Compulsory exercise 2

TMA4267 Linear statistical models spring 2020

Silje Anfindsen and Magnus Wølneberg

13 March, 2020

Problem 1 : Diabetes progression

a)

Interpretation of the print-out from **summary(full)** in Figure 1:

1)

We will now explain the meaning of the quantities presented under "Coefficients" in the summary. The coefficients are the unknown parameters β_0, β_1,... β_p in the linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$.

The first column, **Estimate**, gives the estimates of the parameteres, the intercept β_0 and the slopes β_i. The parameters are usually calculated using the least squares approach, that is to choose β_i such that the sum of squared residuals is minimized $\sum_{i=1}^{n}(y_i - \hat{y_i})^2$. We can interpret the estimates as the expected change in Y for a one unit increase in the variable X_i. In this case a one unit change in **age** gives -0.03529 unit changes in **prog**. We can now find the formula for the full model,

$$\begin{aligned}\textbf{prog} = &-356.643 - 0.035\textbf{age} - 22.792\textbf{sex} + 5.595\textbf{bmi} + 1.115\textbf{map} \\ &+ 1.082\textbf{tc} + 0.739\textbf{ldl} + 0.367\textbf{hdl} + 6.540\textbf{tch} + 157.176\textbf{ltg} + 0.281\textbf{glu}\end{aligned}$$

The second column, **Std. Error** is the standard deviation of an estimate, which in this case measures the average amount that the coefficient estimates vary from the response. In other words, how precisely the model estimates the coefficients unknown values. The standard error is

$$SE = \frac{\sigma}{\sqrt{n}}$$

The third column **t vallue** is given by

$$t = \frac{Estimate}{Std.Error} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

A high t value indicate that we can reject the null hypothesis.

The last column **Pr(>|t|)** inidicates the probability of observing a value equal or larger than t, also called the p value. The p value is helpful when we do a hypothesis test. A low p value allow us to reject the null hypothesis, which means that there is a relationship between the given parameter and the response variable in our model.

2)

The intercept, β_0 gives the measured response Y when all the variables are set to null, $x_i = 0$, that means we do not consider any variables when calculating the response.

3)

The estimated regression coeffisients for **bmi** is positive, therefore and increasing BMI value for the patient will contribute to an increasing response, which in this case is the measurement of disease progression of diabetes after one year. This means that for a patient with a high BMI value, progression of diabetes one year after baseline will be worse.

4)

The estimated error variance is denoted as the Residual Standard Error in the summary. The Residual Standard Error is the average amount that the response will deviate from the true regression line, and it is given by:

$$RSE = \sqrt{\frac{\sum(\hat{y} - y)}{n - p}}$$

where p is the number of degrees of freedom, and n is number of observations.

5)

The covariates within the given significance level of $\alpha = 0.05$ are **sex**, **bmi**, **map** and **ltg**. In order to determine whether there is a relationship between the response and each of the predictors, we can check if setting $\beta_i = 0$ affects the response or not.

In order to test the significance for each of the covariates we perform a t test where the null hypothesis is,

$H_0 : \beta_i = 0$,

and the alternative hypothesis is,

H_1: $\beta_i \neq 0$,

where $i = \mathbf{age}, \mathbf{sex}, ..., \mathbf{glu}$.

The p value is valid if

$$P(p(\mathbf{Y}) \leq \alpha) \leq \alpha$$

for all $0 \leq \alpha \leq 1$ whenever H_0 is true. This ensures that the probability for a type I error does not exceed α.

b)

In order to evaluate the fit of the full model we will check whether the assumptions of a linear model is satisfied. These are:

1. The expected value of ϵ_i is 0: $E(\epsilon_i) = 0$.
2. All ϵ_i have the same variance: $Var(\epsilon_i) = \sigma^2$.
3. The ϵ_i are normally distributed.
4. The ϵ_i are independent of each other.

Looking at the Residual vs. Fitted plot in Figure 2, there are no presence of a pattern between the residuals and the predicted values $\hat{y}_i$. This helps us validate assumptions 1 and 2. Looking at the top part of Figure 2, we notice that there are no signs of patterns between each of the covariates and the response variable. The QQ-plot in the right bottom of Figure 2 does not show any evidence against assumption 3. In conclusion we have no evidence against the model not satisfying a linear relationship between the response and the covariates. And we can therefore conclude that the linear regression seems to be significant for our data.

