Chapter 20

Multiple Regression

In this chapter we will discuss the following topic:

- How to calculate a Multiple Linear regression line using the R-function **lm()**.
- How to create pairwise scatterplots for all the variables using the R-function **pairs()**.
- How to test for significance of regression using the R-function **summary()** or **anova()**.
- How to test for which regressor(s) that contribute(s) to the model using the R-function **summary()**.

The equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

is called a **multiple linear regression model** with k explanatory variables. The equation is a linear function of the regression coefficients, β_i, for $i = 1, 2, ..., k$. We use the **method of least squares** to estimate the parameters β_i. The error term is assumed to be normally distributed with mean zero and variance σ^2, and all the errors are assumed to be uncorrelated.

The residual is again defined as the difference between an observed value y of the response variable and the predicted value $\hat{y}$ given by the regression line, that is

$$\text{residual} = y - \hat{y}.$$

Please refer to chapter 24 for the conditions for inference of regression.
When we are testing for significance of regression, we are testing the hypothesis,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

against

$$H_1 : \beta_i \neq 0 \text{ for at least one } j.$$

If the null hypothesis is rejected, we know that at least one of the regressors, x_1, x_2, ..., x_k gives a significant contribution to the model. We use **analysis of variance** to test the hypothesis, where the analysis of variance identity is given by

$$SS_T = SS_R + SS_{Res},$$

where SS_T is the **total sum of squares**, SS_R is the **regression sum of squares**, and SS_{Res} is the **residual sum of squares**. The term SS_T measures the total variability in the observed values, SS_R measures the amount of variability in the observed values explained by the regression line, and SS_{Res} is a measure of the residual variation **not** explained by the regression line. For further reading and for formulas, please refer to [4].

It can be shown that $\frac{SS_R}{\sigma^2}$ follows a chi-square distribution with the number of degrees of freedom equal to the number of regressors, k, in the model. Furthermore, it can be shown

that $\frac{SS_{Res}}{\sigma^2}$ follows a chi-square distribution with the number of degrees of freedom equal to $n-k-1$. Here $k+1$ constraints are given in the calculation of the residuals because of the estimation of the parameters β_0,..., β_k. Now define the **Regression mean square** as, $MS_R = \frac{SS_R}{k}$ and the **residual mean square** as $MS_{Res} = \frac{SS_{Res}}{n-k-1}$. If the null hypothesis is true, the statistic

$$F = \frac{\frac{SS_R}{k}}{\frac{SS_{Res}}{(n-k-1)}} = \frac{M_{S_R}}{MS_{Res}}$$

follows a F distribution with k and $n-k-1$ degrees of freedom. We reject the null hypothesis at the $\alpha\%$ significance level if F is greater than the critical value $F_\alpha(k, n-k-1)$. A large value of F indicates that a large proportion of the variation of the observations is explained by the regression line. This means that at least one of the variables, x_1, x_2, ...,x_k, significantly contributes to the model.

If the F test above shows that at least one of the regressors significantly contributes to the model, we will test to see which of the regressor(s) that contribute(s). We will test if x_i contributes to the model given that the other regressions are in the model. We test

$$H_0 : \beta_i = 0 \quad \text{against} \quad H_1 : \beta_i \neq 0.$$

If we don't reject H_0, it indicates that β_i is not significantly different from zero such that we can delete x_i from the regression model. If the null hypothesis is true, the test-statistic

$$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)},$$

where $se(\hat{\beta}_i)$ is the standard error of $\hat{\beta}_i$, follows a t-distribution with $(n-k-1)$ degrees of freedom. If $|t| > t_{\frac{\alpha}{2}}(n-k-1)$ we reject the null hypothesis at the $\alpha\%$ significance level.

Notice the following situations:

- If a regressor is additive in the variables a and b, we will use + between the explanatory variables such that we have $a+b$.
- To add an interaction term between to explanatory variables a and b, we do $a*b$. This is equivalent to $a+b+a:b$, where $a:b$ is the interaction term between a and b.

