Statistical Modeling I
Practical in R – Output

Practical in R – Output

In this practical, we will work with the Stackloss dataset (stackloss.csv). We will look at two different models and their analysis.

The data are obtained in a production process of oxidizing ammonia. The variables of interest are:

- Y: the stack loss, which is the percentage of the ingoing ammonia that escapes unabsorbed;
- X_1: the airflow
- X_2: the cooling water inlet temperature in degrees C;
- X_3: the acid concentration in percent.

1. Fit Model 1: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$ with $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

   First of all, we load the data by using the usual command:

```
> data <- read.csv(’stackloss.csv’)
> attach(data)
```

   Then we fit the model 1 by using the following commands:

```
> fit.lm <- lm(y ~ x1 + x2 + x3)
> summary(fit.lm)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72377 -0.17117 -0.04551  0.23614  0.56978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.61414    8.90214   0.406  0.68982
x1           0.07156    0.01349   5.307  5.8e-05 ***
x2           0.12953    0.03680   3.520  0.00263 **
```

```
x3           -0.15212    0.15629  -0.973  0.34405
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3243 on 17 degrees of freedom
Multiple R-squared:  0.9136,Adjusted R-squared:  0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

2. Check if there are any apparent problems with the residuals;

   In order to check if there are problems with the residuals, we need to create them and then checking the normality and linearity conditions:

```
> stdres1 <-rstandard(fit.lm)
> fits1<-fitted(fit.lm)
> plot(fits1,stdres1, main="Std res vs fits, stackloss")
> qqnorm(stdres1, main="Q-Q Plot, stackloss")
> qqline(stdres1)
```

   In Figure 1.1, we show the standardized residuals versus the fitted values (left) and the QQ plot (right). For the normality probability plot, we do not contradict the normality assumption, which is confirmed by the Shapiro-Wilk test (p-value equal to 0.6451). Regarding the residuals versus the fitted values, the plot casts some doubt on the assumption of a constant variance but not the linearity of the model

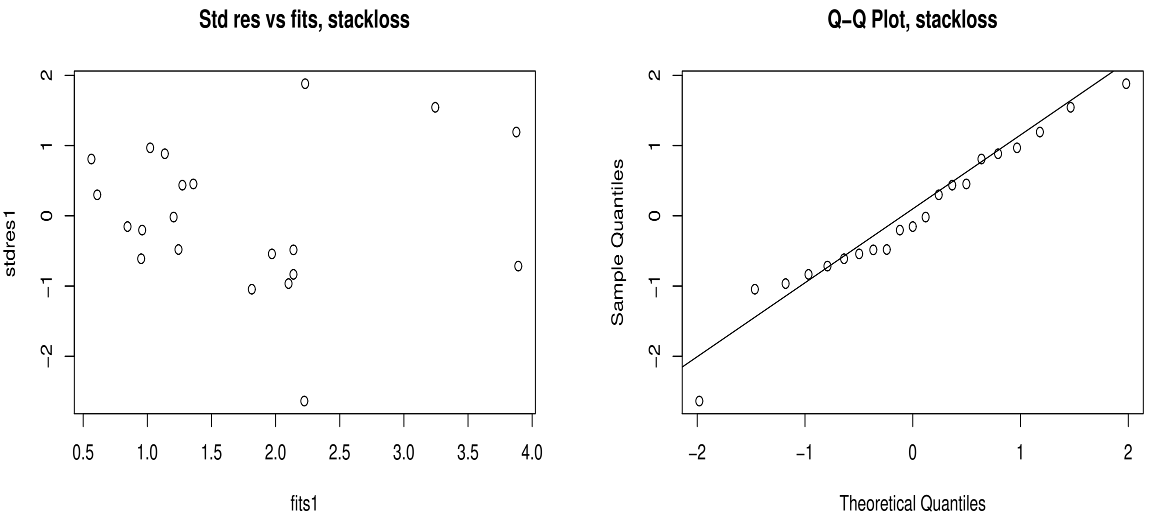

Figure 1.1: Plot of standardized residuals versus fitted values (left) and of the QQ plot for the model with three explanatory variables (right).

3. Test the hypothesis regarding the overall regression by using the F test

   Moving to the F test, we look at the last line of the summary in question 1), which is

   ```
   F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
   ```

   Thus there is strong evidence against the null hypothesis ($F = 59.9$):

   $$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \qquad \text{versus} \qquad H_1 : \text{ one of the } \beta_i \text{ is different from zero}$$

4. Test the hypothesis regarding the parameters β_j for $j = 0, 1, 2, 3$ by using the t tests

   Moving to the parameters, from the summary of the fitted models, we can see the t-tests for each coefficient. In particular, β_1 and β_2, the parameters related to x_1 and x_2 are statistically significant, thus we reject the null hypothesis $\beta_j = 0$ against the alternative $H_1 : \beta_j \neq 0$. This conclusion is not confirmed for the intercept and the parameter related to x_3, thus β_0 and β_3 are not statistically significant and we cannot reject the null hypothesis. Obviously all these considerations are done when all the other parameters are int the model.

5. Fit Model 2: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta x_{2,i} + \varepsilon_i$ with $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

   Moving to the second model, we use the usual command to fit the linear regression model with two explanatory variables:

   ```
   > fit1.lm <- lm(y ~ x1 + x2)
   > summary(fit1.lm)