Another method in order to test if the regression is significant is through a hypothesis test, where the nullhypothesis is

H_0: $\beta_{age} = \beta_{sex} = ... = \beta_{glu} = 0$,

and the alternative hypothesis is

H_1: at least one β_j is non zero.

The Multiple R-squared in Figure 1 can be interpreted as how well the model is fitting the actual data. The value is always defined in the interval between 0 and 1, where 1 indicates a good model fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{e_i}^2}{\sum_{i=1}^{n} (y_i - \bar{y_i})^2}$$

c)

A reduced model can have better performance than a full model when the aim is prediction if any of the covariates of low significance are removed. Then we are left with a model consisting of less covariates where each has a higher correlation to the response variable.

Best subset model selection is a method for selecting subsets of predictors. For each possible combination of p predictors we fit a least squares regression model. Then we look at the resulting models in order to try to identify the best of them. To do this we define adjusted R^2 and BIC criteria. These are techniques that adjust the training error for the model size opposite to for example MSE (mean square error) which decreases as more variables are included in the model, or R^2 which increases in the same case.

In Figure 3a best subset selection has been performed on our dataset. The function identifies the best model that contains a given number of predictors from 1 to 10 in this case. The best model is quantifies using RSS (residual sum of squares).

Now we want to choose a reduced regression model based on the results in Figure 3 and Figure 4. By observing Figure 3 we see that the BIC criteria has a minimal value for 5 predictors, and adjusted R^2 has a maximal value for 8 predictors. Further, we observe the plots in Figure 4 and find that 6 predictors gives both a small BIC value and a large adjusted R^2 value. The predictors in our reduced model is therefore: sex, bmi, map, tc, ldl and ltg from the results in figure 3.

```
ds <- read.csv("https://web.stanford.edu/~hastie/CASI_files/DATA/diabetes.csv", sep = ",")

#fit the reduced model
reduced_6 <- lm(prog ~ sex + bmi + map + tc + ldl + ltg, data = ds)

summary(reduced_6)
```

```
##
## Call:
## lm(formula = prog ~ sex + bmi + map + tc + ldl + ltg, data = ds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -158.818  -39.184   -2.126   37.391  148.910
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -334.9019    25.3094 -13.232  < 2e-16 ***
## sex          -21.5052     5.7045  -3.770 0.000186 ***
## bmi            5.7040     0.7077   8.060 7.45e-15 ***
## map            1.1260     0.2159   5.216 2.83e-07 ***
## tc            -1.0391     0.2206  -4.710 3.33e-06 ***
## ldl            0.8395     0.2296   3.656 0.000288 ***
## ltg          168.5354    16.8138  10.024  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.07 on 435 degrees of freedom
## Multiple R-squared:  0.5148, Adjusted R-squared:  0.5081
## F-statistic: 76.91 on 6 and 435 DF,  p-value: < 2.2e-16
```

The reduced model is:

$$Y = -334.9019 - 21.5052\mathbf{sex} + 5.7040\mathbf{bmi} + 1.1260\mathbf{map} - 1.0391\mathbf{tc} + 0.8395\mathbf{ldl} + 168.5354\mathbf{ltg}$$

We will now compare the estimated regression parameters for the full and reduced model. The t value is a measure of how many standard deviations our coefficent estimate is away from zero. When the parameter is far away from zero we can reject H_0, and conclude that there is a linear relationship between the parameter and the response. We observe that the t values for the reduced model are higher for all the chosen covariates compared to the full model. This means that the covariates are further away from zero, and the chance of rejecting the null hypothesis for the reduced model is higher.

d)

We will now perform a hypothesis test for a new reduced model with the five covariates: **sex**, **bmi**, **map**, **hdl**, **ltg** and the intercept.

```r
#create linear models with a set of covariates
model_reduced <- lm(prog~ sex + bmi + map + ltg + hdl, data=ds)
model_full <- lm(prog ~., data=ds)

#perform hypothesis test for the two models with H0 as defined in problem
anova(model_reduced, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: prog ~ sex + bmi + map + ltg + hdl
## Model 2: prog ~ age + sex + bmi + map + tc + ldl + hdl + tch + ltg + glu
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    436 1288082
## 2    431 1264264  5     23817 1.6239 0.1523
```