Problem. The table below found in the built-in data set **Stackloss** in **R**, shows observations of 21 days of "operation of a plant for the oxidation of ammonia to nitric acid", where *Air Flow* is the "flow of cooling air", *Water Temp* is "the temperature of cooling water", *Acid Conc.* is the "Concentration of acid[per 1000, minus 500]", and *stack.loss* is "an (inverse) measure of the over-all efficiency of the plant" [1], [2], and [3]. The quoted sentences are from [5].

Find the equation that best predict stack.loss (measure of the over-all efficiency of the plant).

Solution. We first call the data.frame **stackloss** and plot pairwise scatterplots:

```
> stackloss
   Air.Flow Water.Temp Acid.Conc. stack.loss
1        80         27         89         42
2        80         27         88         37
..       ..         ..         ..         ..
> par(mex=0.5)
> pairs(stackloss,gap=0,cex.labels=1,col="blue")
```

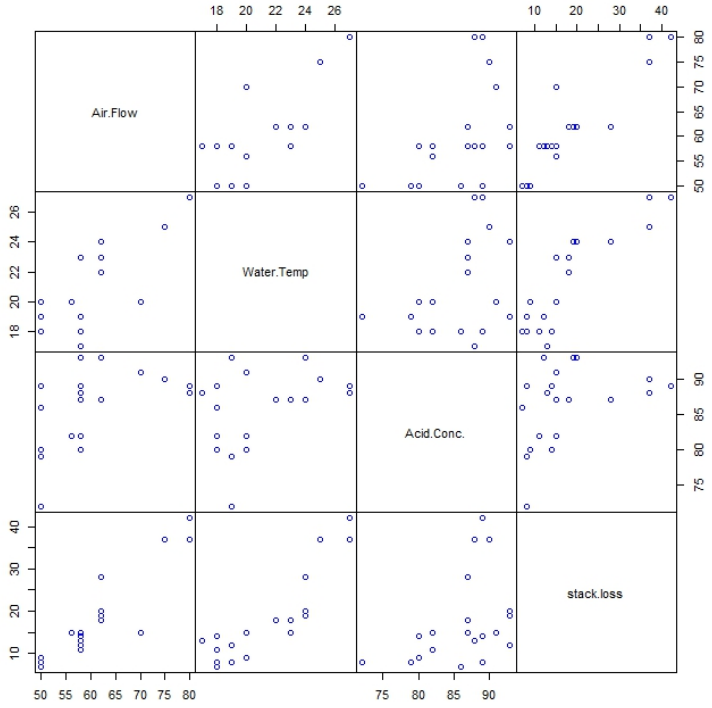

Figure 1: The figure shows the pairwise scatterplot of all the variables including the response variable.

From the graph, it seems like there is a linear relationship between *stack.loss* and *Air Flow* and between *stack.loss* and *Water Temp*. We perform a least squares regression:

```
> attach(stackloss)
> g.fullmodel=lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.)
> summary(g.fullmodel)

Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2377 -1.7117 -0.4551  2.3614  5.6978
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.9197    11.8960  -3.356  0.00375 **
Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680   3.520  0.00263 **
Acid.Conc.   -0.1521     0.1563  -0.973  0.34405
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.243 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
```

The p-value for the F-test is 3.016×10^{-9} so we can conclude that at least one of the regressors is significant. The p-values for the regressors Air.Flow and Water.Temp given the other regressors in the model are smaller than 0.05. The p-value for the regressor Acid.Conc. given the other regressors in the model is greater than 0.05 so we delete that regressor from the model. To obtain a new model:

```
> g=update(g.fullmodel,.~.-Acid.Conc.)
> summary(g)

Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp, data = stackloss)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5290 -1.7505  0.1894  2.1156  5.6588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -50.3588     5.1383  -9.801 1.22e-08 ***
Air.Flow      0.6712     0.1267   5.298 4.90e-05 ***
Water.Temp    1.2954     0.3675   3.525  0.00242 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.239 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
```

All the regressors are significant. The coefficient of determination, $R^2 = 0.9088$, is also large. The equation that predicts stack.loss is:

$$\text{stack.loss} = -50.3588 + 0.6712 \times \text{Air.Flow} + 1.2954 \times \text{Water.Temp}.$$