   Call:
   lm(formula = y ~ x1 + x2)

   Residuals:
        Min       1Q   Median       3Q      Max
   -0.75290 -0.17505  0.01894  0.21156  0.56588

   Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   (Intercept) -5.03588    0.51383  -9.801 1.22e-08 ***
   x1           0.06712    0.01267   5.298 4.90e-05 ***
   x2           0.12954    0.03675   3.525  0.00242 **
   ---
   Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

   Residual standard error: 0.3239 on 18 degrees of freedom
   Multiple R-squared:  0.9088,Adjusted R-squared:  0.8986
   F-statistic: 89.64 on 2 and 18 DF,  p-value: 4.382e-10
   ```

6. Check if there are any apparent problems with the residuals;

   As done above, we need to define the standardized residuals in order to check the normality and linearity assumptions:

```
> stdres2 <-rstandard(fit1.lm)
> fits2<-fitted(fit1.lm)
> plot(fits2,stdres2, main="Std res vs fits, stackloss2")
> qqnorm(stdres2, main="Q-Q Plot, stackloss2")
> qqline(stdres2)
```

   Figure 1.2 shows in the right panel that there is no problem with the normality assumption, which is also confirmed from the Shapiro-Wilk test (p-value of 0.7321). Regarding the standardized residuals versus the fitted values, the left panel shows some doubts on the assumption of a constant variance but not regarding the linearity assumption.

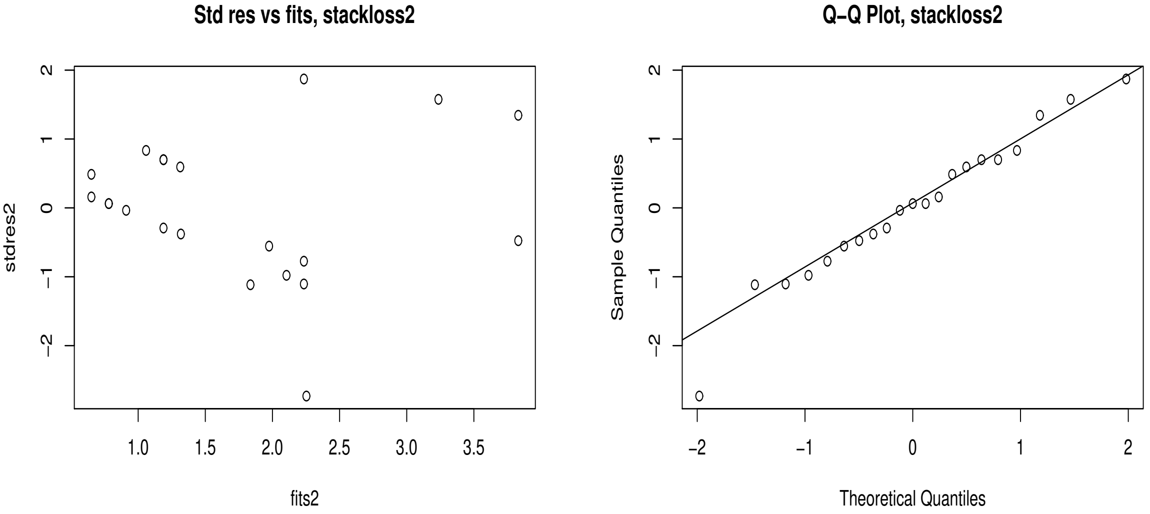

Figure 1.2: Plot of standardized residuals versus fitted values (left) and of the QQ plot for the model with two explanatory variables (right).

7. Test the hypothesis regarding the overall regression by using the F test

   To test the overall regression, we use the F test from the summary of the linear regression and in this case the values of the F test is 89.64, thus the overall regression is highly significant.

8. Test the hypothesis regarding the parameters β_j for $j = 0, 1, 2$ by using the t tests

   Moving to the parameters, from the summary of the linear regression, we can see that all the three parameters are statistically significant and thus we can reject the null hypothesis that the coefficients are equal to zero.

9. Which is the best model between Model 1 and Model 2 and why?

   Looking at the adjusted R^2 we can see that the Model 2 is better than Model 1 since the $adj(R^2) = 89.86\%$ that is high and comparable with the adjusted coefficient of determination for the full model, which is $adj(R^2) = 89.83\%$