From the code above we observe that the p value for the test is $p = 0.1523$ and $F = 1.6239$. The function above tests whether the removed covariates in the reduced model are zero (H_0). If so, including these covariates in our reduced model would probably be a mistake. We observe that the p value is greater than

the usual significance level of 0.05. This indicates that the null hypothesis is true, which means that we should choose the reduced model that excludes the covariates in H_0.

We can also look at the F-statistic which is a good indicator of whether there is a relationship between our predictors and the response variable. A higher value of F gives a better reason for rejecting the null hypothesis. As the F statistics has a value close to zero, this comfirms the choice of approving H_0 and concluding with the reduced model as a better model compared to the full model.

Problem 2 : Multiple testing

Let us start with creating the code to analyze the given data set of p-values.

```
#loading data
pvalues <- scan("https://www.math.ntnu.no/emner/TMA4267/2018v/pvalues.txt")

#split data set into positive and negative findings from 2c)
true <- pvalues[1:900]
false <- pvalues[901:1000]

#initilize steps for true findings
i <- 1

#remove all findings smaller than significance level
while (i <= length(true)){
  if (true[i] < (0.05)){
    true <- true[-i]
  }
  else{
    i <- i+1
  }
}

#initilize i again for false findings
i <- 1
while (i <= length(false)){
  if (false[i] < (0.05)){
    false <- false[-i]
  }
  else{
    i <- i+1
  }
}

cat("Number of rejections", 1000 - (length(true) + length(false)))
```

```
## Number of rejections 155
```

```
cat("\nAmount of type I errors", 900-length(true))
```

```
##
## Amount of type I errors 55
```

```
cat("\nAmount of type II errors", length(false))
```

```
##
## Amount of type II errors 0
```

a)

We find that there is 155 rejections of the null-hypothesis H_0. A false positive finding (type I error) is when H_0 is rejected when it is actually true. We do not know the number of false positive findings, but the probability of at least one false positive finding with significance value 0.05 is between 23% and 50%.

b)

The familywise error rate, FWER, is the probability of finding at least one false positivie finding. To control the FWER at level 0.05 mean that you have the control of all overall type I error rate under any combination of true and false hypothesis tests. The Bonferroni method use

$$\alpha_{loc} = \frac{\alpha}{m}$$

where m is the total number of tests. In our problem $m = 1000$ which gives $\alpha_{loc} = 5 * 10^{-5}$. We apply the Bonferroni method.

```
true1 <- pvalues[1:900]
false1 <- pvalues[901:1000]
i <- 1

while (i <= length(true1)){
  if (true1[i] < (5 * 10^(-5))){
    true1 <- true1[-i]
  }
  else{
    i <- i+1
  }
}

i <- 1
while (i <= length(false1)){
  if (false1[i] < (5 * 10**(-5))){
    false1 <- false1[-i]
  }
  else{
    i <- i+1
  }
}
cat("Number of rejections", 1000 - (length(true1) + length(false1)))
```

```
## Number of rejections 50
```

```
cat("\nAmount of type I errors with Bonferroni", 900-length(true1))
```

```
##
## Amount of type I errors with Bonferroni 0
```

```
cat("\nAmount of type II errors with Bonferroni", length(false1))
```

```
##
## Amount of type II errors with Bonferroni 50
```

After applying the Bonferroni method the number of rejections is lowered from 155 to 50.

c)

We split the test in the way described in the exercise and calculate the type I and II errors.

```
smoke <- matrix(c(155, 50, 55, 0, 0, 50),ncol=3,nrow=2)
colnames(smoke) <- c("Rejections","Type I ","Type II")
rownames(smoke) <- c("Original","With Bonferroni")
smoke <- as.table(smoke)
smoke
```

```
##                 Rejections Type I  Type II
## Original               155      55       0
## With Bonferroni         50       0      50
```

We can know observe that the Bonferroni reduces the type I errors, but the type II errors increases.