We next test the conditions for inference by drawing residual plots and a QQ-plot:

```
> res=rstandard(g)
> opar=par(mfrow=c(2,2),mex=0.5)
> plot(Air.Flow,res,col="green")
> abline(0,0,col="red")
> plot(Water.Temp,res,col="blue")
> abline(0,0,col="red")
> plot(fitted(g),res,col="brown")
> abline(0,0,col="red")
> qqnorm(res,col="blue")
> qqline(res,col="red")
```

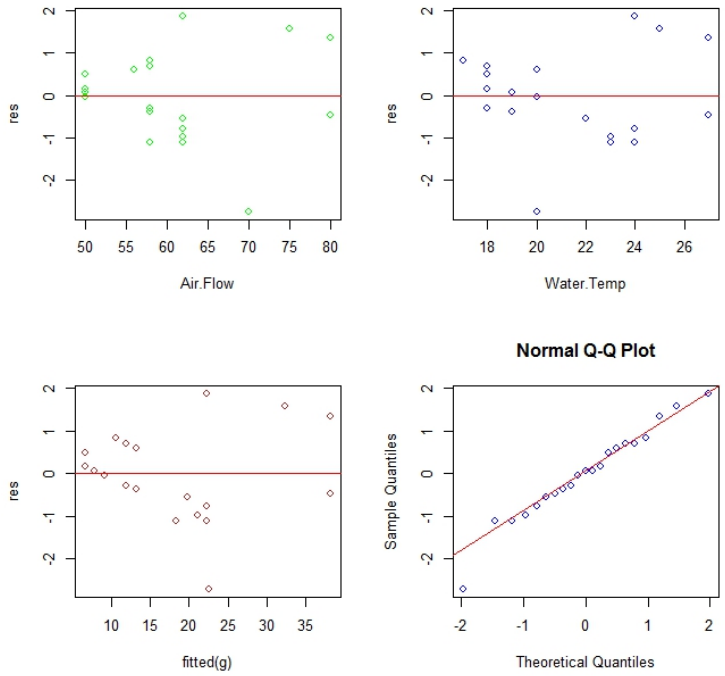

Figure 2: The figure in the upper left corner shows a residual plot against the variable Air.Flow. The figure in the upper right corner shows a residual plot against the variable Water.Temp. The figure in the lower left corner shows a residual plot against the predicted values and the figure in the lower right corner shows the QQ-plot of the residuals.

It seems that the variance of the residuals increases somewhat with increasing values for the explanatory variables. A transformation that stabilize the variance might give a more accurate result. The QQ-plot does not reveal any departures from normality.

Explanation. The code can be explained as follows:

- The command **stackloss** calls the data set **stackloss**.
- The command **par(mex=0.5)** initialize the window for the plot.
- The function **pairs(stackloss,gap=0,cex.labels=1)** plots a pairwise scatterplot between all the variables. The command **gap** sets the distance between each individual "picture". The command **cex.labels** sets the font size.
- The command **g=update(g.fullmodel,.~.-Acid.Conc.)** deletes *Acid.Conc* from the regression model and computes a new model.
- The command **opar=par(mfrow=c(2,2))** initialize a 2×2 window for the plot.
- The function **fitted(g)** computes the predicted values, $\hat{y}$, of the regression line.

Example. In this example, we will consider the data set called **kfm** [5] which is located under the library called **ISwR**. This library needs to be installed from **Packages** in R. (please refer to the Introduction chapter in the tutorial for how to install libraries). The **kfm** data in R was originally collected by Kim Fleischer Michaelsen. It provides data for 50 infants of age approximately 2 months who were weighed immediately before and after each breast feeding. The intake of breast milk in dl/24h (dl.milk) was measured along with weight of child in kg (weight), supplementary milk substitute in ml/24h (ml.suppl), weight of mother in kg (mat.weight), height of mother in cm (mat.height), and sex (boy and girl). Fit a regression line that explains the intake of breast milk. Is sex a significant factor?

