Demo 2.2

Nguyen Xuan Binh

11/12/2021

When cement hardens, heat is produced. The amount of heat depends on the composition of the cement. From file hald.txt, you can find the following information regarding 13 different batches of cement: HEAT =heat energy (cal/g) CHEM1, CHEM2, CHEM3, CHEM4 =ingredients of cement (% of the dry substance)

Solution. The goal of the exercise is to find out which of the explanatory variables CHEM1, CHEM2, CHEM3, CHEM4 are significant in explaining the behavior of the response variable HEAT. First, we import the data and install the package car for later use.

install.packages("car")

```
library(car)

## Loading required package: carData
```

a) Estimate a linear regression model with all explanatory variables. Compare statistical significance of the regression coefficients and examine the variance inflation factors of the corresponding explanatory variables.

Estimation of the full model In situations, where it is not known which of the explanatory variables affect the response variable, it is first usually reasonable to estimate the full model, i.e. the model with all candidates for explanatory variables. First, we should examine the correlations between the different variables.

```
hald=read.table("hald.txt",header=T)
cor(hald)

##              CHEM1      CHEM2      CHEM3      CHEM4       HEAT          SUM
## CHEM1  1.00000000  0.2285795 -0.8241338 -0.2454451  0.7307175  0.05010722
## CHEM2  0.22857947  1.0000000 -0.1392424 -0.9729550  0.8162526 -0.26044918
## CHEM3 -0.82413376 -0.1392424  1.0000000  0.0295370 -0.5346707 -0.11025122
## CHEM4 -0.24544511 -0.9729550  0.0295370  1.0000000 -0.8213050  0.32907694
## HEAT   0.73071747  0.8162526 -0.5346707 -0.8213050  1.0000000 -0.16458053
## SUM    0.05010722 -0.2604492 -0.1102512  0.3290769 -0.1645805  1.00000000
```

The variable HEAT correlates strongly with all explanatory candidates. Correlation is positive with the variables CHEM1 and CHEM2, and negative with CHEM3 and CHEM4. There is a strong negative correlation between variables CHEM1 and CHEM3, as well as between variables CHEM2 and CHEM4. We begin by estimating the full model:

$$HEAT = \beta 0 + \beta 1 CHEM1 + \beta 2 CHEM2 + \beta 3 CHEM3 + \beta 4 CHEM4 + \varepsilon \quad (1)$$

```
fullmodel=lm(HEAT~CHEM1+CHEM2+CHEM3+CHEM4,data=hald)
summary(fullmodel)
```

```
##
## Call:
## lm(formula = HEAT ~ CHEM1 + CHEM2 + CHEM3 + CHEM4, data = hald)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1750 -1.6709  0.2508  1.3783  3.9254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.4054    70.0710   0.891   0.3991
## CHEM1         1.5511     0.7448   2.083   0.0708 .
## CHEM2         0.5102     0.7238   0.705   0.5009
## CHEM3         0.1019     0.7547   0.135   0.8959
## CHEM4        -0.1441     0.7091  -0.203   0.8441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.446 on 8 degrees of freedom
## Multiple R-squared:  0.9824, Adjusted R-squared:  0.9736
## F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

The model (1) has a high coefficient of determination (98.2%). The value of the F-test statistics for the null hypothesis H0 : $\beta1 = \beta2 = \beta3 = \beta4 = 0$ is 111.5 and the p-value is close to zero, i.e. the model is statistically significant and at least one of the regression coefficients $\beta0$; $\beta1$; $\beta2$; $\beta3$ deviates from zero. However, none of the explanatory variables of the model (1) is statistically significant with a 5%:n level of significance. This is due to the multicollinearity of the explanatory variables. Multicollinearity of the explanatory variables can be measured with VIF-coefficients. The VIF-coefficient is 1 for an explanatory variable whose sample correlation is 0 with other explanatory variables. The stronger a variable is linearly dependent on the other variables, the larger the VIF-coefficient of the variable is. If VIF > 10; then multicollinearity might be a problem. VIF-coefficients can be computed with the function vif of the package car.

