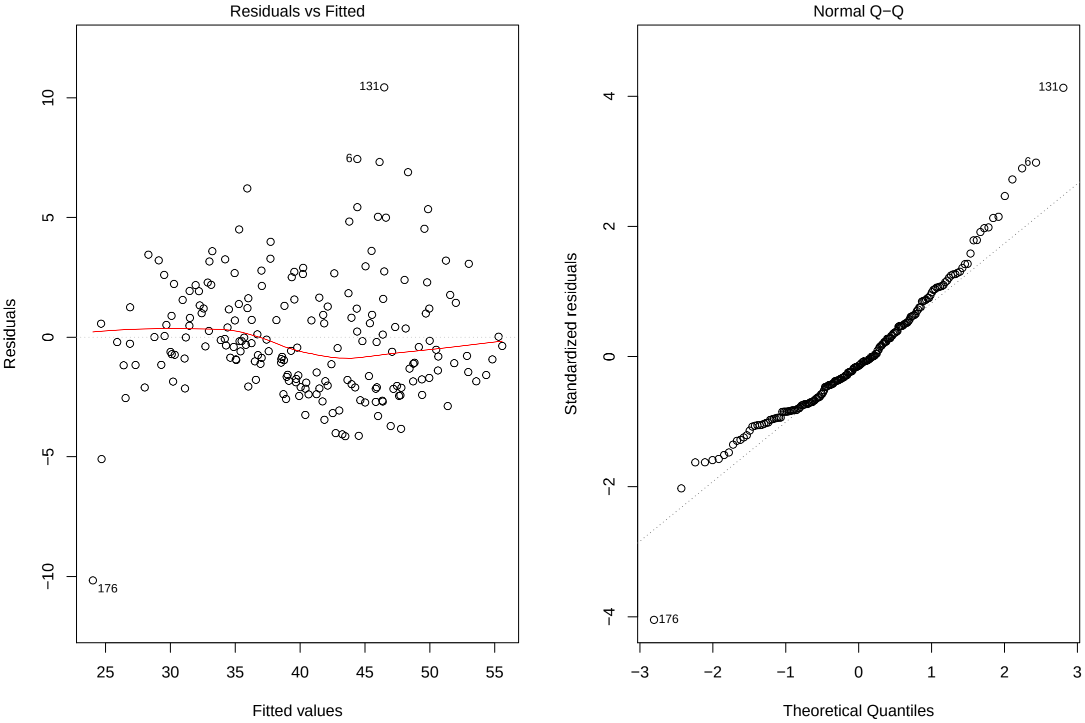

Figure 2: Residual plots

Table 5: Shapiro-Wilk normality test

```
Shapiro-Wilk normality test


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

(g) [15 marks]

From Table 6, we can predict that the mean number of fish larvae when temp=17.5, PE = 300 and PP=50 is 28.93 with a margin of error of 0.86, which is stated with 95% confidence (i.e. 95% CI is (28.07, 29.79)).

Similarly, we can predict that the mean number of fish larvae when temp=20.5, PE = 300 and PP=50 is 29.75 with a margin of error of 1.36, which is stated with 95% confidence (i.e. 95% CI is (28.39, 31.11))

Table 6: Means predicted response and 95% CI

```
    fit   lwr   upr
28.93 28.07 29.79
29.75 28.39 31.11
```

Note that the margin of error, and consequently the interval width, for the predicted response when temp = 20.5 is much larger than the prediction when tem = 17.5.

From the scatter plot (Figure 3), we can see that the data point temp = 17.5, PP=50, PPE=300 (green circle) is within the range of the observed data , however the point temp=20.5, PP=50, PPE=300 (red circle) is not. Alternatively, we can check the range for temp: 12.3 - 18.8.

So the prediction for the point at temp=20.5 is an extrapolation, which is not reliable, and thus should be avoided.

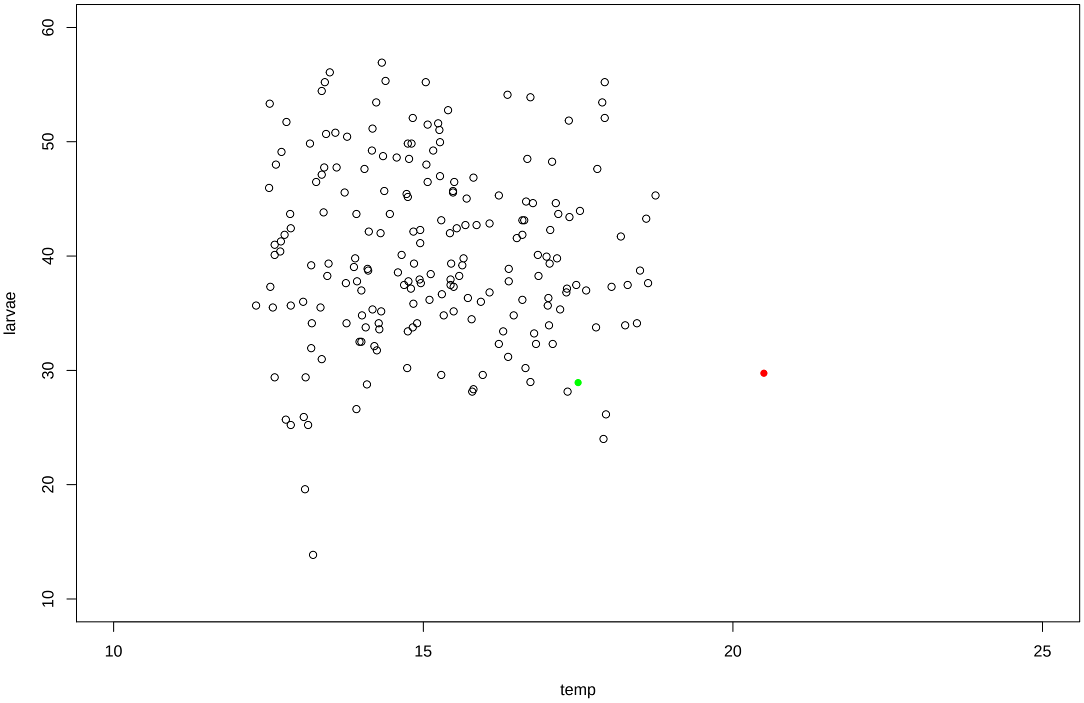

Figure 3: Scatter plot of larvae versus temp

(j) **Conclusion/Summary** [10 marks]

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. increase when temperature increase, given the amount of polyethylene and polypropylene microplastics remains the same. 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=4, show.signif.stars=F)
## Rfunctions.r required to produce pairs plot and confidence intervals
source("Rfunctions.r")
## 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
pairs(microP[, c(2:6)],lower.panel=panel.smooth,upper.panel=panel.cor)
## correlations and p-values
cor.prob(microP)
## 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)

## 95\% CIs for parameters
betaCI(mod2)
## Residuals plots
par(mfrow =c(1,2))
plot(mod2, which=1:2)
shapiro.test(mod2$residuals)
## predicted values
predict(mod2,list(temp=17.5,PP=50,PE=300),interval="confidence")
predict(mod2,list(temp=20.5,PP=50,PE=300),interval="confidence")

## Scatter plot of larvae vs temp
par(mfrow =c(1,1))
```

```
plot(larvae~temp, xlim=c(10,25), ylim=c(10,60), data=microP)
points(17.5, 28.93, col = "green", pch=16)
points(20.5, 29.75, col = "red",pch=16)

## check the range of temperature
summary(microP$temp)
```