Solution. We first need to install the library **ISwR** under **Packages** in R. The data set **kfm** is contained in this library. We turn **sex** into a factor with levels boy=0 and girl=1. (Here **sex** is already a factor so this step is in this case not necessary). We first try to fit a full model with interaction terms between all the variables:

```
> library(ISwR)
> kfm
    no dl.milk  sex weight ml.suppl mat.weight mat.height
1    1    8.42  boy   5.002      250         65       173
2    4    8.44  boy   5.128        0         48       158
..   ..    ..    ..   ..          ..        ..         ..
49  104    6.97  girl  4.890      30         67       165
50  105    5.82  girl  4.339      95         47       163

> kfm$Sex=factor(kfm$sex,level=c("boy","girl"),label=c(0,1))
> kfm$Sex
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1
[39] 1 1 1 1 1 1 1 1 1 1 1 1
Levels: 0 1
```

Here we added the factor **Sex** to the data frame **kfm** with levels 0 and 1. To see this type:

```
> kfm
    no dl.milk  sex weight ml.suppl mat.weight mat.height  Sex
1    1    8.42  boy   5.002      250         65        173    0
2    4    8.44  boy   5.128        0         48        158    0
..   ..    ..    ..   ..          ..        ..         ..     ..
49  104    6.97  girl  4.890      30         67       165     1
50  105    5.82  girl  4.339      95         47       163     1

> g.fullmodel=lm(dl.milk~Sex*weight*ml.suppl*mat.weight*mat.height,data=kfm)
> summary(g.fullmodel)
```

We do not here include the output (you should check). The p-value for the F-test is 0.04 so we can conclude that at least one of the regressors are significant. However, none of the individual variables given the other variables are significant in this model since their p-values are all greater than 0.05. Thus, a full model is not appropriate. You should also try to include an interaction term between some of the variables and see what happens. Next, we fit an additive model. We add the command **data=kfm** into the **lm()** function since we have not attached the data frame.

```
> g=lm(dl.milk~Sex+weight+ml.suppl+mat.weight+mat.height,data=kfm)
> summary(g)

Call:
lm(formula = dl.milk ~ Sex + weight + ml.suppl + mat.weight +
    mat.height, data = kfm)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74201 -0.81173 -0.00926  0.78326  2.52646

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.681839   4.361561  -2.678 0.010363 *
Sex1         -0.499532   0.312672  -1.598 0.117284
weight        1.349124   0.322450   4.184 0.000135 ***
ml.suppl     -0.002233   0.001241  -1.799 0.078829 .
mat.weight    0.006212   0.023708   0.262 0.794535
mat.height    0.072278   0.030169   2.396 0.020906 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.075 on 44 degrees of freedom
Multiple R-squared: 0.5459,     Adjusted R-squared: 0.4943
F-statistic: 10.58 on 5 and 44 DF,  p-value: 1.03e-06
```

Here the regressors **mat.weight**, **ml.suppl**, and **Sex** are not significant. We first delete the regressor **mat.weight** from the model:

```
> g2=update(g,.~.-mat.weight,data=kfm)
> summary(g2)

Call:
lm(formula = dl.milk ~ Sex + weight + ml.suppl + mat.height,
    data = kfm)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77312 -0.81196 -0.00683  0.76988  2.52240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.112571   3.997860  -3.030  0.00405 **
Sex1         -0.494675   0.308875  -1.602  0.11626
weight        1.372524   0.306612   4.476 5.14e-05 ***
ml.suppl     -0.002313   0.001190  -1.943  0.05824 .
mat.height    0.076363   0.025560   2.988  0.00454 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.064 on 45 degrees of freedom
Multiple R-squared: 0.5452,     Adjusted R-squared: 0.5047
F-statistic: 13.48 on 4 and 45 DF,  p-value: 2.658e-07
```

The regressors **Sex** and **ml.suppl** are still not significant. We next delete **Sex** from the previous model:

```
> g3=update(g2,.~.-Sex,data=kfm)
> summary(g3)

Call:
lm(formula = dl.milk ~ weight + ml.suppl + mat.height, data = kfm)

Residuals:
     Min       1Q   Median       3Q      Max
-2.06540 -0.74758 -0.02408  0.67488  2.79882

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.064926   4.020073  -3.250  0.00216 **
weight        1.464781   0.306231   4.783 1.81e-05 ***
ml.suppl     -0.002237   0.001209  -1.850  0.07074 .
mat.height    0.077600   0.025979   2.987  0.00451 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 1.082 on 46 degrees of freedom
Multiple R-squared: 0.5192,     Adjusted R-squared: 0.4879
F-statistic: 16.56 on 3 and 46 DF,  p-value: 1.953e-07
```