```
vif(fullmodel)
```

```
##     CHEM1     CHEM2     CHEM3     CHEM4
##  38.49621 254.42317  46.86839 282.51286
```

In model (1), the VIF-coefficients of the variables CHEM2 and CHEM4 are larger than 200, which indicates that strong multicollinearity is present in the model. Next, we further study the existing multicollinearity by estimating two regression models, where CHEM2 and CHEM4 are explained with all the other explanatory variables of the original model (1). Consider the model:

$$CHEM2 = \alpha0 + \alpha1CHEM1 + \alpha3CHEM3 + \alpha4CHEM4 + \delta; \quad (2)$$

which can be estimated using,

```
model2 <- lm(CHEM2 ~ CHEM1+CHEM3+CHEM4,data=hald)
summary(model2)
```

```
##
## Call:
## lm(formula = CHEM2 ~ CHEM1 + CHEM3 + CHEM4, data = hald)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2494 -0.7280  0.3881  0.7033  0.9512
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 96.59382    2.16253   44.67 7.06e-12 ***
## CHEM1       -0.97860    0.10602   -9.23 6.94e-06 ***
## CHEM3       -1.00350    0.09443  -10.63 2.15e-06 ***
## CHEM4       -0.97759    0.02111  -46.30 5.12e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.126 on 9 degrees of freedom
## Multiple R-squared:  0.9961, Adjusted R-squared:  0.9948
## F-statistic: 760.3 on 3 and 9 DF,  p-value: 3.864e-11
```

The coefficient of determination of the model is 99.6% implying that CHEM2 is strongly linearly dependent on the other explanatory variables. Note that the VIFcoefficient of CHEM2 in the model (1) is VIF2 = 1/(1 − R2^2); where R2^2 is the coefficient of determination for model (2). Consider the model,

CHEM4 = α0 + α1CHEM1 + α2CHEM2 + α3CHEM3 + δ; (3)

which can be estimated using,

```
formula <- lm(CHEM4 ~ CHEM1+CHEM2+CHEM3,data=hald)
summary(formula)
```

```
##
## Call:
## lm(formula = CHEM4 ~ CHEM1 + CHEM2 + CHEM3, data = hald)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3264 -0.6836  0.4439  0.7463  1.0379
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 98.65079    1.94627  50.687 2.27e-12 ***
## CHEM1       -1.00504    0.10175  -9.878 3.96e-06 ***
## CHEM2       -1.01865    0.02200 -46.303 5.12e-12 ***
```

```
## CHEM3       -1.02809    0.09187 -11.191 1.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.15 on 9 degrees of freedom
## Multiple R-squared:  0.9965, Adjusted R-squared:  0.9953
## F-statistic: 844.5 on 3 and 9 DF,  p-value: 2.413e-11
```

The coefficient of determination of the model is 99.7% implying that CHEM4 is strongly linearly dependent on the other explanatory variables. Note that the VIF-coefficient of CHEM4 in the model (1) is VIF4 = 1/(1 − R3^2); where R3^2 is the coefficient of determination of the model (3). Multicollinearity of the model (1) is explained by noting that cement consists almost entirely of the substances CHEM1, CHEM2, CHEM3 and CHEM4. The sum of these variables is somewhere between 95-99%. Therefore, by increasing the amount of a substance, we have to reduce the amount of some other substances in the mixture. This explains the strong negative correlations between the variable pairs (CHEM1, CHEM3) and (CHEM2, CHEM4).

b) Find the best combination of explanatory variables by using Akaike information criterion (AIC).

There exists different strategies for choosing the explanatory variables of a regression model. When searching for the best combination of explanatory variables, different models are compared to each other by using some criterion for model selection.

Some well-known criteria for model selection are, e.g., Akaike information criterion (AIC), Schwarz bayesian information criterion (SBIC) and Hannan-Quinn criterion (HQ). The criterion functions of model selection methods are of the form, $\min_{M \subseteq (1,\ldots,q)} C(|M|; (\sigma M^2)$, where M is a combination of explanatory variables and $(\sigma|M|^2$ is the maximum likelihood estimator for the variance of the residuals of the corresponding model. Furthermore, C is an increasing function with respect to the two arguments. In general, we expect the following from a criterion function:

• Maximal coefficient of determination,

• Using as few explanatory variables as possible. In R, the function step() gives the combination of explanatory variables that minimizes the value of AIC. Note that step() computes AIC by assuming normally distributed residuals.

```
step(fullmodel)
```

```
## Start:  AIC=26.94
## HEAT ~ CHEM1 + CHEM2 + CHEM3 + CHEM4
##
##         Df Sum of Sq    RSS    AIC
## - CHEM3  1    0.1091 47.973 24.974
## - CHEM4  1    0.2470 48.111 25.011
## - CHEM2  1    2.9725 50.836 25.728
## <none>               47.864 26.944
## - CHEM1  1   25.9509 73.815 30.576
```

```
##
## Step:  AIC=24.97
## HEAT ~ CHEM1 + CHEM2 + CHEM4
##
##         Df Sum of Sq    RSS    AIC
## <none>                47.97 24.974
## - CHEM4  1      9.93  57.90 25.420
## - CHEM2  1     26.79  74.76 28.742
## - CHEM1  1    820.91 868.88 60.629

##
## Call:
## lm(formula = HEAT ~ CHEM1 + CHEM2 + CHEM4, data = hald)
##
## Coefficients:
## (Intercept)        CHEM1        CHEM2        CHEM4
##     71.6483       1.4519       0.4161      -0.2365
```

The output can be interpreted as follows. The AIC of the full model is 26.944. When CHEM3 is omitted from the model, the AIC is 24.974. When CHEM4 is omitted, the AIC is 25.011. When CHEM2 is omitted, the AIC is 25.728 and when CHEM1 is omitted, the AIC is 30.576. We wish to minimize the model selection criterion and hence, we estimate the model without CHEM3.

Consider the model,

HEAT = β0 + β1CHEM1 + β2CHEM2 + β4CHEM4: (4)

Now the AIC of model (4) is 24.974. From the output of R, we see that omitting any of the remaining explanatory variables (CHEM1, CHEM2, CHEM4) would increase the AIC value. Next, we estimate the model (4)

```
model4 <- lm(HEAT ~ CHEM1 + CHEM2 + CHEM4 , data=hald)
summary(model4)

##
## Call:
## lm(formula = HEAT ~ CHEM1 + CHEM2 + CHEM4, data = hald)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0919 -1.8016  0.2562  1.2818  3.8982
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.6483    14.1424   5.066 0.000675 ***
## CHEM1         1.4519     0.1170  12.410 5.78e-07 ***
## CHEM2         0.4161     0.1856   2.242 0.051687 .
## CHEM4        -0.2365     0.1733  -1.365 0.205395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.309 on 9 degrees of freedom
## Multiple R-squared:  0.9823, Adjusted R-squared:  0.9764
## F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

Note that the variables CHEM2 and CHEM4 are not statistically significant with 5% significance level. Figure 5 illustrates the estimated residuals of the full model. The shape of the histogram indicates that the normality assumption does not hold, which on the other hand means that AIC is not a reliable method for model selection. In homework assignment 2.3, the model selection is done using the permutation test. The permutation test does not require normality and thus, it is the safer alternative here.

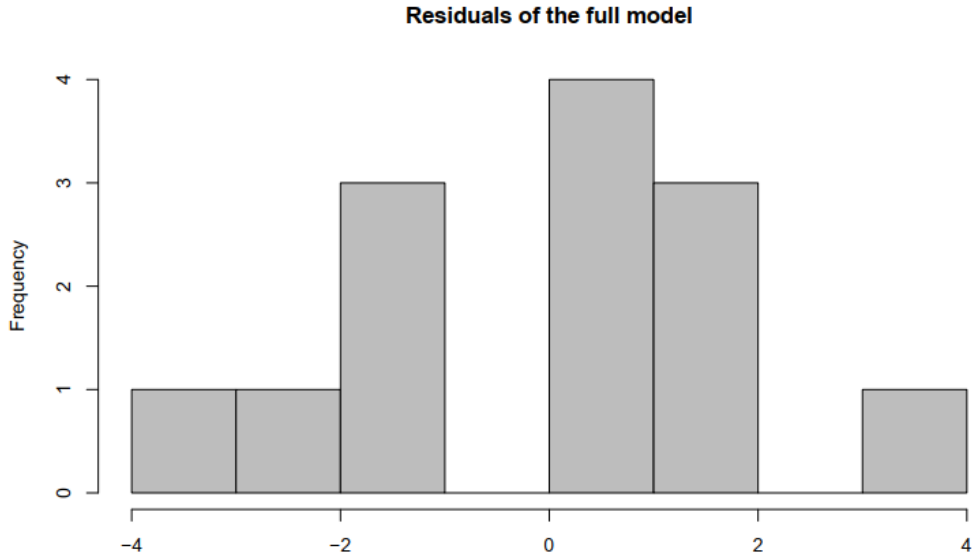

Figure 5: The residuals of the full model.

Remark: It is not possible to use the error sum of squares or the coefficient of determination as a criterion for model selection, since minimizing the error sum of squares as well as maximizing the coefficient of determination always leads to the full model (the model with all possible explanatory variables).