The regressor **ml.suppl** is not significant, so we delete that regressor from the previous model:

```
> g4=update(g3,.~.-ml.suppl,data=kfm)
> summary(g4)

Call:
lm(formula = dl.milk ~ weight + mat.height, data = kfm)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19598 -0.82149  0.01822  0.75582  2.83375

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.92014    4.07325  -2.926  0.00527 **
weight        1.42862    0.31338   4.559 3.67e-05 ***
mat.height    0.07063    0.02636   2.680  0.01013 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.109 on 47 degrees of freedom
Multiple R-squared: 0.4835,     Adjusted R-squared: 0.4615
F-statistic:    22 on 2 and 47 DF,  p-value: 1.811e-07
```

Now all the regressors are significant. To check for significant difference between models g and $g4$, we do:

```
> anova(g,g4)
Analysis of Variance Table

Model 1: dl.milk ~ Sex + weight + ml.suppl + mat.weight + mat.height
Model 2: dl.milk ~ weight + mat.height
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 50.840
2     47 57.826 -3   -6.9861 2.0154 0.1256
```

The explanation for this ANOVA table is beyond this course. It shows that there is a difference of 3 degrees of freedom between model **g** and model **g4** since they differ by three parameters. The p-value of 0.1256 shows that there is not a significant difference between

these two models. Thus, we will keep the simpler model with fewer parameters. Thus, we obtain the regression line:

$$\text{dl.milk} = -11.92014 + 1.42862 \times \text{Weight} + 0.07063 \times \text{mat.height}.$$

We next plot **dl.milk** against **mat.height** and **height** with separate colors and symbols for **girl** and **boy**.

```
> par(mfrow=c(1,2))
> plot(dl.milk~mat.height,pch=c(1,2)[Sex],col=c("blue","red")[Sex],data=kfm)
> plot(dl.milk~weight,pch=c(1,2)[Sex],col=c("blue","red")[Sex],data=kfm)
```

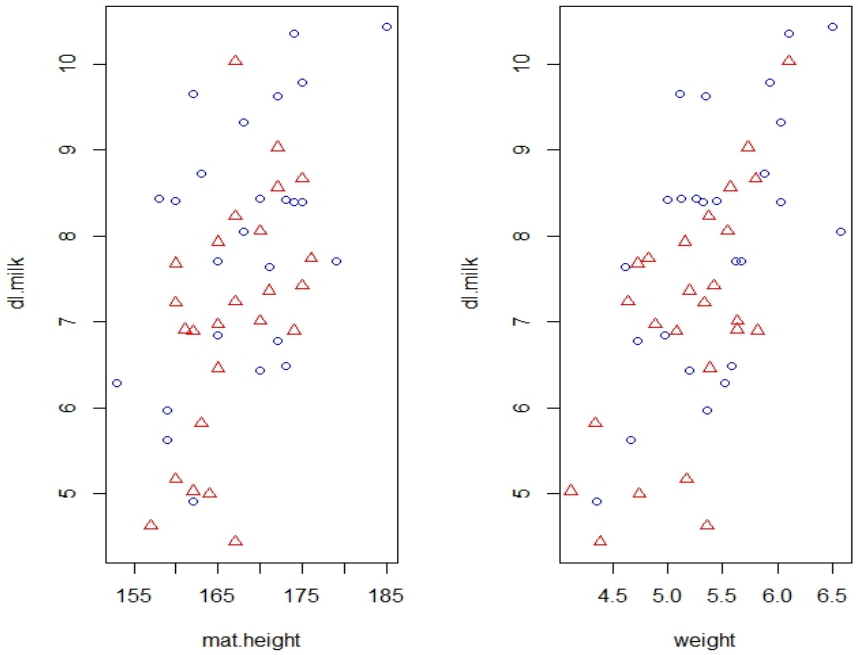

Figure 3: The blue color is for the factor **boy** and the red color for the factor **girl**.

The factor **sex** is here not significant and is therefore deleted from the model.
We check the conditions for inference:

```
> res=rstandard(g4)
> attach(kfm)
> opar=par(mfrow=c(2,2),mex=0.6)
> plot(weight,res,col="green")
> abline(0,0,col="red")
> plot(mat.height,res,col="purple")
> abline(0,0,col="red")
> plot(fitted(g4),res,col="blue")
```

```
> abline(0,0,col="red")
> qqnorm(res,col="darkgreen")
> qqline(res,col="red")
```

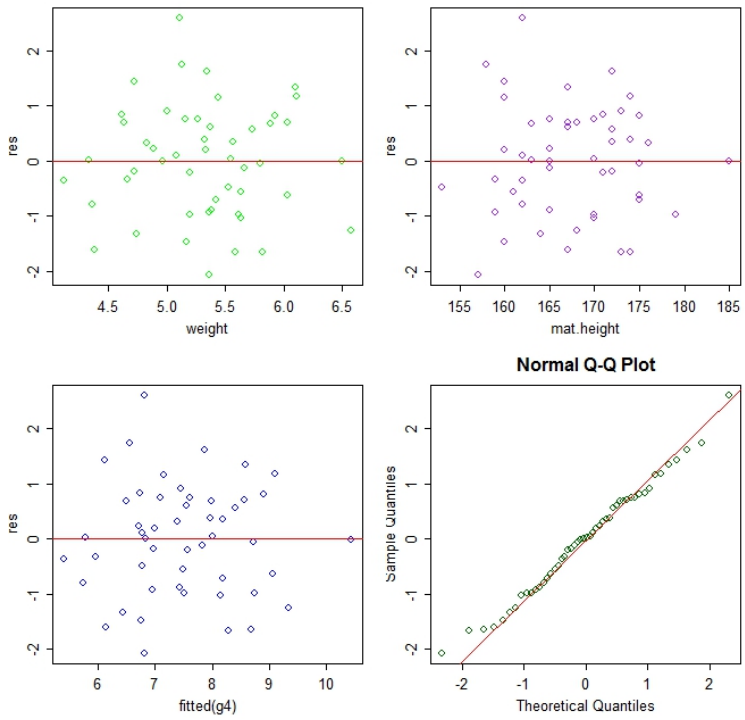

Figure 4: The figure in the upper left corner shows a residual plot against the variable **weight**. The figure in the upper right corner shows a residual plot against the variable **mat.height**. The figure in the lower left corner shows a residual plot against the predicted values and the figure in the lower right corner shows the QQ-plot of the residuals.

There are no unusual pattern in the regression plots. The QQ-plot does not reveal any departures from normality.

Explanation. The code can be explained as follows:

- The command **library(ISwR)** calls the package **ISwR**.
- The function

  ```
  factor(kfm$sex,level=c("boy","girl"),label=c(0,1))
  ```

  converts the variable **sex** into a factor or change the name of the levels of an existing factor. The entries in **level** are the original levels of the factor. The command **label** changes the names of the levels. In this case, the names of the old levels were **"boy"** and **"girl"** and the names of the new levels are **0** and **1**. If **level** and **label** are not set, the default in R is to give the original levels in alphabetic order.
- The function **anova(g,g4)** performs an ANOVA and computes the differences between models **g** and **g4**.

References

[1] R. A. Becker, J. M. Chambers, and A. R. Wilks. *The New S Language. Wadsworth Brooks/Cole*, 1988.

[2] K. A. Brownlee. *Statistical Theory and Methodology in Science and Engineering*, New York: Wiley. pp. 491500, 1960, 2nd ed. 1965.

[3] Dodge, Y. (1996) *The guinea pig of multiple regression In: Robust Statistics, Data Analysis, and Computer Intensive Methods; In Honor of Peter Huber's 60th Birthday*, Lecture Notes in Statistics 109, Springer-Verlag, New York, 1996.

[4] D. C. Montgomery, E. A. Peck, G. Geoffrey Vining. *Introduction to linear regression analysis*, Fourth Edition, John Wiley Sons, INC. Publication, 2006.

[5] See R-documentation at

`http://127.0.0.1:31382/library/ISwR/html/kfm.html`

00